UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Igor Wescley Silva de Freitas

Um estudo comparativo de técnicas de detecção de *outliers* no contexto de classificação de dados

Mossoró-RN

2019

Igor Wescley Silva de Freitas

Um estudo comparativo de técnicas de detecção de *outliers* no contexto de classificação de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Profº Dr. Daniel Sabino Amorim de Araújo

Mossoró-RN

2019

F866e Freitas, Igor Wescley Silva de.
Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados / Igor Wescley Silva de Freitas. - 2019.
97 f. : il.

Orientador: Daniel Sabino Amorim de Araújo.
Dissertação (Mestrado) - Universidade Federal Rural do Semi-árido, Programa de Pós-graduação em --Selecione um Curso ou Programa--, 2019.

1. Outliers. 2. Detecção de Outliers. 3. Classificação. 4. Metodologia. I. de Araújo, Daniel Sabino Amorim, orient. II. Título.

Igor Wescley Silva de Freitas

Um estudo comparativo de técnicas de detecção de *outliers* no contexto de classificação de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 25 / 01 / 2019.

BANCA EXAMINADORA

Profº Dr. Daniel Sabino Amorim de Araújo
Orientador

Profª Dra. Isabel Dillmann Nunes
Avaliadora Externa

Profº Dr. Araken de Medeiros Santos
Avaliador do Programa

A meu Pai, **Francisco Telmo de Freitas** [In Memoriam], que se estivesse presente, de certo estaria muito feliz por mais uma etapa da vida acadêmica de seu filho concluída.

Agradecimentos

Sem a presença de Deus todos os dias, seria impossível estar realizando este trabalho, superando todos os obstáculos e chegando até o fim. Obrigado Senhor pelo cuidado, proteção e pela saúde.

Dona Joana D'arc sempre me apoiou em meus planos e sonhos, me incentivando e aconselhando em todos os momentos. Obrigado minha Mãe, não teria finalizado sem sua presença, esse trabalho também é seu.

Minha profunda gratidão ao professor Daniel Sabino por me orientar e por dar todo apoio necessário para o desenvolvimento deste trabalho. Ter a oportunidade de ter sido seu aluno e agora trabalhar com você foi muito gratificante.

Aos professores deste programa de pós-graduação, pela valiosa missão de compartilhar o conhecimento, sou muito grato. Esse agradecimento se estende também para os funcionários técnicos administrativos que contribuem para o funcionamento do programa, sempre prestativos e dedicados.

Ao amigos de pós-graduação, por essa caminhada juntos, pelo apoio e sugestões nos momentos de dificuldade.

Aos amigos do trabalho e chefias da UFERSA Campus Angicos, pela compreensão. Sou muito grato a Deus por fazer parte dessa família. O apoio de vocês foi imprescindível para que eu pudesse chegar até o fim.

Aos grandes amigos da vida, Kelvin, José Antônio, Kelson, Nicácio, Fernando, Gilvan e Giovanna, por entender a ausência e compartilhar momentos de alegria e distração.

À toda minha família, pelo amor e carinho. Sempre me incentivaram a trilhar o caminho certo, dos estudos e da responsabilidade. Meu agradecimento especial.

”O pouco que podemos ver do futuro, é o suficiente para perceber que há muito a fazer. ”

Alan Turing

Resumo

Outliers são objetos que se desviam consideravelmente dos demais em relação a alguma medida, e promovem grande influência na análise dos dados. Na estatística, essa influência pode induzir uma análise equívoca dos dados, neste caso, os *outliers* constituem dados que precisam ser removidos. Para outras aplicações, o *outlier* pode representar alguma informação valiosa, tratando-se de algum tipo de fraude, intrusão em sistemas, anomalias em redes de computadores, falhas mecânicas e condição clinica crítica. Para todo caso, os *outliers* precisam ser identificados, independente de seu tratamento. A literatura fornece diversas técnicas para detecção de *outliers*, cada uma com suas características e especificidades, que por sua vez foram aplicadas em diversos domínios, tendo em vista resolver problemas singulares. Precisar qual técnica tem melhor desempenho para determinado domínio de dados, constitui um desafio ainda pouco explorado na literatura e provoca o desenvolvimento de estratégias, para mensurar a performance de técnicas de detecção de *outliers*. Nesse sentido, a proposta deste trabalho é apresentar um estudo comparativo de técnicas de detecção de *outliers*, através de uma metodologia que permita uma análise uniforme e objetiva. As técnicas utilizadas na análise comparativa estão distribuídas em técnicas baseadas em métodos estatísticos, proximidade e distância. Como parte da metodologia, elas são aplicadas no pré-processamento dos dados, onde seu desempenho é mensurado analisando o efeito desta aplicação na indução de classificadores. As métricas de avaliação de classificadores funcionam como indicadores de desempenho das técnicas. De acordo com os resultados dos experimentos realizados, foi possível analisar efetivamente o desempenho das técnicas de detecção de *outliers* para diferentes domínios, e confirmar a validade da metodologia.

Palavras-chave: *Outliers*; Detecção de *Outliers*; Classificação; Metodologia.

Abstract

Outliers are objects that deviate considerably from others in relation to some measure, and promote great influence in the analysis of the data. In statistics, this influence may induce an equivocal analysis of the data, in which case the outliers constitute data that need to be removed. For other applications, the outlier may represent some valuable information, dealing with some type of fraud, system intrusion, computer network anomalies, mechanical failures and critical clinical condition. In any case, outliers need to be identified, regardless of their treatment. The literature provides several techniques for detection of outliers, each with its characteristics and specificities, which in turn have been applied in several domains, in order to solve singular problems. To specify which technique performs better for a particular data domain is a challenge that is still little explored in the literature and causes the development of strategies to measure the performance of outliers detection techniques. In this sense, the proposal of this work is to present a comparative study of outliers detection techniques, through a methodology that allows a uniform and objective analysis. The techniques used in the comparative analysis are distributed in techniques based on statistical methods, proximity and distance. As part of the methodology, they are applied in the pre-processing of the data, where their performance is measured by analyzing the effect of this application on the classifier induction. Classifier evaluation metrics serve as performance indicators for classifiers. According to the results of the experiments, it was possible to effectively analyze the performance of outliers detection techniques for different domains, and confirm the validity of the methodology.

Keywords: Outliers; Outlier Detection; Classification; methodology.

Lista de ilustrações

Figura 1 – Componentes envolvidos no problema de detecção de *outliers*. Adaptado de (CHANDOLA; BANERJEE; KUMAR, 2009) 23
Figura 2 – Ilustração de um *outlier* global 24
Figura 3 – Exemplo de um *outlier* contextual. Adaptado de (SINGH; UPADHYAYA, 2012) 25
Figura 4 – Ilustração de um *outlier* coletivo. Adaptado de (HAN; PEI; KAMBER, 2011) 26
Figura 5 – Visualizando *outliers* usando *box-plot*. Adaptado de (HAN; PEI; KAMBER, 2011) 33
Figura 6 – Ilustração de distância *mahalanobis*. Adaptado de (DEBES; KOENIG; GROSS, 2005) 35
Figura 7 – Demonstração visual do método de Probabilidade de Correlação *Outlier* (COP). Adaptado de (KRIEGEL *et al.*, 2012) 36
Figura 8 – Ilustração de um histograma com dados *outliers* 38
Figura 9 – Detecção de *outliers* em grupos de diferentes densidades. Adaptado de (BREUNIG *et al.*, 2000) 41
Figura 10 – Funcionamento das propriedades do LOF. Adaptado de (BREUNIG *et al.*, 2000) 43
Figura 11 – Modelo de Classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009) 46
Figura 12 – Abordagem geral para a construção de um modelo de classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009) 47
Figura 13 – Exemplo de Bia e Variação. Adaptado de (LAROSE, 2014) 48
Figura 14 – O nível ideal de complexidade do modelo corresponde a menor taxa de erro no conjunto de testes. Adaptado de (LAROSE, 2014) 49
Figura 15 – Ilustração do método *Holdout*. Adaptado de (HAN; PEI; KAMBER, 2011) 51
Figura 16 – Metodologia proposta para comparar técnicas de DO. 64
Figura 17 – Demonstração do *Z-Score* na distribuição normal 65
Figura 18 – Parâmetros do limiar crítico para os experimentos 74

Lista de tabelas

Tabela 1 – Tabela de confusão para um modelo de classificação binário. Adpatado de (TAN; STEINBACH; KUMAR, 2009) 51
Tabela 2 – Resumo dos trabalhos relacionados 61
Tabela 3 – Resumo de informações sobre os conjunto de dados 68
Tabela 4 – Características do recurso computacional utilizado nos experimentos . 74
Tabela 5 – Valores estabelecidos dos parâmetros do algoritmos de DO. 75
Tabela 6 – Taxas de *outliers* identificados pelos algoritmos de DO 78
Tabela 7 – Taxas de *outliers* identificados pelos filtros de classificação 78
Tabela 8 – Taxa de melhoria,piora e vitórias após a fase de pré-processamento. Taxa de casos em que não foi identificado *outliers* (Coluna "*Outliers*"). Resultado total por algoritmos de DO. 79
Tabela 9 – Desempenho geral dos classificadores antes e após o pré-processamento das instâncias *outliers*, identificadas pelos algoritmos de DO. 80
Tabela 10 – Taxa de melhoria e piora após a fase de pré-processamento. Correlação de classificadores com bases e taxa total por base. 81
Tabela 11 – Taxa de melhoria e piora após a fase de pré-processamento. Resultado total por classificador. 82
Tabela 12 – Taxa de melhoria após a fase de pré-processamento. Correlação classificador com algoritmos de DO. 82
Tabela 13 – Taxa de piora após a fase de pré-processamento. Correlação classificador com algoritmos de DO. 82
Tabela 14 – Taxa de melhoria após a fase de pré-processamento. Correlação de algoritmos de DO com bases 83
Tabela 15 – Taxa de piora após a fase de pré-processamento. Correlação de algoritmos de DO com bases. 83
Tabela 16 – Desempenho geral dos classificadores antes e após o pré-processamento das instâncias *outliers*, identificadas pelos filtros de classificação. 85
Tabela 17 – Taxa de melhoria,piora e vitórias após a fase de pré-processamento. Taxa de casos em que não foi identificado *outliers* (Coluna "*Outliers*"). Resultado total por filtros de classificação. 86
Tabela 18 – Taxa de melhoria e piora após a fase de pré-processamento. Resultado total por classificador. 86
Tabela 19 – Taxa de melhoria após a fase de pré-processamento. Correlação de filtros de classificação com bases. 87
Tabela 20 – Taxa de melhoria após a fase de pré-processamento. Correlação classificador com filtros de classificação. 87

Lista de abreviaturas e siglas

DO - Detecção de *Outliers*

AD - Árvores de Decisão

KNN - K-Vizinho mais Próximo

NB - Naive Bayes

SVM - Máquina de Vetores de Suporte

LOF - Fator *Outlier* Local

LDOF - Fator Outlier Baseado na Distância Local

LDF - Fator de Densidade Local

LOCI - Integral de Correlação Local

MDEF - Fator de Desvio de Multi-Granularidade

TCM-KNN - *Transductive Confidence Machines for K-Nearest Neighbors*

INFLO - *InfLuenced Out-lierness*

NHL - Liga de *Hockey* Nacional

EEG - Eletroencefalograma

EM - Expectation-Maximization

LRD - Fator de Acessibilidade Local

MSE - Erro Médio Quadrático

PAAF - Punção aspirativa de agulha fina

VP - Verdadeiro Positivo

VN - Verdadeiro Negativo

FP - Falso Positivo

FN - Falso Negativo

NS - Nível de Significância

TWX - Teste de Wilcoxon

TFR - Teste de Friedman

DWOF - Fator *Outlier* de janela dinâmica

COF - Fator *Outlier* Baseado em Conectividade

LOOP - Probabilidade *Outlier* Local

COP - Probabilidade de Correlação Outlier

DBScan - *Density-Based Spatial Clustering of Aplications with Noise*

JRP - JRipper

IBK - Classificador baseado em K-instâncias

FR - Classificador *Fuzzy* Baseado em Regras

FURIA - Algoritmo de Indução de Regras *Fuzzy* Não-Ordenadas

Sumário

1 INTRODUÇÃO 15
1.1 Motivação 17
1.2 Objetivos 18
1.3 Contribuições 18

2 REFERENCIAL TEÓRICO 20
2.1 *Outliers* 20
2.1.1 Causas de Anomalias 20
2.1.2 Detecção 21
2.1.3 Aspectos Relacionados ao Problema de Detecção de *Outliers* 22
2.1.3.1 Natureza dos dados 23
2.1.3.2 Tipos de *Outliers* 24
2.1.3.3 Disponibilidade de Rótulos 26
2.1.3.4 Classificação de Saída 27
2.1.3.5 Aplicações 27
2.2 Abordagens e Métodos para Detecção de *Outliers* 29
2.2.1 Abordagens para Problemas de Classificação 30
2.2.1.1 Abordagem Supervisionada 30
2.2.1.2 Abordagem Não Supervisionada 30
2.2.1.3 Abordagem Semissupervisionada 31
2.2.2 Métodos para Detecção de *Outliers* 31
2.2.2.1 Métodos Estatísticos 32
2.2.2.1.1 Métodos Paramétricos para Dados Univariados 33
2.2.2.1.2 Métodos Paramétricos para Dados Multivariados 34
2.2.2.1.3 Métodos Paramétricos para Múltiplas Distribuições 36
2.2.2.1.4 Métodos não Paramétricos 37
2.2.2.2 Métodos baseados em Proximidade 39
2.2.2.2.1 Métodos baseados em Distância 39
2.2.2.2.2 Métodos baseados em Densidade 41
2.2.2.3 Métodos baseados em Agrupamento 44
2.3 Classificação 45
2.3.1 Abordagem Geral para Classificação 46
2.3.2 Bias e Variação 47
2.3.3 Métricas de Avaliação 49
2.3.3.1 Avaliando a Performance de um Modelo 50

2.3.4 *Holdout* e Outros Métodos para Estimar a Precisão 51
2.3.4.1 Validação Cruzada 52
2.3.5 Métodos Estatísticos para Comparar Modelos de Classificação 52
2.3.5.1 Teste de *Wilcoxon* 53
2.3.5.2 Teste de *Friedman* 54
2.4 Considerações Finais 55

3 TRABALHOS RELACIONADOS 56
3.1 Considerações Finais 60

4 METODOLOGIA PARA COMPARAR TÉCNICAS DE DETECÇÃO DE *OUTLIERS* 62
4.1 Descrição da Metodologia 63
4.2 Materiais e Métodos 65
4.2.1 Bases de dados 66
4.2.2 Técnicas de DO 68
4.2.3 Algoritmos Indutores 72
4.2.4 Configuração do Experimento 73
4.2.5 Métricas de Avaliação 75
4.3 Considerações Finais 76

5 RESULTADOS EXPERIMENTAIS 77
5.1 Considerações Finais 88

6 CONSIDERAÇÕES FINAIS 89

REFERÊNCIAS 91

1 Introdução

Na natureza, na sociedade e no domínio virtual de dados e informação, eventos ou objetos, por definição, devem possuir um padrão comum. Podemos abstrair tudo como um processo, cujo seu fluxo principal é definido como "normal". Por sua vez, sabendo o que é "normal", temos uma sensível percepção de que eventos que se desviam do fluxo principal, ou objetos que se distingue dos demais sejam identificados como diferentes ou "anormais". Podemos incluir como "anormais", por exemplo, estações de seca prolongada, atletas que superam recordes e um atributo com valor muito menor que os demais. Nesse contexto, eventos e objetos anômalos tem mais importância dos que os normais. Uma seca tem graves consequências para a colheita, habilidades especiais de um atleta podem levar a vitória e valores de atributos distantes dos demais pode influenciar um experimento (TAN; STEINBACH; KUMAR, 2009). Por outro lado, para algumas aplicações, *outliers* podem dificultar a análise dos dados, alterando resultados de sumarizações e estimativas. Logo, detectar eventos e objetos anômalos é o primeiro passo para obtenção de informação e ganho de conhecimento.

A detecção de anomalia é o processo de encontrar objetos de dados com comportamento muito diferente do esperado. Tais objetos são comumente chamados de *outliers*. Podemos também entender um *outlier* como um objeto que se desvia de forma significativa dos demais, como se fosse produzido por um mecanismo diferente, ou tenha recebido interferência externa ao processo normal de seu domínio (HAN; PEI; KAMBER, 2011). São diversos os fatores responsáveis pelas anomalias. Para Tan, Steinbach e Kumar (2009) as mais comuns estão definidas como: dados de classes diferentes, em que um objeto anômalo é um tipo ou classe distinta; variação natural, onde a anomalia se caracteriza por alguma variação intrínseca do dado; e medidas de erro e erros de coleta, em que a causa da anomalia encontra-se no processo de coleta ou medição do dado. Suas origens são muitas vezes desconhecidas, no entanto, elas só são importantes quando se diz respeito à aplicação.

Dependendo do domínio de aplicação, o *outlier* pode ser referido como observações discordantes, exceções, aberrações, peculiaridades ou até mesmo como contaminantes. Podemos citar alguns contextos onde processos para detectar anomalias são usados, como na descoberta de fraudes, em cartões de crédito por exemplo, onde comportamento atípico de compra pode ser considerado como uma anomalia; na detecção de intrusos, no qual sistemas são monitorados para prevenir ataques, ou até mesmo falhas em caso de sistemas críticos; em distúrbios no ecossistema, com objetivo de prever a probabilidade de eventos naturais catastróficos acontecerem; na saúde pública, onde o aparecimento de doenças em determinadas localidades pode indicar problemas sanitários; na medicina, em que sintomas

incomuns ou testes clínicos pode indicar problemas de saúde; no contexto militar, onde bases inimigas são monitoradas(RANA; PAHUJA; GAUTAM, 2014)(TAN; STEINBACH; KUMAR, 2009). O fato é que detectar anomalias, para inúmeras e diversas aplicações é um processo indispensável e valioso.

A maior parte das técnicas de detecção de *outliers* (DO) foram desenvolvidas para domínios de aplicação específicos, enquanto outras são mais gerais, mas não tão genéricas de forma a ser eficiente para qualquer tipo de aplicação (NIU *et al.*, 2011). Dessa forma, diversos aspectos são considerados quando se categoriza técnicas de DO. Dentre elas estão o tipo de *outliers* encontrado, dimensionalidade, natureza dos dados, disponibilidade de rótulos e necessidade ou não de parâmetros (ZAMONER, 2013).

Para o problema de classificação, existem três abordagens comuns para DO, são elas: a não supervisionada, quando os rótulos das classes não estão disponíveis; a abordagem supervisionada, no qual todo o conjunto de treinamento contém objetos com rótulos normais e anômalos; e a abordagem semissupervisionada, quando é disposta uma situação onde o conjunto contém objetos com rótulos normais, referentes ao problema de classificação, e informações de dados anômalos é inexistente (TAN; STEINBACH; KUMAR, 2009)(WU; BANZHAF, 2010). É importante ressaltar, que dados rotulados que envolvem anomalias são caros e muitas vezes difíceis de adquirir. Dados anômalos tem natureza dinâmica, envolvem comportamentos inesperados e podem surgir somente em acontecimentos catastróficos, como na queda de um avião por exemplo. (CHANDOLA; BANERJEE; KUMAR, 2009)

Deste modo, diversas técnicas para detecção de anomalias foram implementadas com objetivo de serem empregadas nos inúmeros domínios de aplicação, diminuindo os problemas ocasionados pela característica e circunstância pelo qual o conjunto de dados foi obtido. Diferentemente da classificação, quando não existe informação prévia dos dados, as principais técnicas compreendem as baseadas em modelos, ou métodos estatísticos, fazem suposições sobre a normalidade dos dados. Eles assumem que objetos que não se enquadram no modelo estatístico gerado são *outliers*; nos métodos baseados em proximidade, onde assumem que determinado objeto é anômalo, caso seus vizinhos mais próximos estiverem suficientemente distantes. Esse método é divido em dois outros métodos, os baseados em distância e em densidade; e nos métodos baseados em agrupamento, que assumem objetos como normais aqueles que fazem parte de grandes *clusters*, enquanto objetos anômalos compõem pequenos *clusters* ou não fazer parte de nenhum. (HAN; PEI; KAMBER, 2011)

Cada técnica possui sua particularidade e se diferenciam quanto à eficiência computacional, a capacidade de operar com conjuntos de grandes dimensões, parametrização e aplicabilidade (WU; BANZHAF, 2010). No entanto, seja nos casos em que técnicas sejam desenvolvidas para solucionar um problema específico, ou aquelas mais genéricas, validar a sua eficácia ainda constitui um obstáculo, visto que o problema de detecção de *outliers* é, na maioria dos casos, um problema não supervisionado (CHANDOLA; BANERJEE;

KUMAR, 2009). Ainda assim, o grande desafio de um estudo comparativo entre tais técnicas não somente medir desempenho em determinado conjunto de dados, e sim medir a eficiência que cada técnica produz em diferentes contextos de aplicação.

1.1 Motivação

Um método ou abordagem para DO genérica ou universalmente aplicável é inexistente ou escasso. Os principais métodos existentes na literatura formam um amplo acervo de conhecimento e relevante para diferentes contextos e domínios. Todavia, o desenvolvedor deve selecionar a técnica adequada ao seu conjunto de dados, em relação ao modelo de distribuição, tipos de atributos, escalabilidade, desempenho e capacidade incremental, a fim de condicionar novos exemplos e a precisão do modelo. O desenvolvedor também deve considerar qual será a abordagem adequada, visto que a mesma depende do tipo dado, se estão rotulados ou não, integridade dos rótulos, de que maneira as anomalias devem ser detectadas e o quanto que a informação gerada pela detecção é útil para a aplicação (HODGE; AUSTIN, 2004). Desta maneira, definir um método que seja aplicável a um domínio ou contexto específico não é uma tarefa trivial. Quando na literatura, estudos que comparam estatisticamente diversas técnicas de DO aplicadas a diferentes domínios são raros, o processo para definir o uso de tal técnica é empírico ou muitas vezes intuitivo. Bem como a motivação da escolha, no qual pode não estar fundamentado.

Segundo Chandola, Banerjee e Kumar (2009), ao aplicar uma determinada técnica em determinado domínio, os pressupostos gerados podem ser usados como diretrizes para avaliar a eficácia da técnica. Idealmente, uma pesquisa abrangente sobre DO deve possibilitar ao leitor não apenas compreender a motivação por trás em usar uma técnica específica, como também fornecer uma análise comparativa de diferentes técnicas. Por outro lado, para realizar uma análise comparativa de forma efetiva, é necessário definir uma metodologia que possibilite mensurar as diversas técnicas uniformemente, com intuito de tornar a análise objetiva. Um estudo comparativo, baseado em uma metodologia com essa característica, possibilita não só medir eficiência de técnicas em contextos específicos, mas selecionar a melhor técnica para determinada situação.

Nos problemas de classificação, algoritmos indutores geram modelos de classificação baseando-se em conjunto de dados de treinamento. Esse processo está bem definido na literatura. Dessa forma, mensurar o desempenho das técnicas de DO, utilizando o processo de avaliação de classificadores, constitui uma estratégia para contornar os problemas supracitados, para comparar técnicas de DO em diferentes domínios.

A importância deste trabalho sustenta-se no fato de que, analisar de forma objetiva o desempenho de técnicas de DO em ambientes supervisionados, e em sua grande maioria não supervisionados, consiste em uma tarefa difícil de se atingir. Dado que, a subjetividade

gerada pelo contexto não supervisionado impede a avaliação efetiva do desempenho de cada técnica. Um trabalho com essa dimensão, traz uma nova perspectiva na análise comparativa de técnicas de DO, não somente por possibilitar a quantificação do desempenho individual, como também por permitir que técnicas de diferentes abordagens e métodos possam ser analisadas e comparadas de forma efetiva.

1.2 Objetivos

Como visto, a principal dificuldade em comparar o desempenho de técnicas de DO é a análise subjetiva dos resultados. Por se tratar, na grande maioria, de técnicas não supervisionadas, as métricas de avaliação não são efetivas. Partindo disso, este trabalho tem o propósito de realizar um estudo comparativo de diferentes técnicas de DO, em uma ampla diversidade de domínio de dados, de maneira que seja possível mensurar o desempenho por domínio de dados e realizar uma análise estatística dos resultados.

Para realizar um estudo dessa dimensão, vários desafios precisam ser superados. Podemos listar os objetivos específicos a serem realizados:

- Construir uma metodologia para análise comparativa, baseada na classificação de dados;
- Quantificar o desempenho das técnicas de DO através da metodologia proposta;
- Entender a relação entre os domínio de dados e os algoritmos de DO;
- Compreender como as características das bases afetam o desempenho das técnicas de DO;
- Estabelecer e entender como se comporta a eficácia e eficiência das técnicas de DO, no que tange a detecção e posteriormente remoção de *outliers*, como parte da metodologia proposta;
- Experimentar os filtros de classificação como técnicas supervisionadas de DO, e compreender a eficácia e eficiência no uso dessa estratégia;
- Perceber qual o efeito que os algoritmos indutores provocam no desempenho das técnicas de DO;

1.3 Contribuições

Este trabalho teve o propósito de comparar diversas técnicas de DO através de uma metodologia baseada em classificação de dados. Mensurar o desempenho das técnicas de DO, utilizando o processo de avaliação de classificadores, constitui uma estratégia

para contornar os problemas supracitados e uma valiosa contribuição para a literatura, possibilitando uma análise comparativas de diferente técnicas de DO, para diferentes domínios.

Diante a aplicação da metodologia, este trabalho conseguiu estabelecer quais técnicas foram mais eficientes, quais alcançaram eficácia relevante e quais técnicas obtiveram resultado inferior. Conseguir determinar de forma objetiva esse tipo de resultado para técnicas de DO pode ser considerado uma novidade para a literatura, visto que até então, a trabalhos publicados apresentam uma análise subjetiva, sem critérios específicos (AGGARWAL, 2015). A metodologia proposta também possibilitou entender, como determinadas características das bases de dados podem influenciar o desempenho das técnicas, como a quantidade de atributos, número de rótulos e desbalanceamento.

Para viabilizar a análise comparativa e aplicação da metodologia, foi necessário implementar o filtro *RemovedByElki* na ferramenta WEKA, que por sua vez, recebe os resultados gerados pelas técnicas de DO, nativas da ferramenta ELKI, e realiza de pré-processamento dos dados, removendo as instâncias detectadas como *outliers*. Podemos considerar o filtro *RemovedByElki* como uma contribuição significante para o presente trabalho.

Outra contribuição importante apresentada, foi a experimentação dos filtros de classificação como técnicas de DO. Originalmente técnicas supervisionadas, sua utilização como filtros de classificação é uma conhecida estratégia para detecção de *outliers* (GAMBERGER; LAVRAC; GROSELJ, 1999). A avaliação dos filtros de classificação através da metodologia proposta, possibilitou não só quantificar o desempenho dessas estratégias como técnicas de DO, mas também permitiu observar um novo viés para a otimização da classificação de dados.

2 Referencial Teórico

Este capítulo é dedicado à apresentação da fundamentação teórica, utilizada na construção da proposta deste trabalho. Dessa forma, ela é composta por assuntos que fazem parte do escopo de Detecção de *Outliers* (DO) e Classificação. O conhecimento extraído desses temas, serão os "pilares" no desenvolvimento de uma nova metodologia, para comparar técnicas de DO voltadas para o problema de classificação.

2.1 *Outliers*

Intuitivamente, podemos entender um *outlier* ou anomalias quando consideramos um processo estatístico aplicado a um conjunto de dados. Neste processo é observado que existem pontos que se desviam do padrão. Essa situação ilustra o conceito de *outliers*, onde podemos definir como todo aquele objeto que se distância de forma considerável do restante dos dados (HAN; PEI; KAMBER, 2011). Hawkins (1980) definiu *outlier* como uma observação que se desvia dos demais, como fosse produzido por um mecanismo distinto. Este assunto é amplamente discutido na comunidade acadêmica, e é objeto de estudo em diversas áreas como na análise de tráfego em redes computadores, na segurança da informação, fraudes financeiras, etc.

O termo *outlier*, apesar de sua definição estar relacionada a um objeto que foge da distribuição normal dos dados, para alguns autores, *outliers* são diferentes de dados ruidosos. Han, Pei e Kamber (2011) atribui o ruído a um erro ou uma variância aleatória percebida por alguma medição. Para Quinlan (1986) *outlier* é um conceito mais amplo e ruído está inserido neste escopo. Alpaydin (2014) descreve o termo ruído como algo indesejável no conjunto de dados, causado por inúmeros fatores, como inconsistência na fase de *input*, erro na fase de rotulação e até mesmo por valores que estão fora do domínio dos dados. De fato, existe muita semelhança entre *outlier* e ruído. O que os separam é forma de tratamento, geralmente o ruído é removido para uma melhor análise dos dados, e o *outlier* é um dado valioso para seu domínio de aplicação. Todavia, seja um dado *outlier* ou ruído, sua detecção é de suma importância independente do tratamento ou aplicação.

2.1.1 Causas de Anomalias

Algumas causas comuns que ocasiona o surgimento de anomalias nos dados são descritos por Tan, Steinbach e Kumar (2009). Dentre elas está a variação natural, medidas de dados e erros de coleta dados de diferentes classes.

Variação Natural: A variação natural na distribuição dos dados, está relacionado

a situação em que alguma característica ou atributo do dado se difere dos demais. Por exemplo, um carro extremamente caro não o transforma em uma classe diferente, mas é considerado um dado "anormal"apenas no sentido do preço. Ou seja, a maioria das instância de dados estão próximas de seu centro, e a probabilidade de um devido dado se distancie da média dos outros objetos é pequena.

Medidas de Dados e Erros de Coleta: O processo de medição ou de coleta de dados pode ser uma causa para produção de dados anômalos. Um erro humano pode causar o registro equivocado de um dado ou um dispositivo com defeito pode efetuar uma medição errada. A presença de ruído também é um fator que ocasiona medição errada dos dados. A remoção desse tipo anomalia é objeto de estudo no que tange o pré-processamento dos dados.

Dados de Classes Diferentes: No problema de classificação, existem dados que divergem das classes normais. Neste caso, dados que se inserem nesta situação são considerado anômalos. Suponhamos que existam usuários de cartões de crédito que fazem uso de forma legítima, logicamente usuário que usam de forma fraudulenta pertencem a classe distinta. A mesma ilustração se aplica na detecção de intrusos em sistemas, no monitoramento ambulatorial e em falhas em componentes mecânicos por exemplo. Tais anomalias caracterizam-se muitas vezes como o objetivo da DO, e trata-se de dados valiosos para a mineração de dados.

As causas de anomalias no conjunto de dados podem ter diversas origens e dependem muito do domínio em que está inserido. Em muitos casos, anomalias específicas podem ser desconhecidas e não fazer parte das causas descritas supracitadas. Na prática, as técnicas de DO, de forma genérica, objetivam identificar instâncias que difiram substancialmente dos outros dados.

2.1.2 Detecção

Podemos abstrair uma anomalia, como um padrão diferente do comportamento normal esperado. Um viés para uma abordagem de DO, seria definir uma região que corresponde o comportamento normal, e rotular como anomalia qualquer instância que não pertencesse a esta região. Embora essa estratégia pareça ser simples, envolve muitos desafios. Um deles é definir a região normal dos dados, uma vez que identificar todo comportamento normal não é um tarefa trivial; Outro fator é quando anomalias são resultados de ações maliciosas. Pessoas responsáveis por essas ações, fazem com que tais comportamentos anômalos pareçam ser normais; Em muitos domínios o comportamento normal é constante evolução; Dependendo do domínio, o entendimento de anomalia é diferente; A disponibilidade de dados rotulados para treinar e validar modelos utilizados nas técnicas de DO as vezes é um problema. (CHANDOLA; BANERJEE; KUMAR, 2009)(HAN; PEI; KAMBER, 2011)

Diante esses desafios, o problema de DO em sua forma genérica não é simples de resolver. Normalmente, as técnicas de DO são desenvolvidas para tratar anomalias em sua formulação específica. Su e Tsai (2011) estreita esse problema identificando e descrevendo tópicos vinculados as estratégias de DO. Podemos dividir o problema de DO em **Detecção de Novidades**, **Detecção de Anomalias** e **Remoção de Ruídos**.

Detecção de Novidades: Esse tópico resume-se a identificar instâncias ainda não observadas. Técnicas para esse tipo de detecção normalmente estão inserida do contexto de aprendizado de máquina. Instâncias novas são aquelas que não fizeram parte do conjunto de treinamento usado para construir o modelo, e difere substancialmente dos dados conhecidos.

Detecção de Anomalias: Nesse tipo de detecção o objetivo é identificar padrões que se desviam do comportamento normal esperado. Nesse caso, as anomalias são conhecidas ou tem um comportamento esperado. Embora para alguns domínios o comportamento anômalo seja raro, para outros casos as anomalias são frequentemente observadas.

Remoção de Ruídos: O ruído em conjunto de dados geralmente existe e dificilmente é evitado. Para Han, Pei e Kamber (2011), eles podem distorcer os dados, impossibilitando a distinção de objetos normais e anômalos. Basta um *outlier* estar presente no conjunto de dados para afetar a média e o desvio padrão. Técnicas para remoção de ruído visam remover o máximo de informação inútil, a fim de minimizar o prejuízo na análise dos dados.

A formulação do problema de DO envolve diversos fatores. Destaca-se como principais: a natureza dos dados, disponibilidade ou não de dados rotulados, os tipos de *outliers* e a classificação da saída de uma DO. Esses fatores, muitas vezes são determinados pelo domínio de aplicação, no qual tais anomalias são identificadas. Os conceitos adotados por cientistas de diversas áreas, como aprendizado de máquina, estatística e mineração de dados, foram aplicados nas formulações do problema de DO. A Figura 1 apresenta uma adaptação de Chandola, Banerjee e Kumar (2009), onde sumariza os componentes envolvidos em qualquer técnica de DO. A próxima subseção será dedicada a descrição e discussão das característica relacionadas ao problema de DO para o desenvolvimento de técnicas.

2.1.3 Aspectos Relacionados ao Problema de Detecção de *Outliers*

Esta subseção descreve e discute diversos aspectos que envolvem a detecção de *outliers*. O conhecimento desses aspectos é requisito indispensável na formulação do problema, e representa o principal subsídio para o desenvolvimento de técnicas de DO.

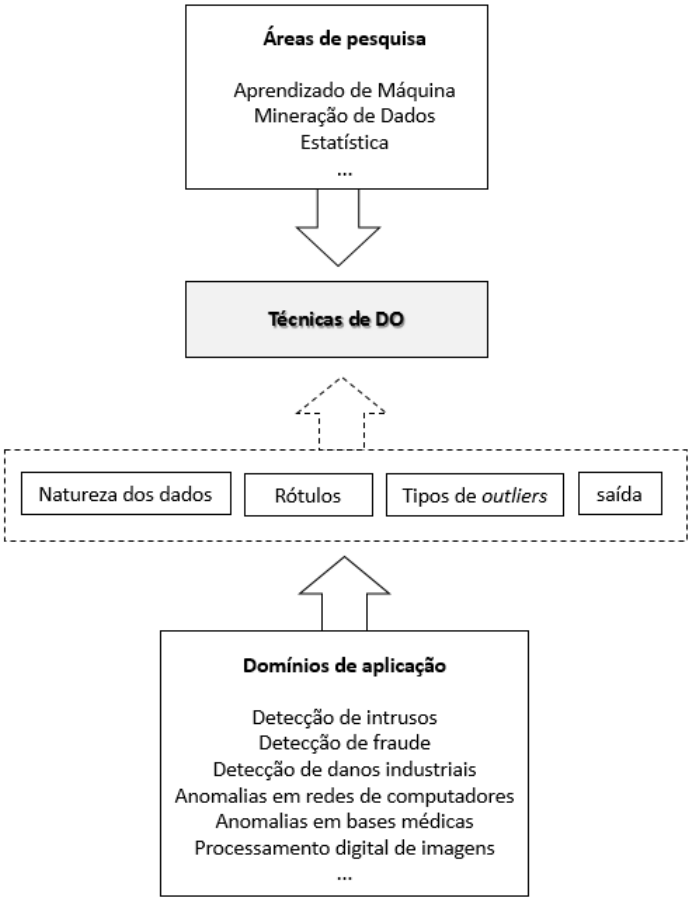


Figura 1 – Componentes envolvidos no problema de detecção de *outliers*. Adaptado de (CHANDOLA; BANERJEE; KUMAR, 2009)

2.1.3.1 Natureza dos dados

A natureza dos dados de entrada é um dos principais fatores na seleção de técnicas de DO. A entrada para essas técnicas consiste em um conjunto de instâncias de dados, que por sua vez pode ser compreendida como um conjunto de atributos ou características. Os atributos podem ser tipificados, como binários, categóricos ou contínuos. As instâncias de dados podem ser univariáveis, quando são compostas por somente um atributo, ou multivariáveis, quando são compostas por vários atributos do mesmo tipo ou de diferentes tipos. (TAN; STEINBACH; KUMAR, 2009)

Ter conhecimento sobre a natureza dos dados implica em poder determinar a aplicabilidade das técnicas de DO. Como exemplo, técnicas estatísticas, os modelos estatísticos se limitam a dados contínuos e categóricos. Da mesma maneira são as técnicas baseadas em vizinhança, as medidas de distância são baseadas nos tipos dos atributos. Existem situações em que a instância de dados é composta por uma matriz de distância ou similaridade, nesses casos, técnicas estatísticas e baseadas em classificação não são aplicáveis. (CHANDOLA; BANERJEE; KUMAR, 2009)

Existem outros aspectos que definem a natureza dos dados, onde as instâncias podem estar relacionadas de alguma forma. Os dados sequenciais são ordenados de forma linear, como exemplo os dados do genoma, séries temporais e sequências proteicas. Instâncias de tráfego veicular e climáticas configuram-se como dados espaciais, e em toda vizinhança dos dados existe alguma relação. Os dados em formato de grafos, são representados como vértices que estão conectados em outros (SINGH; UPADHYAYA, 2012). Diante o exposto, é inteligível perceber que compreender precisamente a natureza dos dados em questão, caracteriza-se como o primeiro passo desenvolver ou selecionar uma técnicas para DO.

2.1.3.2 Tipos de *Outliers*

Segundo Han, Pei e Kamber (2011), *outliers* podem ser classificados em três categorias, conhecidos como globais (ou pontuais), contextuais (ou condicionais) e coletivos. A seguir uma maior explanação sobre eles.

***Outlier* Global**: O caso mais simples dos três, ocorre quando um determinado objeto se desvia significativamente dos demais. Muitas vezes é chamado de *outliers* pontuais, devido a característica singular. A maioria das técnicas de DO propõem-se a identificar esse tipo de objeto, e o maior desafio para elas é aplicar a medida de desvio adequada em relação ao domínio de aplicação. A Figura 2 ilustra bem um *outlier* global, enquanto os dados normais tendem a ficar próximos entre si, o *outlier* global desvia-se consideravelmente do restante dos dados.

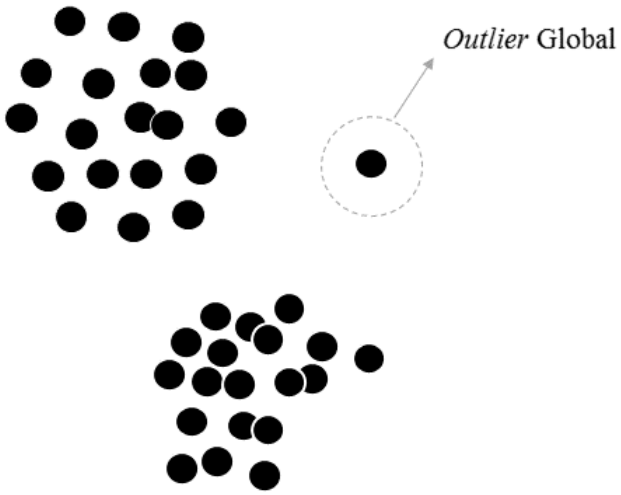


Figura 2 – Ilustração de um *outlier* global

***Outlier* Contextual**: Se uma instância de dados for anômala somente em um determinado contexto, então ela é um *outlier* contextual. Essa noção de contexto é proveniente da estrutura que consiste o conjunto de dados e é descrita como parte da formulação do problema. Podem ser denominados como *outliers* condicionais, porque estão condicionados a um contexto específico. Na detecção de *outliers* contextuais, cada

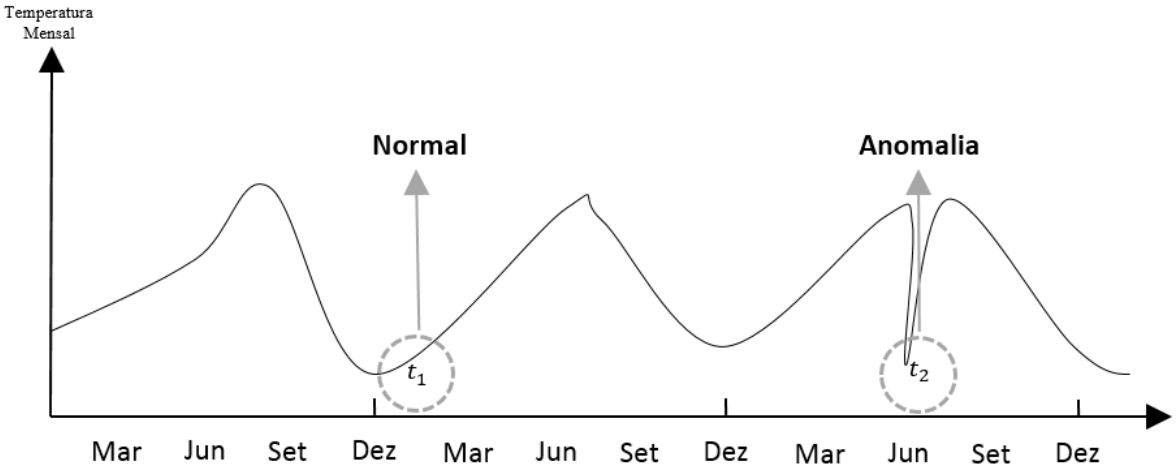


Figura 3 – Exemplo de um *outlier* contextual. Adaptado de (SINGH; UPADHYAYA, 2012)

instância é analisada através dos atributos contextuais e comportamentais.Os **Atributos Contextuais** são aqueles que caracterizam o objeto em seu contexto. Por outro lado, os **Atributos Comportamentais** definem as características do objeto e são úteis para avaliar se este objeto trata-se de um *outlier* no contexto em que pertence. Na Figura 3, Singh e Upadhyaya (2012) ilustra bem um exemplo de um *Outlier* Contextual. A série de tempo que representa a temperatura mensal de uma região nos últimos anos. Uma temperatura de 5^oC pode ser normal durante o inverno (no instante $t1$), no entanto o mesmo valor durante o verão (no instante $t2$) é um valor anormal. Outro exemplo seria uma pessoa com $1.70m$ de altura, pode ser considerado normal, mas quando é observado no contexto de idade, uma criança com essa mesma altura definitivamente é um *outlier*.

***Outliers* Coletivos**: Refere-se a um subconjunto dos dados considerado anormal quando formam um grupo. Em outras palavras, o *outlier* coletivo é atribuído ao subconjunto de objetos que se desviam como um todo, do conjunto de dados. A detecção desse tipo de *outlier* é muito importante para aplicações em que a anomalia está relacionada ao comportamento de várias instâncias. Para isso, é necessário o conhecimento aprofundado do relacionamento entre os objetos, como medidas de distância ou similaridade. A Figura 4 representa uma graficamente um *outlier* coletivo.

Em uma situação real, pode estar presente em um conjunto de dados mais de um tipo de *outlier*. O valor informacional que um *outlier* carrega é de grande importância para inúmeras organizações e negócios. Eles podem ser usados em diversas aplicações e para diferentes fins. A DO globais é o tipo mais simples. A DO contextual requisita mais informações para determinar atributos contextuais e comportamentais. A DO coletiva

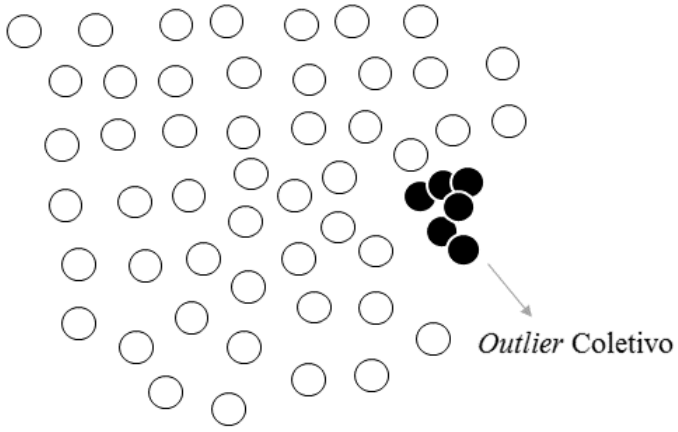


Figura 4 – Ilustração de um *outlier* coletivo. Adaptado de (HAN; PEI; KAMBER, 2011)

necessita de informações mais complexas, para modelar o relacionamento entre objetos, a fim de encontrar grupos *outliers*.(HAN; PEI; KAMBER, 2011)(CHANDOLA; BANERJEE; KUMAR, 2009)(SINGH; UPADHYAYA, 2012)

2.1.3.3 Disponibilidade de Rótulos

A disponibilidade de rótulos nos dados para o problema de DO, remete a um conjunto de dados que contenha instâncias classificadas como normal e/ou anormal. Muitas vezes, rótulos que representem o comportamento preciso do dado é difícil. Isso porque o método mais comum para classificação de dados reais é feita manualmente, por um perito da área. Isso encarasse o processo podendo impossibilitar a aquisição dos rótulos. Outro fator é rotular dados anômalos, dado que esse processo envolve todo possível comportamento anormal, como o surgimento de novos tipos de anomalia, eventos catastróficos ou raros. Normalmente, classificar dados com comportamento normal é mais simples.(CHANDOLA; BANERJEE; KUMAR, 2009)(SINGH; UPADHYAYA, 2012)

Dependendo do grau em que os rótulos estão disponíveis, a DO pode operar de três maneiras: O modo **supervisionado** assume que o conjunto de treinamento seja constituído de instâncias com rótulos normais e anormais. O modo **Semissupervisionado** compreende que o conjunto de dados contenha somente dados rotulados com comportamento normal. E no modo **Não Supervisionado** não requerem rótulos nos dados de treinamento, mas assumem que os dados normais são mais frequentes do que dados anômalos. Isso faz com que as técnicas não supervisionadas tenham uma maior aplicabilidade.(TANG *et al.*, 2002)(HAN; PEI; KAMBER, 2011)

2.1.3.4 Classificação de Saída

O objetivo principal de uma técnica de DO, além de identificar anomalias, é reporta-las. Chandola, Banerjee e Kumar (2009) descrevem duas formas mais comuns que as técnicas de DO reportam as anomalias:

Pontuação ou Escore: Um grau de anormalidade é dados para cada instância de teste. Uma etapa posterior é realizada para que tal instância, dependendo da sua pontuação, seja dada como *outlier* ou não. Técnicas que usam pontuação, normalmente retornam uma lista com *outliers* ranqueados e um *limiar* define quem de fato é uma anomalia.

Rótulo: Técnicas que geram esse tipo de saída, atribuem para cada instância um rótulo (Normal ou Anormal). Podem também definir a quantidade de *outliers* esperados ou um limiar que separa os dados normais dos anômalos.

As técnicas de DO baseadas em escore, tem uma maior liberdade para definir o que é um *outlier* ou não, permitindo ao analista definir um *limiar* para considerar as anomalias mais relevantes. Embora as técnicas baseadas em rótulos possam parametrizar tais condições, as baseadas em escore tem uma maior flexibilidade.

2.1.3.5 Aplicações

As aplicações que envolve DO geralmente estão associadas a minimização de prejuízos, e é objeto de interesse nos negócios do ramo industrial e comercial, bem como na área médica, promovendo automação. Apesar de algum tempo atrás, a presença de *outliers* estava associada a informações inúteis, onde o objetivo das aplicações de DO era de pré-processamento dos dados. Hoje, na mineração de dados, *outliers* constituem dados valiosíssimos e motiva estudos em diversos domínios.

Detecção de Intrusos: Atualmente, ataques a sistemas e redes de computadores são comuns. A detecção de intrusos refere-se a identificação de comportamento malicioso em um sistema informático, caracterizando uma invasão (PHOHA, 2007). O grande volume de dados é o grande desafio para esse domínio e aplicação, uma vez que a entrada de dados é relativamente alta e as técnicas precisam ser computacionalmente eficientes para atender uma análise *on-line*. Dados rotulados com comportamento normal geralmente tem maior disponibilidade, enquanto os rótulos para intrusões são mais raros. Dessa forma, técnicas de DO semissupervisionadas e não supervisionadas são mais comuns neste domínio.(PANNU *et al.*, 2017)(MANDIGOBINDGARH,)

Detecção de Fraude: Remete a identificação de crimes que ocorrem em estabelecimentos comerciais, bancos, cartões de crédito, seguros, telefonia, mercado de ações, etc. A fraude ocorre quando usuários clientes ou falsos clientes consumem recursos da organização de forma fraudulenta. A detecção imediata de tais crimes é a melhor forma

de evitar perdas econômicas, por isso boa parte das técnicas operam em modo *on-line*, monitorando a atividade do usuário. Casos investigados são usados para a construção de modelos supervisionados. Já as técnicas não supervisionadas são usadas para identificar fraudes desconhecidas. (TAN; STEINBACH; KUMAR, 2009)(KOU *et al.*, 2004)

Detecção de Danos Industriais: O uso contínuo de unidades responsáveis pelo processo industrial acarreta no desgaste normal de componentes. A captura desses dos dados é feita através de sensores instalados nas unidade onde monitora a atividade dos componentes. As técnicas visam identificar anomalias em componentes mecânicos, motores, em estrutura físicas e até mesmo em processos químicos. Para esse domínio, a detecção é feita antes do defeito ou problema se consumar, com objetivo de evitar perdas catastróficas. (CATENI; COLLA; VANNUCCI, 2008)(CHANDOLA; BANERJEE; KUMAR, 2009)

Detecção de Anomalias em Redes de Computadores: Apesar deste domínio de aplicação estar muitas vezes associado a detecção de intrusos, nem sempre uma anomalia percebida no tráfego da rede de computadores indica uma atividade maliciosa. Existem duas categorias que divide a detecção de anomalias neste domínio: A detecção de anomalia relacionada ao desempenho e a relacionada a segurança. Falha em servidores, tempestade de *broadcast* e congestionamento estão associados a anomalias de desempenho na rede. Já anomalias relacionadas a segurança são aquelas que podem ter origem em atividade mal-intencionadas, como a produção de tráfego desnecessário para sequestrar a largura da banda e deixar serviços indisponíveis. (BHUYAN; BHATTACHARYYA; KALITA, 2014)(THOTTAN; JI, 2003)

Detecção de Anomalias em Bases Médicas: A detecção de anomalias neste domínio atua através do monitoramento das atividades vitais de um paciente. As anomalias podem ser caracterizadas como uma condição anormal do paciente, erros de instrumentação ou gravação dos dados. Por se tratar de saúde, e envolver a vida de pessoas, os problemas para este domínio de aplicação geralmente são críticos e requerem técnicas com alto grau de precisão. Normalmente os dados consistem em registros de diferentes medidas, como peso, idade e tipo sanguíneo, ou podem ter aspectos temporais e para *outliers* coletivos, como no caso de aplicações para ECG. Geralmente, os dados rotulados são oriundos de pacientes com boa saúde, configurando dados normais, tornando mais comum o uso de técnicas com abordagem semissupervisionada. O maior desafio neste tipo de domínio é o custo e a disponibilidade de profissionais que classifiquem os dados. (CHANDOLA; BANERJEE; KUMAR, 2009)(SINGH; UPADHYAYA, 2012)(RANA; PAHUJA; GAUTAM, 2014)

Processamento Digital de Imagens: Detecção de anomalias neste domínio buscam alterações de imagem em uma série temporal, como na detecção de movimento por exemplo, e na identificação de regiões anormais em imagens estáticas. Algumas aplicações incluem detecção de anomalias em imagens de satélite (AUGUSTEIJN; FOLKERT, 2002) e em imagens hiperespectrais(BANERJEE; BURLINA; DIEHL, 2006). Os dados anômalos

possuem características espaciais e temporais, e normalmente são causados pelo movimentos, inserção artificial de um objeto ou erro instrumental. As instâncias de dados possuem atributos contínuos, como cor, nitidez e textura. O grande volume de dados de entrada é o grande desafio para técnicas desenvolvidas para este domínio. Uma aplicação que detecta movimento por exemplo, tem como entrada arquivos de vídeo e devem detectar *outliers* em tempo real. (CHANDOLA; BANERJEE; KUMAR, 2009)(SINGH; UPADHYAYA, 2012)

Existem outros domínios menos frequentes na literatura que abordam a detecção de *outliers*, como no monitoramento de tráfego de trânsito (DANG; NGAN; LIU, 2015), comportamento robótico (CROOK *et al.*, 2002)(TING; D'SOUZA; SCHAAL, 2007), aplicações *web* (HE; CHEN, 2017), distúrbios em ecossistema (KOU; LU; CHEN, 2006) e em dados biológicos (SUN; CHAWLA; ARUNASALAM, 2006).

2.2 Abordagens e Métodos para Detecção de *Outliers*

Esta seção é dedicada a apresentação de métodos de DO. Existem diversos métodos para detecção de anomalias na literatura, utilizadas em aplicações. Apresentamos aqui duas formas de categorização e descrição de métodos de DO. Primeiramente categorizamos de acordo com a disponibilidade de conjunto de dados rotulados. Métodos desta esfera constroem modelos baseando-se no conjunto de treinamento, com dados classificados por especialistas do domínio. Posteriormente, separamos os métodos segundo suas inferências feitas em relação aos dados normais e anormais. Destacamos aqui a descrição do funcionamento das principais técnicas para cada método. Por outro lado, os métodos aqui explanados generalizam a maneira que a maioria das técnicas existentes detectam anomalias.

Antes de iniciar a apresentação dos métodos de DO, objetivo desta seção, existem diversas questões importantes que necessitam ser abordadas quando se fala em métodos que lidam com anomalias. Essas questões foram amplamente discutidas na subseção 2.1.3, que trata dos aspectos relacionados quando se planeja selecionar ou desenvolver uma técnicas para DO, como entender a natureza dos dados do domínio, o tipo de **outlier** e a disponibilidade de rótulo nos dados. Outros aspectos importantes para ter conhecimento, constitui-se em saber se o problema se trata de identificação de anomalias de forma singular ou em conjunto. Isso está relacionado diretamente ao custo computacional envolvido, consequentemente a eficiência da técnica (HAN; PEI; KAMBER, 2011).

Técnicas que objetivam detectar um *outlier* por vez, estão sujeitas ao problema denominado de **mascaramento**, onde a presença de várias anomalias disfarça a presença de todas. No entanto, técnicas desenvolvidas para detectar múltiplos *outliers* de uma vez, podem sofrer **sobrecarga**, afetando a precisão da classificação. Por isso, a eficiência das técncas de DO deve ser levado em consideração. Abordagens baseadas em classificação

são mais custosas computacionalmente na fase de treinamento, mas depois que o modelo está pronto, a fase de aplicação tem menor custo. As abordagens estatísticas conseguem categorizar instâncias em tempo constante. Já abordagens baseadas em proximidade, naturalmente têm complexidade $O(n^2)$, fazendo com que a eficiência de sua execução dependa da dimensionalidade do conjunto de dados. (TAN; STEINBACH; KUMAR, 2009)

2.2.1 Abordagens para Problemas de Classificação

Dependendo do grau de instâncias rotuladas como normais e/ou anormais, esse conjunto pode ser usado na construção de um modelo para classificar dados anômalos. As abordagens básicas utilizadas são a **Supervisionada**, a **Não Supervisionada** e a **Semissupervisionada**.

2.2.1.1 Abordagem Supervisionada

Técnicas que utilizam a abordagem supervisionada, assumem que no conjunto de treinamento existem dados rotulados como normais e anormais. Tais técnicas constroem um modelo preditivo, onde a tarefa é identificar *outliers* através do classificador. Existem dois obstáculos para essa abordagem. O primeiro é o fato de que geralmente dados normais existem em maior quantidade, criando uma situação onde existam conjunto de dados desbalanceados. Em segundo lugar, a tarefa de adquirir uma quantidade representativa de dados anômalos é considerada difícil, visto que anomalias podem estar relacionadas a catástrofes ou ainda não aconteceram. (CHANDOLA; BANERJEE; KUMAR, 2009)(TAN; STEINBACH; KUMAR, 2009)

Em síntese, os método baseados em abordagem supervisionada constituem estratégias eficazes para a detecção de DO, porém dependem fortemente da condição do conjunto de dados. Caso esse conjunto esteja desbalanceado, o classificador resultante irá fazer interpretações imprecisas.(HAN; PEI; KAMBER, 2011)

2.2.1.2 Abordagem Não Supervisionada

Algumas situações fazem com que a indisponibilidade dos rótulos no conjunto de dados ocorra. Nesse caso, o objetivo de técnicas que seguem abordagem não supervisionada, é atribuir um grau de anormalidade para determinada instância. No entanto, para uma técnica de DO não supervisionada ser bem sucedida, as anomalias devem ser distintas entre sim, bem como os dados normais. Isso porque muitas anomalias semelhantes podem fazer com que elas sejam rotuladas como normais, ou ter um grau de anomalia baixo (TAN; STEINBACH; KUMAR, 2009). Muitas técnicas semi-supervisionadas podem ser adaptadas de forma a operaram em modo não supervisionado, desmarcando rótulos de uma amostra do conjunto de dados para servir como dados de treinamento. Essa adaptação é feita quando os dados de teste contém poucas anomalias, e o modelo aprendido no

treinamento é robusto demais para essa situação (CHANDOLA; BANERJEE; KUMAR, 2009).

Métodos não supervisionados partem de um pressuposto: Os dados normais estão mais agrupados do que os dados anômalos. Esses métodos esperam que os dados normais tenham um padrão com maior frequência. Isso não quer dizer que eles tenham que compartilhar o mesmo grupo, mas espera-se que um *outlier* esteja distante de dados normais, que por sua vez podem formar um ou mais grupos distintos. Em alguns casos os dados normais são diversamente distribuídos e os dados anômalos compartilham de alta similaridade em uma pequena área, como em aplicações de *outliers* coletivos, ilustrado na Figura 4. Em tais cenários, os método não supervisionados não são indicados, dado que eles podem erroneamente identificar objetos normais como *outliers*. (HAN; PEI; KAMBER, 2011)

2.2.1.3 Abordagem Semissupervisionada

Existem situações onde o conjunto de dados consiste em dados rotulados como normais e não rotulados. Nesta configuração, o objetivo de técnicas semissupervisionadas é identificar anomalias usando somente informações dos dados normais do conjunto de treinamento. Objetos que não se adéquam ao modelo construído somente com dados normais são classificados como *outliers*. A vantagem dessa abordagem, é que a presença de muitos dados anômalos a serem rotulados não impactam na precisão do modelo (TAN; STEINBACH; KUMAR, 2009).

Em cenários que existam somente poucos dados anômalos rotulados, o uso singular de técnicas semissupervisionadas é mais difícil. Uma vez que é improvável que esse pequeno conjunto dados anômalos tenha representatividade diante o problema em questão, inviabilizando a construção de um modelo. Por outro lado, para melhorar a precisão na DO, podemos combinar abordagens, construindo modelos de objetos normais rotulados por métodos não supervisionados. Essa também constitui-se de uma estratégia semissupervisionada. (HAN; PEI; KAMBER, 2011)

2.2.2 Métodos para Detecção de *Outliers*

Esta subseção descreve e discute os principais métodos para detecção de *outliers*: **Métodos Estatísticos**, **Métodos Baseados em Proximidade** e **Métodos baseado em Agrupamento**. Os métodos estatísticos, também conhecidos como métodos baseados em modelo, fazem inferência sobre a normalidade dos dados, assumem que os dados normais são gerados por um processo estocástico e que seguem um modelo de anormalidade. Os métodos baseados em proximidade, entendem que os dados tem uma relação de vizinhança e que um objeto é considerado anormal quando se distância consideravelmente de toda vizinhança. Os métodos baseados em agrupamento, assumem que objetos normais

pertencem a grandes grupos de dados e que dados anômalos se encontram em pequenos grupos e escassos, ou esteja fora de qualquer tipo de agrupamento. (HAN; PEI; KAMBER, 2011)(TAN; STEINBACH; KUMAR, 2009)

2.2.2.1 Métodos Estatísticos

A maior parte dos métodos estatísticos são baseados em modelos. Esse modelo é construído a partir do conjunto de dados, e representa uma distribuição de probabilidade sobre eles. Para Anscombe (1960) um candidato a *outlier* pode ser identificado quando o mesmo não foi gerado pelo modelo estocástico assumido ou construído. Uma estratégia trivial para técnicas estatísticas de DO seria analisar a probabilidade de objetos pertencerem ou se adaptarem a esse modelo. Objetos com baixa probabilidade são candidatos a *outliers*. A criação desse modelo depende também de parâmetros de distribuição fornecidos pelo o usuário. Digamos que os dados tenham uma distribuição Gaussiana, nesse caso o cálculo da média e desvio padrão dos dados representam uma análise de sumarização da distribuição, e a probabilidade de cada objeto diante a distribuição pode ser examinada (TAN; STEINBACH; KUMAR, 2009).

É importante destacar alguns pontos relevantes sobre os métodos estatísticos. Técnicas estatísticas de DO são sensíveis a quantidade de instância no conjunto de dados. Quanto maior o número de registros, maior representatividade estatística a amostra terá. Outra questão é que os modelos estatísticos normalmente são adequados para conjuntos de dados quantitativo de valor real ou, pelo menos de dados ordinais. Os dados ordinais podem ser transformados em valores numéricos se adequando ao processamento estatístico. Isso limita a aplicabilidade e aumenta o tempo de processamento, caso um pré-processamento dos dados seja necessário. (HODGE; AUSTIN, 2004)

Tan, Steinbach e Kumar (2009) descreve algumas questões importantes para serem consideradas antes de selecionar ou desenvolver algumas técnica estatística para DO. Uma delas é **identificar a distribuição**, embora a maioria dos tipos de dados possam ser representados pelas distribuições mais comuns, como a Gaussiana, de Poisson ou binominal. Existem conjuntos que seguem uma distribuição atípica. Se uma distribuição errada for considerada, a análise de dados *outlier* pode ser imprecisa. Outra questão é sobre o **número de atributos dos dados**, onde implica saber se os dados são univariáveis ou multivariáveis. Por fim, a **mistura de distribuições**, que está relacionado a casos em que os dados podem estar modelados como uma mistura de distribuições. Nesta situação, as técnicas de DO devem ser desenvolvidas baseando-se nas distribuições envolvidas. Embora tais técnicas sejam mais robustas, esses modelos são mais complicados de entender.

Em síntese, os métodos estatísticos se encaixam em um modelo estatístico, geralmente para comportamento normal dos dados, posteriormente aplicam um teste de inferência estatística para determinar se uma instância é *outlier*. No entanto, existem

diversas formas de aprender e especificar modelos generativos. Geralmente, os métodos estatísticos para a detecção DO podem ser divididos em duas categorias principais: métodos paramétricos e métodos não paramétricos (CHANDOLA; BANERJEE; KUMAR, 2009)(HAN; PEI; KAMBER, 2011). As técnicas paramétricas usam conhecimento implícito da distribuição e estimam os parâmetros dos dados dados (ESKIN, 2000), as técnicas não paramétricas geralmente não consideram tal conhecimento (DESFORGES; JACOB; COOPER, 1998).

2.2.2.1.1 Métodos Paramétricos para Dados Univariados

Método para dados univariados assumem que cada instância possui somente um atributo ou trata cada atributo de forma isolada. Pela simplicidade, vamos assumir aqui que os dados possuem uma distribuição normal e pontos com baixa probabilidade são considerados *outliers*. Formalmente, podemos entender os métodos paramétricos assumindo que os dados normais são gerados por uma distribuição paramétrica seguindo o parâmetro Θ. A função de densidade de probabilidade paramétrica $f(\mathbf{X}, \Theta)$ fornece a probabilidade da instância $\mathbf{X}$ pertença a distribuição. Quanto menor o valor da função, maior a chance de $\mathbf{X}$ ser *outlier*. (CHANDOLA; BANERJEE; KUMAR, 2009)(HAN; PEI; KAMBER, 2011)

O método *boxplot* traça os dados de entrada univariada usando um resumo de cinco atributos. O menor valor não anômalo (Min), quartil inferior ($Q1$), mediana, quartil superior ($Q3$) e maior valor não anômalo (Max). O intervalo interquartil (IQR) é definido pela subtração ($Q3$) - ($Q1$). Como exemplo, sabendo que a região entre $Q1 - 1.5 * IQR$ e $Q3 + 1.5 * IQR$ contém 99,3% das observações. Uma instância situada a mais de $1,5 * IQR$ inferior a $Q1$ ou $1,5 * IQR$ superior a $Q3$, é considerada como uma anomalia (HAN; PEI; KAMBER, 2011). A Figura 5 ilustra bem este exemplo.

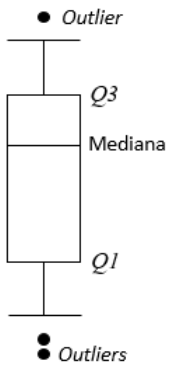


Figura 5 – Visualizando *outliers* usando *box-plot*. Adaptado de (HAN; PEI; KAMBER, 2011)

Outro método estatístico simples para a detecção de *outliers* univariados seguindo uma distribuição normal é o teste de *Grubbs* (GRUBBS, 1969). Para cada objeto X definimos um escore z, definida pela equação 2.1.

$$z = \frac{|x - \bar{x}|}{s}, \tag{2.1}$$

sabendo que $\bar{x}$ é a média, e s compreende o desvio padrão da entrada de dados, X é considerado *outlier* se:

$$z > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}, \tag{2.2}$$

onde N é o tamanho dos dados e $t^2_{\alpha/(2N),N-2}$ é o limite usado para declarar uma instância como *outlier*. Este limiar é o valor de uma distribuição-t em um nível de significância de $\alpha/(2N)$. O nível de significância reflete a confiança associada ao limiar e de certa forma tem controle no número de instâncias *outliers*.(CHANDOLA; BANERJEE; KUMAR, 2009)

2.2.2.1.2 Métodos Paramétricos para Dados Multivariados

Métodos para dados multivariados assumem que as instâncias são compostas de mais de um atributo e com alguma relação entre si. Para Han, Pei e Kamber (2011), muitos métodos de DO univariados podem ser estendidos para lidar com dados multivariados. A ideia básica é transformar a tarefa de DO multivariados em um problema de DO univariados. A distância *Mahalanobis* é uma medida de de distância que considera o formato da distribuição de dados. Dessa forma, conseguimos definir um *limiar* para separar objetos anômalos. A equação 2.3 apresenta como se calcula a distância *Mahalanobis* entre um ponto X e a média dos dados $\bar{X}$.

$$mahalanobis(X, \bar{X}) = (X - \bar{X})S^{-1}(X - \bar{X})^T, \tag{2.3}$$

onde S é uma matriz de covariância dos dados. É simples perceber que a distância *Mahalanobis* de um objeto até a média da distribuição, e está diretamente relacionada com a probabilidade do objeto. Podemos afirmar, que a distância *Mahalanobis* é igual ao *log* da densidade da probabilidade do objeto com maior frequência (TAN; STEINBACH; KUMAR, 2009). A Figura 6 ilustra como a distância *mahalanobis* é computada.

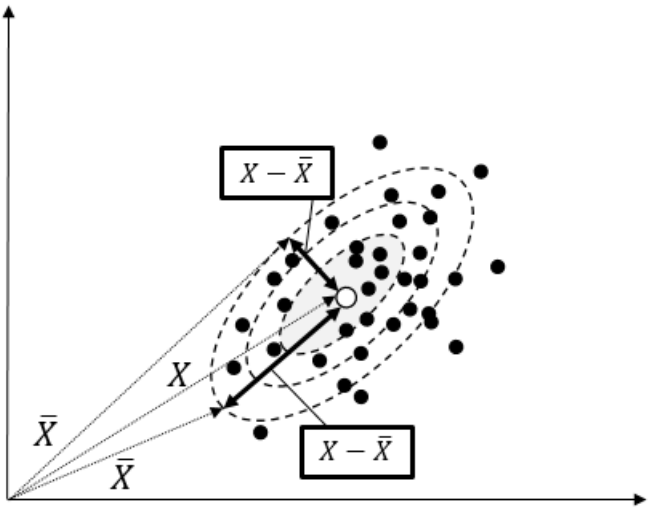


Figura 6 – Ilustração de distância *mahalanobis*. Adaptado de (DEBES; KOENIG; GROSS, 2005)

A saída da $mahanobolis(X, \bar{X})$ é uma variável univariada, assim o teste de *Grubbs* pode ser aplicado a esta medida, transformando-a numa tarefa de DO multivariado, executando os seguintes passos:

1. Calcule o vetor médio do conjunto de dados multivariados.
2. Para cada objeto X, calcule $mahanobolis(X, \bar{X})$.
3. Detectar *outliers* no conjunto de dados univariados transformados, $mahanobolis(X, \bar{X})|X \in D$.
4. Se $mahanobolis(X, \bar{X})$ indicar como um dado anômalo, então X também é considerado um *outlier*.

O método de Probabilidade de Correlação *Outlier* (COP), usa um processo estatístico para detectar *outliers* para dados multivariados. O método parte do pressuposto que os objetos de dados gerados pelo mesmo mecanismo apresentam uma correlação estatística entre alguns atributos, de forma que essa correlação forma um hiperplano no espaço de dados. A distância de *mahalanobis* é usada para identificar se o objeto em observação faz parte de algum hiperplano. Dessa forma é possível detectar dados *outliers* aos dados que estão correlacionados (KRIEGEL *et al.*, 2012). Na Figura 7 podemos observar três mecanismos, que geraram grupos de dados correlacionados, e por sua vez hiperplanos distintos. Os objetos O_1, O_2 e O_3 não estão correlacionados com nenhum mecanismo, por isso são considerados *outliers*.

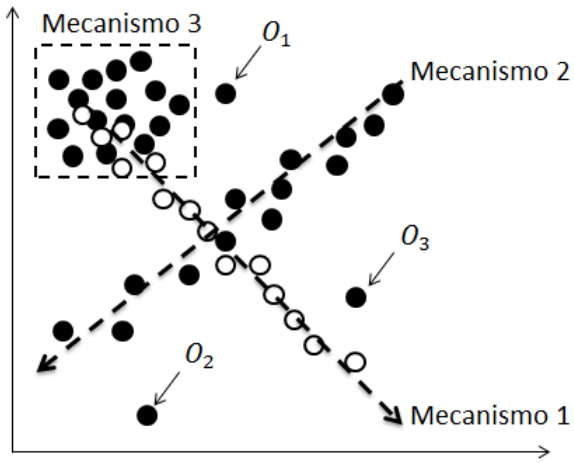


Figura 7 – Demonstração visual do método de Probabilidade de Correlação *Outlier* (COP). Adaptado de (KRIEGEL *et al.*, 2012)

2.2.2.1.3 Métodos Paramétricos para Múltiplas Distribuições

Existem situações em que conjunto de dados são gerados por distribuições diferentes. Considere um caso em que existam dois *clusters* $C1$ e $C2$. Nesta situação, assumir que os dados são gerados somente por uma distribuição normal talvez não seja uma estratégia válida. Isso porque a média estimada está situada entre os dois *clusters*, e os objetos que estão entre esses *clusters* estão próximos da média, fazendo com que *outliers* nesta situação não possam ser classificados como tal. Para solucionar este problema, podemos assumir que os dados são gerados por múltiplas distribuições. Para exemplificar, assumimos duas distribuições dada por $\Theta_1(\mu_1, \sigma_1)$ e $\Theta_2(\mu_2, \sigma_2)$. Para qualquer instância o do conjunto de dados, a probabilidade dela ter sido gerada por uma mistura de modelos é dada por:

$$Pr(\mathbf{o}|\Theta_1, \Theta_2) = f_{\Theta_1}(\mathbf{o}) + f_{\Theta_2}(\mathbf{o}), \tag{2.4}$$

onde f_{Θ_1} e f_{Θ_2} são funções de probabilidade para Θ_1 e Θ_2 respectivamente. Para detectar *outliers*, podemos usar o algoritmo *expectation-maximization* (EM), para a aprender os parâmetros $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$ do conjunto de dados. Cada *cluster* é representado por uma distribuição normal e um objeto o é dado como um *outlier* caso sua probabilidade dada pela combinação das duas distribuições, for baixa.

O algoritmo funciona da seguinte forma. Primeiramente é atribuído aleatoriamente valores iniciais para os parâmetros Θ. Posteriormente os passos *expectation* (E) e *maximization* (M) são iterativamente executados até os parâmetros convergirem. No passo E, para cada objeto $o_i \in O(1 \leq i \leq n)$ é calculado a probabilidade de o_i pertencer a cada

distribuição, dada a equação 2.5.

$$Pr(\Theta_j|o_i,\Theta) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_l)} \tag{2.5}$$

No passo M, os parâmetros Θ são ajustados de forma que $P(O|\Theta)$ seja maximizada, de acordo com a seguinte configuração:

$$\mu_j = \frac{1}{k}\sum_{i=1}^{n} o_i \frac{P(\Theta_j|o_i,\Theta)}{\sum_{l=1}^{n} P(\Theta_j|o_l,\Theta)} = \frac{1}{k}\frac{\sum_{l=1}^{n} o_i P(\Theta_j|o_l,\Theta)}{\sum_{l=1}^{n} P(\Theta_j|o_l,\Theta)} \tag{2.6}$$

e

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} P(\Theta_j|o_i,\Theta)(o_i-\mu_j)^2}{\sum_{i=1}^{n} P(\Theta_j|o_i,\Theta)}} \tag{2.7}$$

Pela simplicidade do algoritmo *Expectation-Maximization* (EM), ele é comumente usado para lidar com muitos problemas de DO. Devido aos valores iniciais serem gerados de forma aleatória, o EM pode não convergir para a solução ideal, mas para um máximo local. Algumas heurísticas foram desenvolvidas para evitar isso, executando o processo EM diversas vezes, usando diferentes valores iniciais aleatórios. Além disso, o algoritmo EM pode ser muito custoso para grandes distribuições.(HAN; PEI; KAMBER, 2011)

2.2.2.1.4 Métodos não Paramétricos

As técnicas de DO não paramétricas assumem que são os dados determinam a estrutura do modelo, e não especificada antecipadamente. Isso não quer dizer que o modelo não tenha parâmetros, seria impossível o aprendizado do modelo somente através dos dados. Os métodos não paramétricos assumem que o número e a natureza dos parâmetros são flexíveis e previamente estabelecido. Histogramas e a estimativa de densidade do *kernel*, são exemplos de métodos não paramétricos.(CHANDOLA; BANERJEE; KUMAR, 2009)(HAN; PEI; KAMBER, 2011)

A DO baseada em histogramas para dados univariados compreende em duas etapas. Primeiramente o histograma é construído com base em em valores obtidos por por diferentes recursos no conjunto de treinamento. Posteriormente, a técnica verifica se uma determinada instância de teste se adapta a alguma das caixas do histograma. Caso positivo, a instância de teste é normal, caso contrário, é anômala. Outra forma seria atribuir uma pontuação *outlier* a cada instância de teste baseando-se na frequência em que ela se adapta as caixas. O tamanho dessas caixas na construção de um histograma é o ponto chave para a DO. Se as caixas forem pequenas, muitas instâncias de teste normais cairão em caixas vazias ou raras, resultando em uma alta taxa de falsos positivos. Da mesma forma, se as caixas forem grandes, muitas caixas serão frequentemente utilizadas por instâncias de testes

anômalas. Sendo assim, o maior desafio técnicas baseadas em histograma é determinar um tamanho ótimo das caixas para construir o histograma. (CHANDOLA; BANERJEE; KUMAR, 2009)

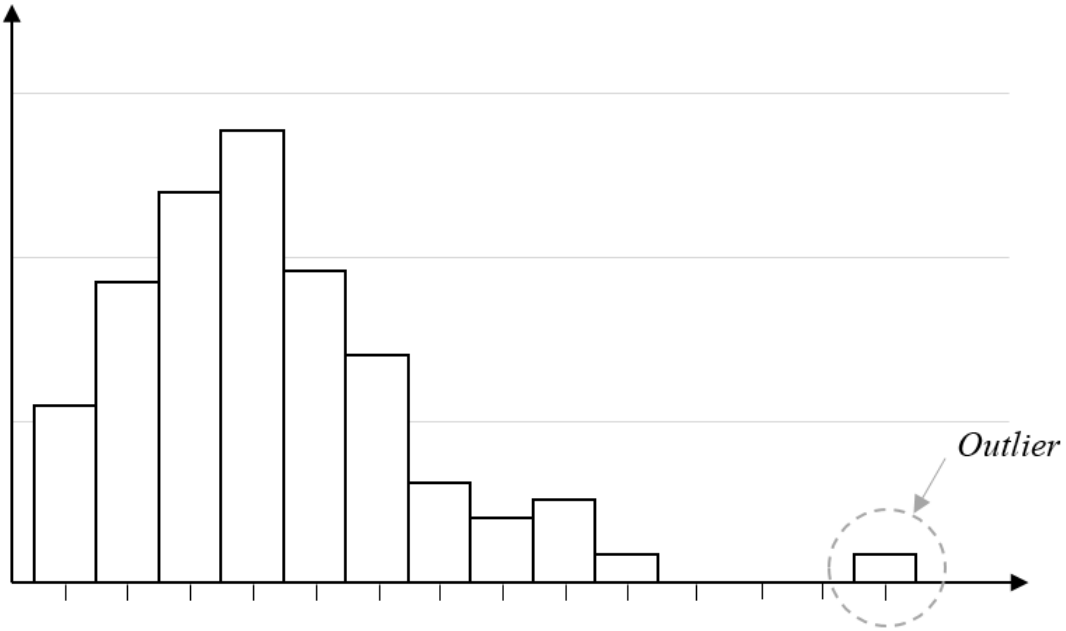


Figura 8 – Ilustração de um histograma com dados *outliers*

Uma maneira de superar esse desafio, é adotando uma estimativa da densidade do *kernel* para prever a distribuição da densidade de probabilidade dos dados. Tratamos um objeto observado como um indicador de densidade com alta probabilidade na região periférica. A densidade de probabilidade para um determinado objeto, depende das distâncias desse objeto para com os outros. Assim, é possível usar uma função de *kernel* para modelar a influência que um determinado objeto tem com sua vizinhança. Um *kernel* $K()$ é uma função de integral, para valores reais não negativos, que satisfaz as seguintes duas condições:

- $\int_{-\infty}^{+\infty} K(u)du = 1$;
- $K(-u) = K(u)$ *para todos os valores de* u.

Normalmente o kernel usado é uma função Gaussiana com média variando de 0 a 1:

$$(\frac{x - x_i}{h}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_j)^2}{2h^2}} \tag{2.8}$$

Seja $X_1, ..., X_n$ uma amostra de dados independente distribuída de forma aleatória pela variável f. A aproximação de densidade do *kernel* é dada pela função de densidade

de probabilidade 2.9:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{2.9}$$

Onde $K()$ é um *kernel* e h é a largura de banda, que serve como um parâmetros de suavização.

Uma vez que é feita aproximação usando uma estimativa de densidade do *kernel*, podemos usar a função de densidade de probabilidade para a DO, semelhante aos métodos paramétricos. Se $\widehat{f}(x)$ for alto, trata-se de um dado normal, caso contrário é um *outlier*. (HAN; PEI; KAMBER, 2011)

Em resumo, métodos estatísticos para DO aprendem modelos partindo do conjunto de dados para distinguir objetos normais de *outliers*. Uma das principais vantagens dos método estatísticos é o fato de que se as proposições relativas à distribuição de dados forem verdadeiras, as técnicas estatísticas constituem uma estratégia eficiente para a DO. Outra vantagem é que se a estimativa de distribuição for robusto para dados anômalos, as técnicas estatísticas podem operar com abordagem não supervisionada. Por outro lado, a eficácia desses métodos é fortemente dependente das premissas relativas as distribuições dos dados. Outro ponto negativo é que as técnicas baseadas em histogramas não são eficazes para dados multivariados, dado que elas não conseguem capturar os relacionamentos entre os diferentes atributos. Anomalias podem ter atributos com valores individualmente mais frequentes, mas quando combinados são raros. Assim, técnicas baseadas em histogramas não são capazes de detectar esse tipo de *outlier*.(HAN; PEI; KAMBER, 2011)(CHANDOLA; BANERJEE; KUMAR, 2009)

2.2.2.2 Métodos baseados em Proximidade

Os métodos de DO baseados em proximidade, identificam pontos anômalos observando sua distância com o restante dos dados. Intuitivamente, objetos com determinada distância dos outros podem ser considerados como *outliers*. Os métodos de DO baseados em proximidade se dividem em dois: Métodos baseado em distância e em densidade. Os baseados em distância observam os vizinhos mais próximos em um determinado raio. Um objeto é considerado anômalo caso sua vizinhança tiver um número baixo de instâncias. O métodos baseados em densidade, observam a densidade de um objeto em relação ao seus vizinhos. Objetos anômalos tem uma densidade menor se comparado com sua vizinhança.(TAN; STEINBACH; KUMAR, 2009)(HAN; PEI; KAMBER, 2011)

2.2.2.2.1 Métodos baseados em Distância

Técnicas baseadas em distância são aquelas que computam a distância de um objeto em relação aos k-vizinhos mais próximos. Tan, Steinbach e Kumar (2009) define a

detecção de objetos anômalos baseados em distância como um grau de elemento estranho, dado pela distância do seu vizinho k mais próximo. Outra definição de DO baseado em distância é feita por Knorr e Ng (1997), onde descreve que *outliers* baseados em distância é um objeto que está a uma distância maior que K em $p\%$ dos demais objetos. Han, Pei e Kamber (2011) formula esse método da seguinte forma: Para um conjunto D de dados, o usuário pode especificar um limiar k que define o que é vizinhança para um determinado objeto. Para cada objeto o é calculado o número de objetos na k-vizinhança de o. Caso a maioria dos objetos em D estiverem consideravelmente distantes de o, então o é um candidato a *outlier*. Dessa maneira, podemos definir formalmente assumindo $k(k \leq 0)$ e $\pi(0 < \pi \leq 1)$, um objeto o é um *outlier* baseado em distância ($DB(k, \pi)$-*outlier*) se:

$$\frac{\|o'|dist(o, o') \leq k\|}{\|D\|} \leq \pi \tag{2.10}$$

onde $dist()$ é a medida de distãncia.

Uma estratégia para computar $DB(k, \pi)$-*outlier*, é expressa pelo algoritmo 2.2.2.2.1, onde usa laços aninhados para verificar a k-vizinhança de cada objeto. Para cada objeto $o_i(1 \leq i \leq n)$ é computado a distância e contado os objetos que que são comparados com o_i.

Algoritmo 1: Detecção de *outlier* baseado em distância, adaptado de (HAN; PEI; KAMBER, 2011)

Entrada : $D = o_1, ..., o_n$, k e π
Saída : $DB(k, \pi)$ *outliers* em D
início
  para $i = 1$ *até* n **faça**
    $count \leftarrow 0$
    para $j = 1$ *até* n **faça**
      se $i \neq j$ ***e*** $dist(o_i, o_j) \leq k$ **então**
        $count \leftarrow count + 1$
        se $count \geq \pi \cdot n$ **então**
          o_i não é um *outlier*
        fim
      fim
    fim
    o_i é um $DB(k, \pi)$-**outlier** de acordo com a equação 2.10
  fim
fim

De acordo com o **algoritmo 1**, podemos perceber que os laços aninhados levam o algoritmo a uma complexidade de $O(n^2)$. Apesar disso o desempenho do método muitas vezes se comporta linearmente, dado que para a maioria dos dados anômalos e na maioria dos casos, o laço interno termina no início quando a quantidade de *outliers* no conjunto de

dados é pequena. A abordagem baseada em distância é muito sensível aos parâmetros de entrada, e não podem lidar com regiões de densidades muito diferentes, uma vez que elas usam limiares globais que não levam em considerações tais variações de densidade. (TAN; STEINBACH; KUMAR, 2009)

2.2.2.2.2 Métodos baseados em Densidade

Técnicas de DO baseadas em densidade tem o objetivo de estimar a densidade da vizinhança para cada instância de dados. Uma instância situada em baixa densidade é identificada como anômala, enquanto uma instância que se encontra em uma vizinhança densa é declarada como normal (CHANDOLA; BANERJEE; KUMAR, 2009). Esse tipo de DO, está altamente relacionado com o fator de proximidade entre as instâncias. Um *outlier* baseado em densidade é identificado de acordo com o grau de anomalia referente ao inverso da densidade em torno dele, como uma distância inversa.(TAN; STEINBACH; KUMAR, 2009)

Para entender melhor as técnicas de DO baseadas em densidade e perceber as limitações que as técnicas baseadas em distâncias sofrem, considere dois *clusters*, C_1 é um agrupamento denso e C_2 é esparso. O objeto o_3 está longe da maioria dos dados, este pode ser classificado como um *outlier* por método baseados em distância. Os objetos o_1 e o_2 numa ótica global, estão mais próximos ao *cluster* C_1, que é mais denso. Assim, os objetos o_1 e o_2 não são *outliers* baseado em distância. De certa forma, se os objetos o_1 e o_2 fossem classificados como *outliers* $DB(k, \pi)$-*outlier*, todos os objetos de C_2 também seriam categorizados da mesma forma. No entanto, numa ótica local de C_1, os objetos o_1 e o_2 se desviam consideravelmente e também estão distantes de C_2. (BREUNIG *et al.*, 2000)(CHANDOLA; BANERJEE; KUMAR, 2009)

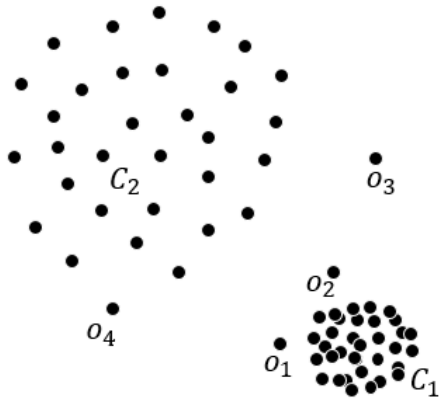


Figura 9 – Detecção de *outliers* em grupos de diferentes densidades. Adaptado de (BREUNIG *et al.*, 2000)

A Figura 9 ilustra bem essa situação. Os métodos baseados em distância não conseguem identificar *outliers* locais. Podemos perceber também que o objeto o_4 tem uma distância maior em relação a sua vizinhança do que o objeto o_1 em relação a sua vizinhança. Contudo, o objeto o_4 é um dado local pertencente a C_2, não sendo considerado um dado anômalo. Para solucionar este impasse, é introduzido a noção e densidade relativa que indica o grau em que um objeto é um *outlier* em situações que existem diferentes grupos com diferentes estados de densidades (HAN; PEI; KAMBER, 2011). O *Fator Outlier Local* (LOF) (BREUNIG *et al.*, 2000) é uma técnica de DO que atribui para cada objeto um fator de anormalidade local, resultado da diferença da densidade do objeto com sua vizinhança.

Antes de definir a formulação do LOF, primeiramente precisamos definir algumas premissas. Tome D como o conjunto de dados e a k-distância denotada por $dist_k(o)$, que trata-se da distância de o e seu vizinho mais próximo de k. Consequentemente, a vizinhança de k contém todos os objetos cuja distância para o não é maior que $dist_k(o)$, assim a distância de k a o é denotada por:

$$N_k(o) = o'|o' \in D, dist(o, o') \leq dist_k(o) \tag{2.11}$$

$N_k(o)$ serve como medida de densidade relativa para o. Porém, o pode ter vizinhos tão próximos de forma que $dist(o, o')$ tenha um valor pequeno, gerando flutuação estatística. Para contornar isso, é preciso adicionar uma suavização a formulação, dada pela equação 2.12. Em sequência definimos o fator de acessibilidade local (LRD), que compara a densidade local com a de seus vizinhos, a fim de quantificar o grau em que o objeto é considerado um *outlier*. Por fim, a formulação do LOF dada pela equação 2.14. (BREUNIG *et al.*, 2000)

$$suavedist_k(o \leftarrow o') = max\{dist_k(o), dist(o, o')\} \tag{2.12}$$

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o'_{\in N_k(o)}} suavedist_k(o \leftarrow o')} \tag{2.13}$$

$$LOF_k(o) = \frac{\sum_{o'_{\in N_k(o)}} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} suavedist_k(o' \leftarrow o) \tag{2.14}$$

Em síntese, o LOF é a média da razão entre a LRD de o e os de k vizinhos mais próximos. Quanto menor a LRD de o maiores as LRD's dos k vizinhos mais próximos de o, consequentemente maior o valor de LOF. Isso identifica exatamente um *outlier* local, no qual sua densidade é relativamente baixa quando comparada com as densidades locais de seus vizinhos (HAN; PEI; KAMBER, 2011). A Figura 10 ilustra como as propriedades do LOF se comportam na DO. Para entende-las melhor, estão descritas nas equações a seguir.

- $direto_{min}(o) = min\{suavedist_k(o' \leftarrow o | o' \in N_k(o)\}$
- $direto_{max}(o) = max\{suavedist_k(o' \leftarrow o | o' \in N_k(o)\}$
- $indireto_{min}(o) = min\{suavedist_k(o'' \leftarrow O' | o' \in N_k(o) \text{ e } o'' \in N_k(o')\}$
- $indireto_{max}(o) = max\{suavedist_k(o'' \leftarrow O' | o' \in N_k(o) \text{ e } o'' \in N_k(o')\}$

Assim, o LOF é delimitado da seguinte forma.

$$\frac{diretriz_{min}(o)}{indireto_{max}(o)} \leq LOF(o) \leq \frac{diretriz_{max}(o)}{indireto_{min}(o)} \tag{2.15}$$

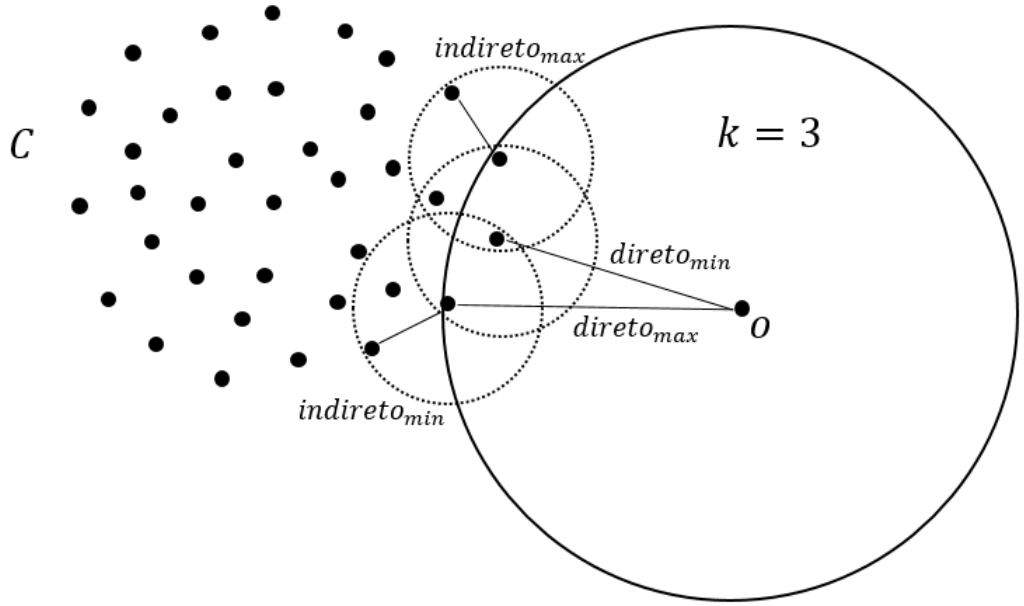


Figura 10 – Funcionamento das propriedades do LOF. Adaptado de (BREUNIG *et al.*, 2000)

O LOF inspirou o desenvolvimento de diversos métodos. O ALOCI (PAPADIMITRIOU *et al.*, 2003), COF (TANG *et al.*, 2002), INFLO (JIN *et al.*, 2006), KDEOS (SCHUBERT; ZIMEK; KRIEGEL, 2014), LDOF (ZHANG; HUTTER; JIN, 2009) e LoOP (KRIEGEL *et al.*, 2009) são exemplos de técnicas que derivaram-se do LOF, utilizando outras heurísticas, otimizando a performance, combinando outras abordagens para obter maior eficácia na detecção de *outliers*.

Os método de DO baseados em densidade, configuram-se como estratégias para atribuir um grau de anormalidade para cada objeto, quando este se encontra em regiões de diferentes densidades. Seguindo o mesmo desempenho computacional que os métodos a distância, sua complexidade assintótica está em $O(n^2)$. Embora possa ter redução para

$O(nlogn)$ para pequenos conjuntos de dados, sua eficiência está fortemente atrelada a dimensão dos dados. A seleção de parâmetros é um desafio para esses métodos. No entanto, a técnica LOF aborda a diversidade de valores para k e consegue determinar os graus de anormalidade mais máximos. (BREUNIG *et al.*, 2000)(TAN; STEINBACH; KUMAR, 2009)

2.2.2.3 Métodos baseados em Agrupamento

O agrupamento de dados encontra *clusters* com objetos fortemente relacionados, enquanto que na DO anomalias são identificadas quando estão altamente relacionadas a outros objetos, fazendo com que pertençam a nenhum grupo ou a grupos muito pequenos (TAN; STEINBACH; KUMAR, 2009). Normalmente técnicas de agrupamento de dados são não supervisionada, embora existam abordagens semissupervisionadas para o problema (BASU; BILENKO; MOONEY, 2004). Para o problema de DO elas se dividem em três categorias.

A primeira categoria assume que os dados normais pertencem a somente um cluster, enquanto os *outliers* não pertencem a *cluster* algum. Técnicas que se enquadram nesta categoria aplicam técnicas de agrupamento no conjunto de dados e instâncias que não pertencem a nenhum grupo são consideradas anomalias. A desvantagens dessas técnicas é que elas não são otimizadas para encontrar *outliers* e sim *cluster* de dados. A segunda categoria assume que os dados normal são adjacentes ao centro do *cluster* mais próximo. As técnicas que assumem tal suposição consistem em duas fases: primeiramente os dados são agrupados aplicando um algoritmo de agrupamento, posteriormente a distância de cada objeto até o seu centroide mais próximo é usado para computar sua pontuação de anormalidade. A terceira categoria compreende anomalias como objetos que pertencem a grupos pequenos e esparsos, enquanto dados normais pertencem a grupos grandes e densos. (CHANDOLA; BANERJEE; KUMAR, 2009)

As técnicas baseadas em agrupamento precisam de medidas de distância para calcular similaridade entre pares de objetos. Tanto para técnicas baseadas em proximidade como as baseadas em agrupamento, a medida de distância é fundamental para o desempenho. Apesar dos dois métodos se assemelharem neste aspecto, a principal diferença entre os dois é que as técnicas baseadas em agrupamento avaliam cada objeto em relação ao *cluster* o qual pertence, enquanto as baseadas em proximidade, avaliam cada objeto em relação a sua vizinhança local.

Uma das técnicas mais exploradas para o agrupamento de dados é o *K-means* (STEINBACH *et al.*, 2003). Trate-se de um agrupamento baseado em centroides que cria uma particionamento a nível dos objetos dos dados. Primeiramente de forma aleatória, é escolhido os K centroides iniciais, onde K é um parâmetro de entrada, especificado pelo usuário. Cada instância é atribuída a um centroide mais próximo, e o conjunto de instâncias

atribuídas a um mesmo centroide formam um grupo. Nesse contexto, o *K-means* pode identificar *outlier* computando um grau de anormalidade para cada instância, podendo ser realizada de duas formas. Uma maneira seria pela distância do objeto até o seu centroide mais próximo, e outra forma seria pela distância relativa do seu centroide mais próximo. Essa distância relativa compreende no cálculo da distância do ponto até seu centroide mais próximo, com a distância média de todos os pontos do mesmo centroide. O **algoritmo 2** apresenta de forma genérica como seria o funcionamento do *K-means* para DO (TAN; STEINBACH; KUMAR, 2009). Considere D como conjunto de dados, K como parâmetro de quantidade de centroides e L conjunto de *outliers*.

Algoritmo 2: Detecção de *outlier* baseado em *K-means*, adaptado de (TAN; STEINBACH; KUMAR, 2009)

Entrada : D conjunto de dados, K quantidade de grupos
Saída : L conjunto de *outliers* em D
início
  Selecione K pontos como centroides iniciais
  repita
    Calcule a distância de cada ponto em D atribuindo-os ao K grupos
    Calcule o grau de anormalidade para cada ponto em D
    Atualize L com os pontos com alto grau de anormalidade
  até *não haver mudanças significativas nos K pontos*;
fim

Diante diversos métodos para DO, elas se destingem em desempenho para determinados domínios, tipos de *outliers* e natureza dos dados. Logo, questões de como quantificar e qualificar uma comparação de técnicas de DO para determinados contextos, são de extrema importância e de grande contribuição científica. Mensurar o desempenho de tantas técnicas para diferentes contextos, de forma parametrizada e uniforme, não é uma estratégia tão simples de elucidar e a literatura não apresenta métodos ou metodologias para que questões como esta possam ser respondidas. A classificação é uma estratégia que pode ser adaptada como uma abordagem para comparar, diversas técnicas de DO.

2.3 Classificação

Extrair modelos de conjunto de dados com objetivo de descrever classes ou categorias, é uma forma de analisar dados denominada de classificação. Para tais modelos é atribuído a função de classificador, no qual prediz rótulos de classes. Em outras palavras, o problema classificação compreende em construir modelos de classificação baseando-se em conjunto de dados, para categorizar objetos ainda não conhecidos (HAN; PEI; KAMBER, 2011). Tan, Steinbach e Kumar (2009) define o conceito de classificação como uma tarefa de

aprender uma função objetivo f, que mapeie todos os conjuntos de atributos x para classes pré-determinadas y. Nesse caso a função alvo corresponde ao modelo de classificação.

Figura 11 – Modelo de Classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009)

De acordo com a Figura 11, podemos entender um modelo de classificação como uma caixa preta capaz de atribuir um rótulo para uma classe, recebendo como entrada o conjunto de atributos de um determinado objeto ainda não observado. Esta seção é dedicada a discussão dos principais conceitos sobre classificação, enfatizando a apresentação de estratégias de avaliação e comparação de classificadores, bem como discutir as medidas de precisão dos modelos e como obter estimativas confiáveis de precisão.

2.3.1 Abordagem Geral para Classificação

Um método ou técnica de classificação compreende numa abordagem sistemática na construção de modelos preditivos, que tem como entrada um conjunto de dados. Cada técnica emprega um algoritmo de aprendizado, a fim de estabelecer o melhor modelo para o relacionamento do conjunto de atributos com os rótulos dos dados de entrada (TAN; STEINBACH; KUMAR, 2009). O ***bias*** ou viés indutivo do algoritmo é responsável por essa melhor adequação, discutido na próxima subseção. Deste modo, o principal objetivo dos métodos de classificação é adaptar-se bem ao dados de entrada e prever precisamente rótulos de dados ainda não conhecidos, para isso é necessário que o algoritmo construa modelos com alto grau de generalização. (ALPAYDIN, 2014)

O processo de classificação de dados normalmente constitui de duas etapas, a primeira corresponde a uma etapa de aprendizado, onde o modelo é construído analisando o conjunto de dados, composto por tuplas ou exemplos de banco de dados, seguidos de seus respectivos rótulos. Um exemplo X corresponde a um vetor de atributos de dimensão n, onde $X = (x_1, x_2, ..., x_n)$ representa n medições feitas em cada exemplo por n atributos $A = (a_1, a_2, ..., a_n)$. Cada exemplo X pertence a uma categoria pré-definida, representada pelo atributo classe C, com valor discreto e não ordenado. O tipo desse atributo é categórico ou nominal, e representa juntamente uma categoria ou classe no qual o exemplo X pertence. Normalmente, esse processo é denominado de **indução**. (HAN; PEI; KAMBER, 2011)

A segunda etapa corresponde a classificação, onde o modelo gerado na primeira etapa é usado para prever rótulos de classes para os dados. Para avaliar o desempenho de

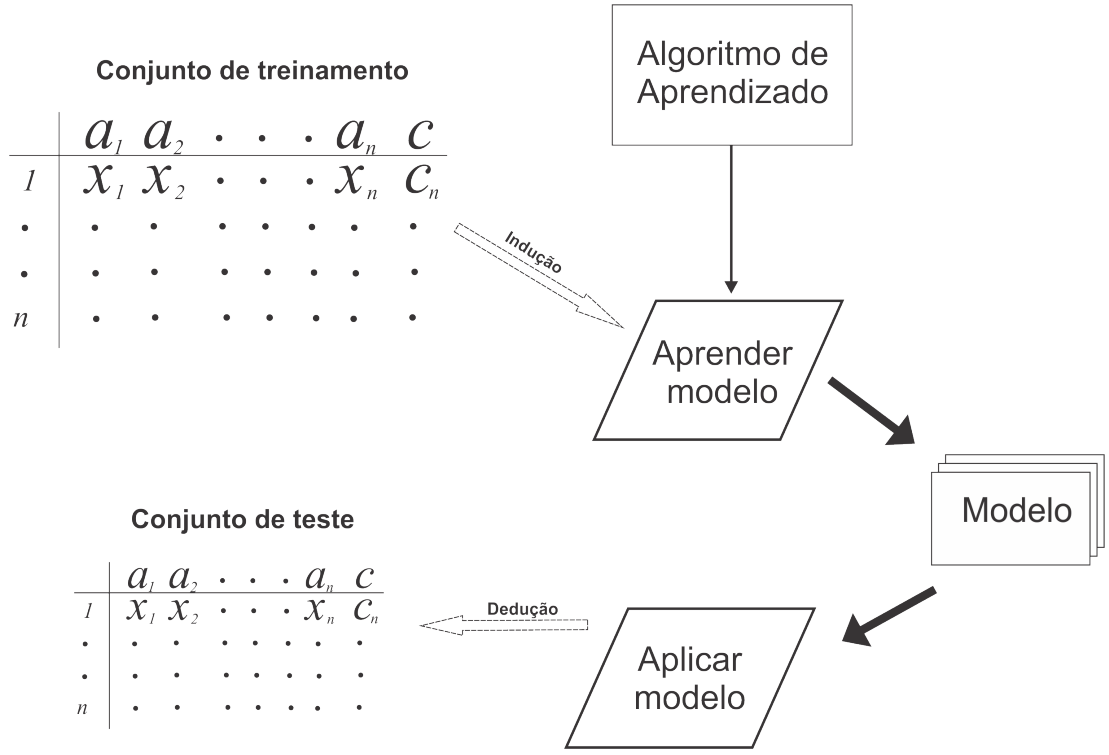


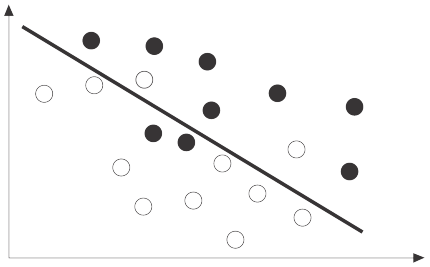
Figura 12 – Abordagem geral para a construção de um modelo de classificação. Adaptado de (TAN; STEINBACH; KUMAR, 2009)

um classificador gerado, um conjunto de teste é usado. Também composto por exemplos com rótulos associados, eles são independentes dos exemplos de treinamento, ou seja, não foram usadas na construção do classificador. Um classificador é avaliado de acordo com a porcentagem de exemplos no conjunto de testes classificados corretamente. O rótulo previsto pelo classificador é comparado com o rótulo do exemplo de teste, gerando uma **precisão** ou **acurácia** do classificador. Essa segunda etapa é normalmente denominada de **dedução** (HAN; PEI; KAMBER, 2011). Esse método de construção do modelo de classificação é conhecido como *holdout*, discutido posteriormente. A Figura 12 ilustra uma abordagem genérica de como é construído um modelo de classificação.

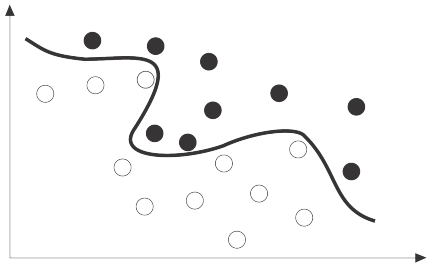
2.3.2 Bias e Variação

Graficamente, podemos entender um modelo de classificação como uma função que projeta uma linha simples (reta) ou complexa (curva), separando os dados de testes. A linha mais simples tem baixa complexidade, porém sofre com alguns dados localizados no lado errado da separação (**erros de classificação**), ilustrado na Figura (a). A Figura (b) ilustra um melhor ajuste da linha, eliminando o erro de separação, no entanto o custo função de separação é maior, projetando uma linha mais complexa. Agora suponhamos que mais dados de testes são inseridos, a separação simples não precisa realizar mudança alguma para acomodar os novos dados, ou seja, tem baixa variação. Por outro lado, o

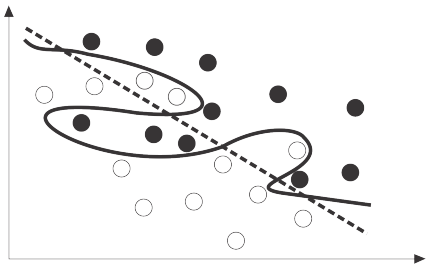
separador mais complexo precisa realizar alterações consideráveis, caso pretenda preservar a taxa de erro original, ou seja, tem alta variação. Essa situação é ilustrada na Figura (c)(LAROSE, 2014).



(a) Separador simples com alta taxa de erros.



(b) Separador complexo sem taxa de erros.



(c) Com mais dados inseridos. Separador não tem variação, mas separador complexo com alta variação.

Figura 13 – Exemplo de Bia e Variação. Adaptado de (LAROSE, 2014)

Apesar do modelo com maior complexidade ter uma *bias* baixo, no que tange ao erro de classificação, ele tem uma alta variância. E embora que o modelo de menor complexidade tenha um *bias* alto, possui uma baixa variância. Isso é conhecido como compensação *bias-variância*. O ideal é que o modelo seja construído de forma que o bias e a variância não sejam muito altos, no entanto, geralmente quando se tenta minimizar um, o outro tende a aumentar. Essa compensação de *bias-variância* também é uma maneira de entender o ***overfitting*** e ***underfitting*** exemplificado na Figura 14. Ela mostra que à medida em que o modelo aumenta a complexidade, as taxas de erro para o conjunto de treinamento e no conjunto de teste caem. No entanto, à medida que a complexidade do modelo vai aumentando, a taxa de erro para o conjunto de teste começa a se achatar e aumentar. Isso porque o modelo memorizou o conjunto de treinamento de tal maneira que não deixou espaço para a generalização para dados desconhecidos.(LAROSE, 2014)

O *overfitting* é ocasionado quando o modelo tenta explicar todas as tendências e estruturas dos dados de treinamento. Aumentar a complexidade do modelo com objetivo de melhorar a precisão para o conjunto de treinamento, inevitavelmente tem efeito degradante na generalização do modelo. Já o *underfitting* está no outro extremo, o modelo representa tendências insuficientes dos dados treinamento, de tal forma que apresenta um *bias* muito alto. (LAROSE, 2014)

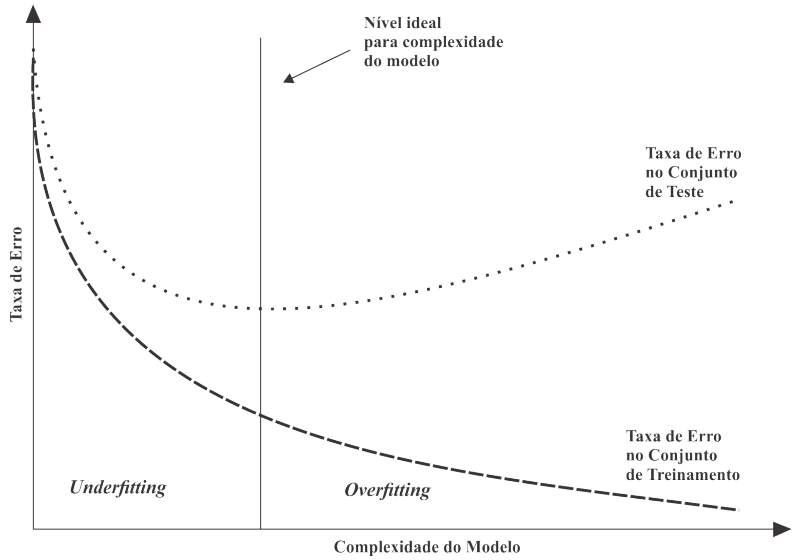


Figura 14 – O nível ideal de complexidade do modelo corresponde a menor taxa de erro no conjunto de testes. Adaptado de (LAROSE, 2014)

O método mais comum para avaliar a precisão de modelos é o **erro médio quadrático**(MSE). O modelo com menor MSE é considerado melhor. O MSE combina o *bias* e variância para calcular a medida avaliativa. Trata-se de uma função do erro de previsão com uma medida de complexidade do modelo, como por exemplo o grau de liberdade. Essa função pode ser representada pela equação 2.16, onde mostra claramente a relação complementar do *bias* com a variância. A próxima seção é dedica a apresentação de métricas e métodos para avaliar modelos de classificação. (HAND; SMYTH, 2001)

$$MSE = variância + bias^2 \tag{2.16}$$

2.3.3 Métricas de Avaliação

Uma avaliação de desempenho para um modelo de classificação, baseia-se no cálculo de registros do conjunto de testes, previstos corretamente e incorretamente. O objetivo dessa avaliação é comparar diferentes modelos de classificação com intuito de selecionar o mais adequado. Existem diversas medidas e métodos que servem de parâmetros e fatores decisórios para uma melhor seleção de modelos. Posto isso, algumas questões são levantadas e discutidas nesta seção, como a definição e como calcular a acurácia de um modelo; Qual medida é mais apropriada e como calcular a acurácia com maior confiança. (TAN; STEINBACH; KUMAR, 2009)(HAN; PEI; KAMBER, 2011)

Na seção 2.3.3.1 é discutida métricas de avaliação para a acurácia de um modelo preditivo. *Holdout* e outros métodos para estimar a precisão na seção 2.3.4. E na seção 2.3.5 é discutido como aplicar testes de significância estatística, para avaliar a precisão de dois ou mais modelos.

2.3.3.1 Avaliando a Performance de um Modelo

No problema de classificação, medir o nível de precisão do modelo faz parte da segunda etapa do processo. Depois que o modelo foi construído é importante mensurar o quão "preciso"o modelo está quando aplicado ao conjunto de testes. A principal métrica de avaliação do classificador é a acurácia, no entanto na a literatura descreve outras, como taxa de reconhecimento, sensibilidade, especificidade, precisão, F_1 e F_β. Medir a acurácia de um modelo envolve aplica-lo a um conjunto de testes, que por sua vez é formado por exemplos que não foram usados no conjunto de treinamento. Estimar a precisão do modelo com dados do conjunto de treinamento, pode resultar em estimativas equivocadas, devido à superespecialização do algoritmo de aprendizado (LAROSE, 2014). Aqui dedicaremos a discussão das métricas de acurácia e taxa de erro.

Antes de discutir as métricas de avaliação, é necessário definir algumas terminologias para o melhor entendimento. Exemplo positivo, trata-se da classe principal no qual é esperado que o classificador estime, e os exemplos negativos são as outras classes. Segundo Han, Pei e Kamber (2011), existem mais quatro termos que estão envolvidos no calculo das métricas, e são descritas da seguinte forma:

- **Verdadeiro Positivo(VP)**: Refere-se aos exemplos positivos corretamente rotuladas pelo classificador;
- **Verdadeiro Negativo(VN)**: Compreende os exemplos negativos corretamente rotuladas pelo classificador;
- **Falso Positivo(FP)**: Está relacionado aos exemplos negativos incorretamente rotuladas como positivas pelo classificador;
- **Falso Negativo(FN)**: Refere-se aos exemplos positivos incorretamente rotuladas como negativas pelo classificador;

Na tabela 1, a matriz de confusão sumariza esses termos em um problema de classificação binária, e é útil para realizar uma primeira análise de desempenho do modelo. VP e VN nos dizem se um classificador obteve um bom desempenho, enquanto FP e FN nos dizem o contrário. Um bom resultado para um modelo de classificação compreende em valores de FP e FN próximos de zero. Temos ainda colunas e linhas que correspondem ao total de instâncias classificadas, representadas por P e N. P' é o total de exemplos rotulados como positivos e N' como negativos. Uma matriz de confusão depende do número de classes envolvidas no problema, onde elas podem ser representadas por $C = (c_1, c_2, ..., c_n)$. Dado m o número de classes, uma matriz de confusão é uma tabela de dimensão $m \times m$. (HAN; PEI; KAMBER, 2011)

		Classe Estimada		Total
		C_1	C_2	
Classe Real	C_1	**VP**	**FN**	**P**
	C_2	**FP**	**VN**	**N**
Total		**P'**	**N'**	**P + N**

Tabela 1 – Tabela de confusão para um modelo de classificação binário. Adpatado de (TAN; STEINBACH; KUMAR, 2009)

A **medida de acurácia** da pela equação 2.17, corresponde a taxa de exemplo rotulados corretamente pelo classificador. Pode ser referido também como taxa de reconhecimento do classificador, no qual reflete a capacidade que o modelo tem de reconhecer classes diferentes. (TAN; STEINBACH; KUMAR, 2009)

$$acuracia = \frac{VP + VN}{P + N} \tag{2.17}$$

$$taxa\ de\ erro = \frac{FP + FN}{P + N} \tag{2.18}$$

2.3.4 *Holdout* e Outros Métodos para Estimar a Precisão

Até agora, somente o método ***holdout*** foi considerado na discussão sobre precisão de modelos de classificação. O *holdout* considera dois conjuntos independentes, um conjunto de treinamento e um conjunto de testes, com dados distribuídos aleatoriamente. Geralmente, dois terços dos dados são alocados para o conjunto de treinamento e um terço para o conjunto de testes. O conjunto de treinamento é usado para construir o modelo. A precisão do modelo é então estimada com base em medidas discutidas na seção anterior. No entanto, a estimativa pode ser irresoluta, porque todo o conjunto não foi usado para derivar o modelo. (HAN; PEI; KAMBER, 2011)

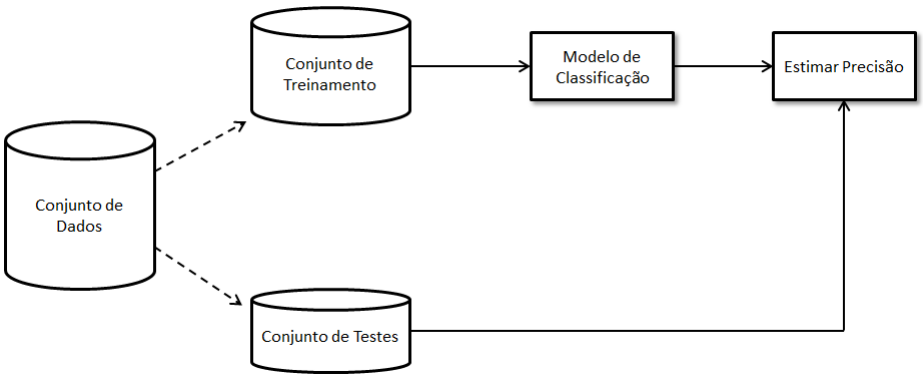


Figura 15 – Ilustração do método *Holdout*. Adaptado de (HAN; PEI; KAMBER, 2011)

Considere agora um conjunto de dados limitado. É preciso garantir que cada classe seja adequadamente representada nos conjuntos de treinamento e teste, com um número razoável de exemplos. Uma estratégia para isso é usar a **subamostragem aleatória** ou validação estratificada. Esse método compreende em usar o método *holdout* repetidas vezes, treinar e testar com diferentes amostras aleatórias. Iterativamente uma porção dos dados é aleatoriamente selecionada para o treinamento, realizando uma estratificação, o restante usada para os testes. As taxas de erro nas diferentes iterações são calculadas para gerar uma média de erro. Todavia, essa estratégia é válida somente para divisão igualitária entre dados de treinamento e de teste, o que geralmente não é possível. Para contornar essa situação, uma técnica estatística denominada de **validação cruzada** pode ser utilizada.(WITTEN *et al.*, 2016)

2.3.4.1 Validação Cruzada

Na validação cruzada, as iterações e particionamento dos dados são parâmetros de entrada, onde cada registro é usado o mesmo número de vezes para treinamento e exatamente uma vez para teste. Considere o processo onde os primeiros dados são divididos aleatoriamente em k subconjuntos mutualmente exclusivos, também conhecido como *folds*, $D_1, D_2, ..., D_k$, cada um com tamanhos equivalentes. O processo iterativo de treinamento e testes é realizado k vezes. Na iteração i, a partição D_i é definida como o conjunto de testes e as partições restantes são usadas de forma unificada para treinar o modelo. Isto é, na primeira iteração, os subconjuntos $D_2, ..., D_k$ formam um único conjunto de treinamento para construir o primeiro modelo, que por sua vez é testado em D_1. Na segunda iteração o treinamento é feito pelos subconjuntos $D_1, D_3, ..., D_k$ e testados em D_2, e assim por diante. Não diferente da subamostragem aleatória, o erro total é calculado pela soma dos erros das k iterações.(HAN; PEI; KAMBER, 2011)(TAN; STEINBACH; KUMAR, 2009)(WITTEN *et al.*, 2016)

2.3.5 Métodos Estatísticos para Comparar Modelos de Classificação

Em muitos casos, é preciso comparar dois ou mais modelos de classificação para o mesmo problema, a fim de avaliar qual tem melhor desempenho. De certa forma, parece ser trivial estimar realizar tal comparação. As métricas e métodos apresentados até aqui, podem ser usados para estimar o erro do modelo para determinado conjunto, e então decidir como "melhor", aquele que obtiver uma estimativa menor do erro. No entanto, pode ser que a diferença que distingue os dois modelos, seja causada somente pela métrica ou método de estimativa. É importante determinar para alguns casos, se determinado modelo é realmente melhor que outro em um problema particular, e não trata-se apenas de um efeito aleatório no processo de estimativa. (WITTEN *et al.*, 2016)

Os testes estatísticos fornecem limites de confiança, que garantem que o desempenho

de um modelo não está atrelado a uma estimativa para um conjunto de dados. Eles buscam determinar se um modelo é melhor ou pior que outro, na média de todos os treinamentos e conjuntos de testes possíveis, extraídos do domínio. Desta maneira, o objetivo é determinar se a média do conjunto de amostras (estimativas geradas por algum método, como validação cruzada ou *bootstrap* por exemplo) tem significância estatística maior ou menor, que a média de outro.(WITTEN *et al.*, 2016)(TAN; STEINBACH; KUMAR, 2009)

Suponha que para cada modelo, é realizada a validação cruzada com 10-*folds*, repetindo 10 vezes. Cada vez usando um particionamento diferente e cada particionamento é esboçada de forma independente. Dessa forma, podemos calcular a média das 10 taxas de erro obtidas respectivamente para M_1 e M_2, a fim de obter uma taxa de erro média para cada modelo. Para um determinado modelo, as taxas de erros individuais, calculadas nas validações cruzadas, podem ser consideradas como amostras diferentes e independentes de uma distribuição de probabilidade. Com isso é possível fazer testes de hipóteses, como o teste de *Friedman* e o de *Wilcoxon*.

2.3.5.1 Teste de *Wilcoxon*

O teste de *Wilcoxon* (TWX) é um teste não-paramétrico, e como TTS compara duas amostras pareadas. O TWX, também conhecido como *Wilcoxon Matched-Pairs* e *Wilcoxon signed-ranks test*, foi desenvolvido por *F. Wilcoxon* (WILCOXON, 1945) e usa os postos das diferenças intra-pares como base de cálculo, onde a variável deve ser mensurada de forma ordinal. O teste classifica em postos a diferença entre os modelos sobre cada amostra, usada para estimar a precisão e avaliar o desempenho. A diferença numérica de cada par, proporciona três condições possíveis: aumento (+), diminuição (-) ou igualdade (=). Dessa forma, soma-se as diferenças positivas 2.19 e negativas 2.20.

$$W^+ = \sum_{d_i>0} r_i + \frac{1}{2}\sum_{d_i=0} r_i, \tag{2.19}$$

$$W^- = \sum_{d_i<0} r_i + \frac{1}{2}\sum_{d_i=0} r_i, \tag{2.20}$$

onde r_i é o posto ou *ranking* da i-ésima amostra estimada, computado pela diferença entre os modelos comparados. A próxima etapa compreende em calcular o valor T, dado pela menor dos somatórios dos *rankings* de mesmo sinal.(DEMŠAR, 2006)

$$T = min(W^+, W^-), \tag{2.21}$$

O passo seguinte é determinar o total das diferenças com sinal e descontado o número de igualdades ($d_i = 0$), dado por n. Caso $n \leq 25$, significa que trata-se de valores críticos de T, e uma tabela de significância apropriada deve ser consultada. Caso $n > 25$,

é preciso utilizar o cálculo estatístico z, dado que trata-se de uma distribuição de dados que é aproximadamente normal.(DEMŠAR, 2006)

$$z = \frac{T - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}, \tag{2.22}$$

O TWX assume que a diferença entre os dois modelos não é significativa, representando uma hipótese nula. Assumindo um nível de significância de 5%, a hipótese nula não é rejeitada se $-1.96 \leq z \leq 1.96$.

2.3.5.2 Teste de *Friedman*

Nenhum dos testes estatísticos descritos anteriormente, consegue fazer suposições estatísticas para múltiplos modelos e amostras. Para Demšar (2006) testes pareados não são recomendados para comparar mais de dois modelos, isso porque múltiplos testes precisam ser realizados, e de certa forma uma proporção das hipóteses nulas são rejeitadas devido ao acaso, fazendo pouco sentido lista-las e compara-las. O teste de *Friedman* (TFR)(FRIEDMAN, 1940) é um teste estatístico não paramétrico, indicado para comparar diversos modelos de classificação sobre múltiplas amostras.

De maneira semelhante ao TWX, no TFR os modelos são organizados em forma de *ranking*, de acordo com o desempenho obtido sobre cada amostra de dados e computado por alguma métrica. Atribui-se 1 ao primeiro colocado, 2 ao segundo e assim sucessivamente. Em seguida, é calculado a média dos *rankings* obtidos pelos modelos sobre todas os conjuntos experimentados. (DEMŠAR, 2006)

Tome r_i^j o j-ésimo *ranking* de k modelos no i-ésimo de N conjunto de dados. O TFR compara as classificações médias dos modelos, $R_j = \frac{1}{N}\sum_i r_i^j$. Para o TFR, a hipótese nula afirma que todos os modelos envolvidos no teste são equivalentes, assim seus rankings devem ser iguais seguindo o cálculo estatístico:

$$X_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{3}], \tag{2.23}$$

onde X_F^2 é um distribuição com $k-1$ graus de liberdade. No entanto, o TFR como descrito, é considerado muito conservador e um novo cálculo estatístico derivado sobre 2.23 é apresentado em 2.24, seguindo uma distribuição F de *Snedcor* com $K-1$ e $(k-1)(N-1)$ graus de liberdade.(SHESKIN, 2003)

$$F_F = \frac{(N-1)X_F^2}{N(k-1) - X_F^2}, \tag{2.24}$$

2.4 Considerações Finais

Este capítulo foi dedicado ao referencial teórico, abordando todo conteúdo teórico-científico para embasar a proposta e o desenvolvimento deste trabalho. Foi apresentado uma revisão geral sobre *outliers* e seus principais conceitos como as causa de origem de anomalias, aspectos relacionados a sua detecção e as principais abordagens e métodos para identificação de *outliers*. Foi abordado também os principais conceitos sobre classificação, como um modelo de classificação é induzido e ocorre a dedução de rótulos, bem como se avalia o desempenho de um classificador e modelos são comparados, com objetivo de seleciona-los.

O desfio deste trabalho é desenvolver uma estratégia em que o desempenho qualitativo de técnicas de DO possam ser mensuradas de forma parametrizada e uniforme. Os princípios e conceitos apresentados neste capítulo, formam diretamente e indiretamente a estrutura necessária, para o desenvolvimento de uma abordagem que possa cumprir tal objetivo. Uma metologia que adapte o problema de classificação para comparar diferentes técnicas de DO, é objeto de discussão neste trabalho.

3 Trabalhos relacionados

Este capítulo é destinado aos trabalhos existentes na literatura sobre detecção de *outliers*, investigando publicações que desenvolveram e aplicaram técnicas em diferentes contextos de dados, ou comparando-as com outras técnicas a um determinado domínio. Este capítulo também tem o objetivo de apresentar trabalhos que contribuíram com metodologias para a análise de desempenho de técnicas de DO. As técnicas aqui apresentadas, estão inseridas em diversas abordagens, como as baseadas em classificação, distância, densidade, agrupamento e estatística.

Técnicas baseadas em classificação são caracterizadas por se aplicarem em conjuntos de dados totalmente rotulados. Por sua vez elas se dividem em duas categorias: modelo para múltiplas classes e para somente uma classe. As técnicas usadas para múltiplas classes, assumem que exemplos rotulados pertencem a várias classes normais, ou na forma de classe normal e *outlier*. Por outro lado, técnicas voltadas para classe única, assumem que todas as instâncias do conjunto possuem somente uma classe normal. Independente da categoria, a estratégia comum é construir um modelo preditivo para classes normais e anormais. Dessa forma, bases de dados balanceadas é imprescindível para construção de modelos efetivos nesse tipo de abordagem.(CHANDOLA; BANERJEE; KUMAR, 2009)

As Árvores de Decisão (AD) são amplamente estudadas nas áreas de aprendizado de máquina, e quando aplicadas a problemas de detecção de anomalias produzem resultados eficazes(AGRAWAL; AGRAWAL, 2015). Amor, Benferhat e Elouedi (2004) fornece um estudo experimental, onde compara o algoritmo *Naive Bayes* com AD aplicado ao problema de detecção de intrusos. O estudo foi realizado com o conjunto de dados KDD'99, considerando três níveis de condições: Ataques inteiros, agrupando em quatro categorias principais e focando somente em comportamentos normais e anormais. Segundo os resultados obtidos, a AD apresentou maior precisão. Tambem aplicado na detecção de intrusos, Sindhu, Geetha e Kannan (2012) propõe um algoritmo otimizado baseado em AD para classificar padrões normais e anômalos. A estratégia consiste em remover instâncias redundantes, aplicar um algoritmo de seleção de características baseado em *wrapper* e realizar a identificação de intrusos usando um procedimento iterativo de AD neural, inspirado em vários classificadores AD. Kumar, Hanumanthappa e Kumar (2012) também sugere usar AD para detecção de intrusos. O objetivo do trabalho deles foi propor um modelo para detectar ataques desconhecidos de forma eficiente.

Técnicas para detectar *outliers* baseadas no k-vizinho mais próximo (KNN) exigem medidas de distâncias ou similaridade para classificar, calculada entre duas instâncias de dados. Diferentemente das técnicas baseadas em agrupamento, as medidas precisam

ser positivas e simétricas(CHANDOLA; BANERJEE; KUMAR, 2009). No contexto de detecção de intrusos em redes de computadores, Li *et al.* (2007) propuseram um método melhorado do TCM-KNN (*Transductive Confidence Machines for K-Nearest Neighbors*). TCM introduziu na computação a medida de confiança baseada na teoria de aleatoriedade algorítmica. Tal medida de confiança usada no TCM é baseada em testes genéricos de aleatoriedade e aproximação. O método demonstrou-se ser efetivo na detecção de anomalias, com baixa taxa de falsos positivos e alta precisão. Barbará, Domeniconi e Rogers (2006) propõe um método também baseado no TCM-KNN para detectar anomalias em bases com *clusters* definidos. O método foi testado em diversos contextos e mostrou-se ser robusto, apresentando bom desempenho em conjunto de dados ruidosos.

Técnicas inspiradas na teoria do conjunto difuso, trata do raciocínio aproximado. Diferente da lógica clássica, aplicações baseadas em lógica *fuzzy* lidam com dados gerados por especialista de situações reais e problemas complexos (AGRAWAL; AGRAWAL, 2015). Cox (1995) aplicou um sistema de detecção de fraude baseado em lógica *fuzzy*, com objetivo de analisar pedidos de provedoras de planos de saúde. Este sistema *fuzzy* se baseia em dados de especialistas humanos para detectar padrões de comportamentos fraudulentos. No artigo de Raj e Portia (2011) várias técnica utilizadas nos para detecção de fraude de cartão de crédito são avaliadas, considerando a metodologia com baseada em determinados critérios de design. Um das técnicas comparadas é o sistema de detecção darwiniana *fuzzy* (BENTLEY *et al.*, 2000), ele usa programação genética para evoluir regras de lógica *fuzzy*, capaz de classificar transações de cartão de crédito em classes "suspeitas"e "não suspeitas". Nos experimentos a técnica darwiniana apresentou maior precisão. Já no trabalho de Pathak, Vidyarthi e Summers (2005), propuseram um sistema especializado baseado em lógica *fuzzy*, no qual identifica e avalia se dados fraudulentos estão envolvidos na liquidação de sinistros de seguros. A metodologia proposta é ilustrada com um exemplo cujo objetivo é modelar reivindicações de seguros em geral.

As métricas de similaridade, são responsáveis por medir a distância ou similaridade entre dois objetos. O resultado desta função é formar grupos onde objetos com determinado grau de homogeneidade estejam inseridos. Da mesma forma, esses objetos deverão ser bem diferentes de objetos de outros grupos (ZAMONER, 2013). Os algoritmos de agrupamento são capazes de identificar *outliers* sem conhecimento prévio dos dados. Existem vários métodos que podem ser aplicados para a detecção de anomalia (CHANDOLA; BANERJEE; KUMAR, 2009).

O algoritmo *k-means* é um método particional para o agrupamento de dados, simples e eficaz. Os métodos particionais tem o intuito de dividir a base em k-grupos, onde k é um parâmetro de entrada no algoritmo. Pamula, Deka e Nandi (2011) exploram o uso do *k-means* na detecção de anomalias em conjunto de dados médicos. O método exclui os pontos que estão próximos dos centros do *cluster* e calcula os *scores* dos candidatos a

outliers baseado nos pontos restantes. O método foi aplicado na base de dados WDBC, que é utilizado para extração de características no diagnóstico de câncer de mama. O resultados demonstram que o método é preciso. Santhanam e Padmavathi (2014) comparam o *k-means* com técnicas estatística para identificação e remoção de *outliers*. O objetivo é identificar e remover dados com valores extremos, promovendo a redução do volume de dados em bases médicas. Na análise comparativa, o método K-means provou-se ser mais eficiente.

Os métodos particionais identificam grupos de objetos com base na proximidade entre eles. No agrupamento por densidade, os grupos são definidos como regiões densas, que está associado a quão próximos os objetos estão. Uma função de densidade define quais regiões mais densas formam grupos, que por sua vez são separados por regiões menos densas.(LINDEN, 2009)

Um dos principais métodos para encontrar a densidade local para uma instância de dados é o LOF (*Local Fator Outlier*). Desenvolvido por Breunig *et al.* (2000), o método LOF atribui a cada objeto um grau de significância para ser *outlier*. O grau vai estar associado o quão isolado este objeto esta da sua vizinhança. Lee *et al.* (2010) aplicaram LOF no monitoramento de processos no campo industrial. Devido aos métodos tradicionais apresentarem deficiência na detecção de falhas de processo, os autores integraram o LOF no Controle de Processo Estatístico Multivariante (MSPC) para resolver as limitações. Os resultados mostraram a eficácia do esquema proposto.

Outra técnica desenvolvida por Zhang, Hutter e Jin (2009), pode ser comparada com o KNN e LOF, mas sendo menos sensível a valores de parâmetros segundo os resultados. O Fator *Outlier* baseado na distância local (LDOF), usa a localização relativa de um objeto com sua vizinhança, determinando assim o grau em que o objeto esta distante. Já o Fator de Densidade Local (LDF) substitui a estimativa de densidade do LOF pela estimativa de densidade de *kernel* Gaussiana com largura variável (KDE), que por sua vez foi modificada para usar a distância de acessibilidade do LOF. A densidade de *kernel* não tem mais o seu estimador no sentido matemático. Agora, a distância original do KDE (Euclidiana) é substituída pela distância de acessibilidade de LOF a fim de produzir uma estimativa de densidade local. Latecki, Lazarevic e Pokrajac (2007) experimenta seu método em vários domínios, comparando com outras técnicas como o LOF. Os resultados mostram que LDF é superior em todas comparações.

Outra técnica baseada em densidade é a Integral de Correlação Local (LOCI). Desenvolvida por Papadimitriou *et al.* (2003), eles introduzem também o Fator de Desvio de Multi-Granularidade (MDEF), que lida com as variações de densidade locais no espaço de dados e detecta valores anômalos isolados como agrupamentos periféricos. O algoritmo LOCI identifica candidatos a *outliers*, avaliando se seu valor MDEF se desvia significativamente das médias locais. Os experimentos foram realizados com o conjunto

de dados da NBA, e mostram que LOCI pode detectar de forma eficiente *outliers* e *microclusters*.

A depender do problema, *outliers* podem estar situados onde a distribuição de densidade na vizinhança seja diferente, por exemplo, objetos de um *cluster* esparso próximo de um mais denso. Essa configuração pode resultar em estimativas erradas. O algoritmo *INFLuenced Out-lierness*(INFLO) desenvolvido por Jin *et al.* (2006), propõe detectar *outliers* locais baseado numa relação de vizinhança simétrica. Ele considera o vizinho e o vizinho reverso para estimar a distribuição de densidade. E para fazer isso de forma eficiente usa alguns algoritmos de mineração para encontrar valores atípicos com abordagem top-n. Os resultados mostram a eficácia do método no conjunto de dados NHL.

Métodos estatísticos não paramétricos, não necessitam de conhecimento prévio dos dados. Muitos conjuntos de dados simplesmente não seguem um modelo de distribuição específico e são distribuídos de forma aleatória. Abordagens não paramétricas são mais flexíveis e autônomas (HODGE; AUSTIN, 2004). Dasgupta e Forrest (1996) introduzem um método não paramétrico para detectar novidades em operação de máquinas. A operação produz uma série de tempo de medidas em máquinas monitoradas, mapeadas em vetores binários usando a quantificação *binning*. Os autores usaram um conjunto de métodos não paramétricos onde conseguem definir a normalidade de novos dados inseridos.

Algumas técnicas podem detectar *outliers* baseando-se somente na distribuição dos dados. O *Boxplot*, introduzido por Tukey (1977), é um método não paramétrico para identificar valores anômalos com base na subtração dos quartis inferior e superior. A distância entre s quartis serve como uma medida de disseminação dos dados, semelhante ao desvio padrão. Já o método *Hampel* (HAMPEL *et al.*, 2011), identifica *outliers* usando somente a mediana da amostra e o desvio mediano absoluto.

Os métodos semi-paramétricos utilizam modelos de *kernel* locais no lugar de somente um modelo de distribuição. O objetivo é combinar a vantagem a eficiência dos métodos paramétricos com a flexibilidade dos métodos não paramétricos (HODGE; AUSTIN, 2004). Roberts e Tarassenko (1994) usam modelos de mistura gaussiana para aprender um modelo de dados normais em dados de eletroencefalograma (EEG). O objetivo é identificar valores anormais que representam condições médicas como a epilepsia. A mistura representa um núcleo cuja largura é determinada pela distribuição dos dados. O método funciona adicionando de forma incremental novos modelos de mistura. Caso a mistura que melhor representa o novo exemplar estiver fora uma distância limiar, então o algoritmo adiciona uma nova mistura. Este limiar é determinado de forma autônoma na fase de treinamento. Uma vez que o treinamento é concluído, o limite de distância final representa o limite "*outlier*"para novas instâncias.

Os trabalhos supracitados, são destinados a abordar estratégias baseadas em diversas abordagens para o problema da detecção de *outliers*. Muitos desses trabalhos

comparam a performance da técnica proposta em diferentes domínio de dados, ou mais de uma técnica em determinado domínio. Eles objetivam principalmente, validar a implementação proposta, comparando de forma pareada com outra técnica ou aplicando-a em diferentes conjuntos de dados. No entanto, são raros os trabalhos que exploram um estudo comparativo mais extenso, utilizando metodologias que consigam comparar com efetividade técnicas de DO baseadas em diferentes abordagens, em diferentes contextos de dados.

O trabalho de Bakar *et al.* (2006), descreve uma análise de desempenho de técnicas baseadas em *Control Chart*, regressão linear e distância, para a detecção de *outliers*. Eles conseguem estabelecer qual técnica obteve melhor desempenho em determinada situação. Todavia, é uma metodologia que usa conhecimento prévio de quais instâncias são de fato *outliers*. Isso faz com que as técnicas e bases a serem utilizadas no experimento estejam restritas a esse cenário. Além disso, a metodologia proporciona somente comparações pareadas.

Campos *et al.* (2016), realizaram um extenso estudo comparativo, com diversas técnicas de DO não supervisionadas, baseadas em na relação de k-vizinho mais próximo. Conjuntos de dados de diferentes contextos foram usados nos experimentos. Para comparar as diferentes técnicas, propuseram ajustar a pontuação *outlier* gerada pelas técnicas, com o ajuste de probabilidade discutido em (HUBERT; ARABIE, 1985), para validação de agrupamentos. Essa metodologia permitiu que comparações significativas dos desempenhos das técnicas em diferentes conjuntos de dados, pudessem ser realizadas e possibilitado conclusões importantes. No entanto, o trabalho se limita a comparar somente técnicas baseadas em distância, e a metodologia não apresenta um *benchmark* final para cada técnica, impossibilitando uma análise de desempenho e estatística de forma objetiva.

3.1 Considerações Finais

Este capítulo foi dedicado a discussão de técnicas de detecção de *outliers* existentes na literatura. De forma empírica, a busca tentou abranger os principais domínios de aplicação, como também as principais técnicas, organizadas por abordagem.

Cada uma das numerosas técnicas apresentadas aqui tem suas vantagens e desvantagens. Saber qual técnica de detecção de *outliers* é mais adequada para um determinado domínio, consiste em um desafio pouco explorado na literatura. Metodologias para comparar tais técnicas também compõe uma contribuição importante para preencher essa lacuna. Para Chandola, Banerjee e Kumar (2009) uma pesquisa abrangente sobre detecção de *outliers* deve possibilitar ao leitor compreender não somente a motivação para detectar anomalias e usar uma determinada técnica para o problema, mas também fornecer uma análise comparativa de várias técnicas nos principais domínios de aplicação. Ao aplicar uma

determinada técnica, toda conjectura pode ser usada como diretrizes para avaliar a eficácia da técnica em determinado contexto. Dessa forma, a falta de trabalhos recentes, que buscam estratégias para avaliar técnicas de DO de forma efetiva, encoraja o desenvolvimento deste trabalho. A tabela 2 apresenta um resumo dos trabalhos aqui citados.

	Classificação	**Agrupamento**	**Proximidade**	**Estatística**	**Estudos comparativos**
Detecção de intrusos	Amor, Benferhat e Elouedi (2004); Sindhu, Geetha e Kannan (2012); Kumar, Hanumanthappa e Kumar (2012)	–	–	–	–
Detecção de fraude	Cox (1995); Bentley *et al.* (2000); Pathak, Vidyarthi e Summers (2005)	–	–	–	Raj e Portia (2011)
Contexto industrial		–	Lee *et al.* (2010)		–
Detecção de anomalias em redes de computadores	Li *et al.* (2007)	–	–	–	–
Detecção de anomalias em bases médicas	–	–	Pamula, Deka e Nandi (2011)	Roberts e Tarassenko (1994)	–
Outros domínios ou diversos contextos	Barbará, Domeniconi e Rogers (2006)	Santhanam e Padmavathi (2014)	Latecki, Lazarevic e Pokrajac (2007); Papadimitriou *et al.* (2003); Jin *et al.* (2006)	–	Bakar *et al.* (2006); Campos *et al.* (2016)

Tabela 2 – Resumo dos trabalhos relacionados

4 Metodologia para Comparar Técnicas de Detecção de *Outliers*

A aplicabilidade em diversos domínios de aplicação, faz o problema da análise de DO ser de grande importância e constitui-se como objeto de estudo em diferentes áreas de pesquisa. Os métodos de DO discutidos até aqui compartilham um objetivo em comum, identificar objetos que se desviam da maior parte dos dados. No entanto, em muitos casos, aplicar um método de DO a um domínio é uma tarefa empírica, e validar o método implica exatamente em analisar os resultados de tal aplicação. Aggarwal (2015) em seu livro, aborda algumas maneiras para avaliar técnicas de DO, porém para ele, a validação de *outliers* é um problema um tanto difícil, devido aos desafios associados à natureza não supervisionada de algumas técnicas de DO e a característica subjetiva do *outlier*. Não obstante, medir o desempenho de técnicas DO é só um aspecto da tarefa de comparar e selecionar diferentes técnicas, em domínios distintos. Para o problema de classificação, avaliar e distinguir estatisticamente modelos é uma atividade bastante discutida, e de certa forma, os métodos para essas tarefas são claros e objetivos. Contudo, para técnicas de DO, ter resultados qualitativos onde seja possível mensurar a eficácia do método ou se é adaptável a um determinado contexto, é ainda um desafio a ser superado.

Uma das etapas da fase de pré-processamento dos dados é remoção de ruídos, e uma consequência da realização dessa etapa, é uma melhor precisão dos modelos de classificação. Isso porque o ruído se trata de um dado que não condiz com a realidade, desviando uma indução correta do modelo (TAN; STEINBACH; KUMAR, 2009)(LAROSE, 2014). Por outro lado, o *outlier* pode ser representado como um dado ruidoso, e os métodos DO são úteis para remover o ruído (AGGARWAL, 2015). Para alguns autores existe um limiar que separa o conceito de *outlier* e ruído (QUINLAN, 1986)(HAN; PEI; KAMBER, 2011)(ALPAYDIN, 2014), no entanto todos são gerados de forma semelhante, seja por erros no processo de medida e coleta, por variação natural, por dados de classes diferentes e podem ter origens desconhecidas. Dessa maneira, *outlier* ou ruído, todos tem a mesma característica, são objetos que se desviam consideravelmente da maioria dos dados do conjunto (TAN; STEINBACH; KUMAR, 2009). O limiar que separa os dois tipos de dados anômalos, não é nosso objeto de estudo e é desconsiderado no desenvolvimento da proposta deste trabalho. Sendo assim, no contexto deste trabalho, os dois se referem ao mesmo tipo de dados.

Sabendo que os métodos de DO podem ser úteis na fase pré-processamento dos dados, removendo *outliers* do conjunto, podemos comparar diferentes técnicas de DO, medindo seu efeito na indução de modelos de classificação. Dessa forma, é possível mensurar

o desempenho das técnicas, através do efeito que elas produzem no pré-processamento, utilizando os métodos tradicionais de avaliação e comparação de classificadores.

4.1 Descrição da Metodologia

Para que técnicas de DO possam ter seu desempenho qualitativo comparado, no que tange a capacidade de identificar *outliers*, é imprescindível que exista uma metodologia para estruturar uma análise comparativa. Uma metodologia válida deve possibilitar uma parametrização das técnicas, para que a avaliação de desempenho seja feita de forma uniforme e assim classifica-las para cada situação. A aplicação das técnicas de DO na fase de pré-processamento dos dados, permite não só parametrizar seu comportamento, mas também: (i) quantificar o desempenho das técnicas de DO; (ii) Entender a relação entre os domínio de dados e os algoritmos; (iii) compreender como as características das bases afetam o desempenho das técnicas; (iv) e estabelecer e entender como se comporta a eficácia e eficiência das técnicas de DO.

Podemos estruturar uma metodologia para comparar diferentes técnicas de DO, da seguinte forma: considere $L = (l_1, l_2, ..., l_i)$ conjuntos de dados de domínios distintos, $D = (d_1, d_2, ..., d_j)$ o conjunto de diferentes técnicas de DO aplicado ao conjunto L, $A = (a_1, a_2, ..., a_k)$ o conjunto de algoritmos indutores, $M = (m_1, m_2, ..., m_n)$ os modelos de classificação gerados por A, antes da fase de pré-processamento, e $M' = (m'_1, m'_2, ..., m'_n)$ os modelos de classificação gerados por A, após a fase de pré processamento por d_i. As estimativas E_m (Estimativa anterior) e E'_m (Estimativa posterior), correspondem as precisões estimadas por cada modelo, antes e depois de aplicado d_i. A análise comparativa está na avaliação da precisão antes e depois dos modelos de classificação. A Figura 16 ilustra a arquitetura da metodologia proposta.

Essa metodologia implica em analisar o efeito das técnicas de DO, através estimativas dos modelos de classificação induzidos. Aqui as técnicas de DO (D) atuam como componentes de pré-processamento, retornam a lista de *outliers* identificados no conjunto de dados (L) para serem removidos. A partir disso, os algoritmos indutores (A) constroem modelos de classificação (M), onde suas precisões serão estimadas. Essas estimativas podem ser realizadas pelas métricas e métodos já discutidos para o problema de classificação (Seção 2.3.3). Conforme as estimativas geradas, podemos então analisar estatisticamente as predições, comparando o desempenho das técnicas de DO em cada cenário (Seção 2.3.5).

As técnicas de DO aqui estudadas, retornam um *ranking* geral das instâncias do conjunto de dados com base em seu fator *outlier*, computado de acordo com a abordagem que a técnica se baseia. Esse fator *outlier*, pode ser entendido também como uma pontuação ou *score* discutida na Seção 2.1.3.4. Na prática, um resultado objetivo esperado pelas

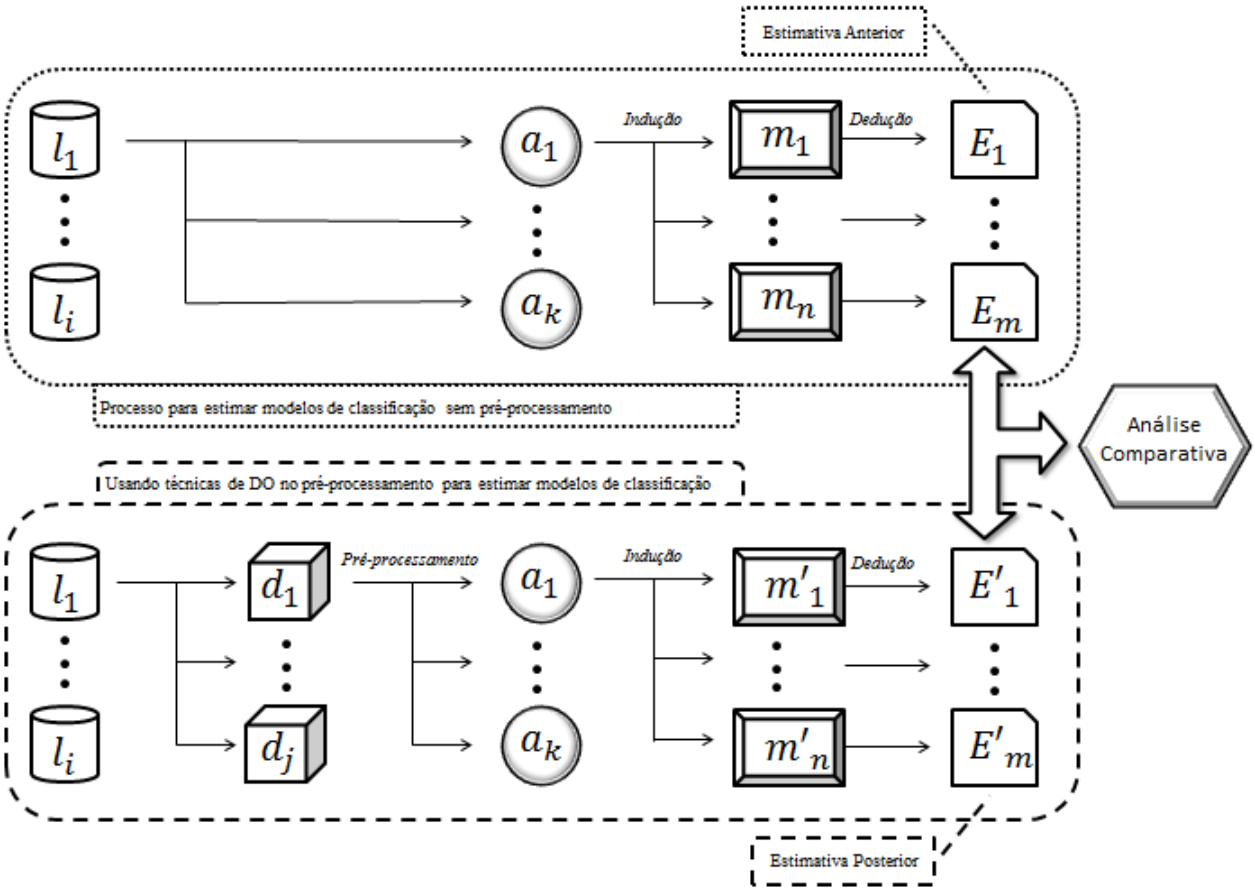


Figura 16 – Metodologia proposta para comparar técnicas de DO.

técnicas, seria um subconjunto de instâncias, que corresponderia aquelas com maiores pontuações *outliers*. No entanto, as pontuações fornecidas pelas mais variadas técnicas, diferem amplamente em sua escala, seu alcance e seu significado. Para muitos métodos, a escala de valores podem diferir até mesmo com mesmo método e conjunto de dados diferentes. Uma pontuação *outlier* em um determinado conjunto de dados pode significar que temos um *outlier*, enquanto em outro conjunto de dados aquela pontuação pode não ter o mesmo significado. Mesmo em casos onde a pontuação *outlier* é idêntica para duas instâncias de conjuntos diferentes, isso pode denotar graus com diferença substancial de confiabilidade, que por sua vez depende das diferentes distribuições de dados. Dessa forma, interpretar e comparar de diferentes técnicas de DO está longe de ser uma tarefa trivial (ZIMEK; SCHUBERT; KRIEGEL, 2012).

Seguindo este raciocínio, definir um limiar fixo para gerar um subconjunto de instâncias, com maiores pontuações *outliers*, não é uma opção. É preciso que o *ranking* gerado pela técnica DO seja padronizado, de tal forma que possibilite definir as pontuações *outliers* que se distanciam das demais, separa-las por um limiar crítico. O *Z-Score* é uma medida que corresponde a quantos desvios padrão, abaixo ou acima da população, uma

determinada pontuação se encontra (ALTMAN; DANOVI; FALINI, 2015). De acordo com a Figura 17 podemos inserir o *Z-Score* (Z) em uma curva de distribuição normal, e definir um limiar crítico (λ) para estabelecer as pontuações que estão distribuídas de forma isolada e em menor quantidade. Nesse caso, para calcular o *Z-Score* é necessário observar a pontuação *outlier* (x_i) para cada instância (i), a média (μ) e desvio padrão (σ) de todas pontuações geradas por técnica, como segue a equação: $Z = (x_i - \mu)/\sigma$. Se $Z > \lambda$, então a instância i é considerada *outlier*.

Como se trata de uma metodologia que está inserida no problema de classificação, é indispensável que os conjuntos de dados estejam rotulados. No entanto, isso não implica que a metodologia restrinja o uso de técnicas supervisionadas de DO. O objetivo das técnicas é identificar os *outliers* em dados rotulados por especialistas ou por processo estocástico, na fase de construção da base. Como função do pré-processamento dos dados removê-los. Esse é um processo não supervisionado, os rótulos não são usados para a detecção de *oultiers*. Isso permite que técnicas dos diferentes métodos de DO possam ser utilizados na análise comparativa.

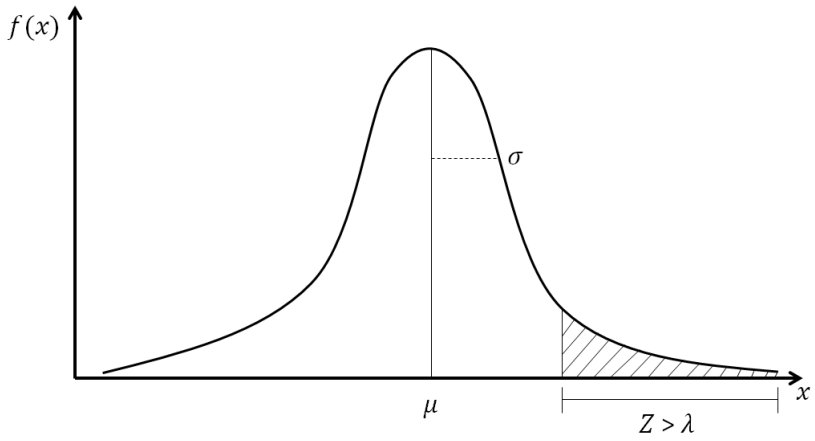


Figura 17 – Demonstração do *Z-Score* na distribuição normal

4.2 Materiais e Métodos

Esta Seção é dedicada a definição e descrição dos métodos, e conjuntos de dados que compõem os experimentos da metodologia proposta. O objeto de estudo em questão é comparar técnicas de DO, avaliando o efeito da remoção dos *outliers* na indução de modelos de classificação. As estimativas geradas na fase de dedução dos classificadores, serão os indicadores usados na análise comparativa. Para planejar os experimentos, primeiramente é preciso definir seu escopo, selecionando os métodos e bases que serão utilizadas.

4.2.1 Bases de dados

Este experimento tem como objetivo abranger grande parte do domínios que a DO se aplica. No entanto, a metodologia proposta restringe a utilização de bases totalmente rotuladas. Neste cenário, entre os exemplos rotulados, podem existir instâncias anômalas a classe em que foram rotuladas. As técnicas de DO identificam esses exemplos, que por sua vez são removidas por parte do pré-processamento, e o efeito deste processo é observado na classificação de dados. Para este experimento, foram selecionadas 14 bases, que podem ser aplicadas na detecção de fraudes financeiras, diagnostico médico, detecção de intrusos, processamento digital de imagens, detecção de anomalias em redes de computadores e detecção de falhas.

Amazon Employee Access: Os dados consistem em dados históricos reais coletados a partir de 2010 e 2011 da Amazon. Os dados são relacionados ao acesso provisionado de funcionários atuais aos sistemas. Ela foi fornecida pela *Amazon*, com objetivo de que pesquisadores construam um modelo, aprendido por esses dados históricos, que determine as necessidades de acesso de um funcionário de acordo com suas atribuições. Modelos preditivos construídos com essa base, vai agir como um detector de novidades, caso algum usuário se comporte como intruso.(GUEVARA; SANTOS; LÓPEZ, 2014)

BNG Breast Cancer Wisconsin: Esta é uma versão maior do conjunto de dados Breast Cancer Wisconsin, gerado por meio de uma rede *bayesiana* treinada no conjunto de dados original e usada para criar pseudo-instâncias. As características são calculadas a partir de uma imagem digitalizada de um material biológico, extraído através de uma punção aspirativa de agulha fina (PAAF), de uma massa mamária. Os atributos descrevem características dos núcleos celulares presentes na imagem. Nesse conjunto a instâncias são rotuladas com diagnóstico final do exame, normal ou anormal.(JR *et al.*, 1997)

Creditcard: Esta base representa dados de transações de credito que ocorreram em dois dias. Foi reduzida para viabilizar os experimentos, ficando com instâncias que representam 492 fraudes de 28.924 transações. O conjunto original englobava 284.807 transações. Os dados são resultados de transformações numéricas de transações financeiras, como "Tempo"e "Valor". (POZZOLO *et al.*, 2015)

DefaultCredit: Esta base foi construída com dados de pagamentos inadimplentes de clientes em Taiwan. O dados foram originados de transformações numéricas sobre dados pessoais e financeiros dos clientes. É aplicável na gestão de risco para estimar clientes confiáveis ou não confiáveis. (YEH; LIEN, 2009)

EEGEyeState: Os dados desse conjunto são referentes a uma medição de EEG contínua com o *EEI Emotiv Neuroheadset*. A duração da medição é de 117 segundos. O estado do olho foi detectado através de uma câmera durante a medição do EEG e

posteriormente adicionado manualmente a base ao analisar os quadros de vídeo. O rótulo com valor '1' indica o olho fechado e '0' o olho aberto. (ASUNCION; NEWMAN, 2007)

KDDCup99: O KDDCup99 foi derivado do conjunto de dados de avaliação DARPA IDS (LIPPMANN *et al.*, 2000), foi utilizado para a competição KDDCup99 (ASUNCION; NEWMAN, 2007). O conjunto de dados completo tem quase 5 milhões de padrões de entrada, onde cada registro representa uma conexão TCP/IP que é composta de 41 recursos que são categóricas e numéricas (STOLFO *et al.*, 2000). Para viabilização dos experimentos, reduzimos este conjunto para 26.648 instâncias, preservando classes com menor número de exemplos.

Letter: Esse conjunto é formado com dados de imagens de letras maiúsculas baseadas em 20 fontes diferentes, tranformado em valores numéricos. Cada letra foi replicada e artificialmente distorcida afim de produzir novas instâncias, resultando um total de 20.000 exemplos. Os rótulos representam 26 letras maiúsculas no alfabeto inglês. Essa base é destinada ao treinamento de modelos para reconhecimento digital de imagens.(ASUNCION; NEWMAN, 2007)

MagicGamma: Este conjunto é formados por dados gerados pelo programa Monte Carlo (HECK *et al.*, 1998), que simula registro de partículas gama de alta energia de um telescópio gama atmosférico *Cherenkov*, usando técnicas de processamento digital de imagens. Os dados numéricos são referentes aos registros capturados, com objetivo de detectar uma partícula gama (sinal interessante) e partículas de *hadron* (fundo). (ASUNCION; NEWMAN, 2007)

Mammography: Mammography consiste em dados retirados de mamografias coletadas em um hospital clínico. Dos dados numéricos são referentes as imagens de mamografia, que por sua vez foram classificadas como benignas ou malignas. Os tipos de tumores estão fortemente relacionados à forma da imagem: As massas benignas têm contornos lisos bem definidos, enquanto as malignas têm os contornos espalhados sobre o parênquima mamário. Possui 11.183 amostras com 260 calcificações. (WOODS *et al.*, 1993)

Mozilla4: No desenvolvimento de software de código aberto, é importante monitorar o tamanho do módulo e entender seu impacto na propensão a defeitos. Esse conjunto formado com dados de projeto de desenvolvimento do produto *Mozilla*. Os dados são originados do monitoramento de manutenção de módulos do software. (KORU; ZHANG; LIU, 2007)

PhishingWeb: Essa base é formada com dados de de sites normais e fraudulentos, com atributos que os caracterizam. Normalmente, os sites de *phishing* se disfarçam como sites confiáveis para obter informações confidenciais. Existem algumas características que distinguem sites de *phishing*, como URL longo, endereço IP em URL, adição de prefixo

e sufixo ao domínio e URL de solicitação, entre outras características. (MOHAMMAD; THABTAH; MCCLUSKEY, 2012)

SkinSegmentation: Este conjunto de dados foi construído usando dados de várias texturas de pele, obtidas de imagens faciais com diversidade de idade, gênero e raça, e por *pixels* obtidos de milhares de amostras aleatórias de texturas que não são de pele. O tamanho original da amostra é de 245057 instâncias, das quais 50859 são as amostras de pele e 194198 são amostras sem pele (BHATT *et al.*, 2009). Para a viabilização do experimento, foi preciso sub-amostrar esse conjunto em 10% para os exemplos que não são pele e 5% para os que são pele.

BNG Spam Base: Esta é uma versão maior do conjunto de dados Spam Base (ASUNCION; NEWMAN, 2007), construída por uma rede *bayesiana* treinada com o conjunto original, no qual gerou instâncias baseadas nas tabelas de probabilidades. O conjunto original é formado por dados de uma coleção de *e-mails* não-*spam* e *spam*. Estes dados são úteis para construção de filtros *spam*.(RIJN *et al.*, 2013)

HRSSanomalous: Esses dados foram coletados de um sistema highstorages, que transporta pacotes entre dois pontos. O arquivo gerado tem o registros das execuções do equipamento, e algumas desses são registros de falhas. Essas falhas são marcadas com valor 1 no rótulo. (HRANISAVLJEVIC; NIGGEMANN; MAIER, 2016)

Nome	Característica	Instâncias	Atributos	Classes	Domínio
Amazon Employee Access	Categórica	32769	10	2	Detecção de intruso
BNG Breast Cancer Wisconsin	Numérica	39366	10	2	Diagnóstico Médico
Creditcard	Numérica	28924	31	2	Fraude Financeira
DefaultCredit	Numérica	30000	24	2	Fraude Financeira
EEGEyeState	Numérica	14980	10	2	Diagnóstico Médico
KDDCup99	Multivariado	26648	42	23	Redes de Computadores
Letter	Numérica	20000	17	26	Processamento de Imagens
MagicGamma	Numérica	19020	11	2	Processamento de Imagens
Mammography	Numérica	11183	7	2	Diagnóstico Médico
Mozilla4	Numérica	15545	6	2	Detecção de falhas
PhishingWeb	Categórica	11055	31	2	Detecção de fraude
SkinSegmentation	Numérica	29592	4	2	Processamento de Imagens
BNG Spam Base	Categórica	10000	58	2	Detecção de intruso
HRSSanomalous	Numérica	23645	20	2	Detecção de falhas

Tabela 3 – Resumo de informações sobre os conjunto de dados

4.2.2 Técnicas de DO

O objeto de estudo deste experimento é comparar as técnicas de DO descritas aqui. O critério da escolha das técnicas aqui apresentadas, se deu por serem mais citadas na literatura e ter disponibilidade de sua implementação. Outro fator importante na escolha das técnicas, foi ter uma representação mínima de técnicas para cada método. Os desempenhos serão avaliados de acordo com o domínio e o algoritmo indutivo utilizados.

Para este experimento, foram selecionados 16 técnicas de DO, baseadas em métodos baseados em distância, densidade, agrupamento e estatística; e 3 Filtros de Classificação.

KNNOutlier (KNN) : Técnica baseada no algoritmo KNN tradicional, no qual detecta valores anômalos localizando a k vizinhança mais próxima de cada objeto. Como sabemos, os objetos de uma vizinhança partilham semelhanças entre seus principais atributos, enquanto objetos *outliers* são diferentes. O *KNNOutlier* é uma técnica baseado em distância, e mais detalhes estão descritos na Seção 2.2.2.2.2 .(YANG; HUANG, 2008a)

ODIN (ODN): Esta técnica baseia-se na estratégia de grafo kNN para detectar *outliers*. Nesta versão, em vez de usar um limiar para obter uma decisão binária, é gerado um escore *outlier*, normalizado por k para possibilitar comparações em diferentes parametrizações. Um grafo KNN define cada vértice como um vetor de dados, e as arestas são ponteiros para o K vetor vizinho. O fator *outlier* que o ODIN atribui para cada instância x_i é definido como: $O_i = 1/(ind(x_i) + 1)$, onde $ind(x_i)$ é o grau do vértice x_i, que corresponde o número de arestas apontando para x_i.(ISMO *et al.*, 2004)(HAUTAMÄKI *et al.*, 2005)

OPTICSOF (OPT): O OPTICSOF é um algoritmo de DO baseado no tradicional OPTICS. O OPTICS por sua vez é uma derivação do DBScan, onde dispensa o parâmetro *Epsilon*. Ele faz uma hierarquização dos *Epsilon* possíveis e seleciona o melhor para determinada distribuição. Para este algoritmo é necessário somente o parâmetro $MinPts$. A ideia básica do algoritmo DBSCAN original é que, para cada objeto de um *cluster*, a vizinhança de um determinado raio (*Epsilon*) deve conter pelo menos um número mínimo de objetos ($MinPts$). Se na vizinhança-*Epsilon* de um objeto p, contenha pelo menos $MinPts$ de instâncias, p é então considerado um objeto central, não *outlier*.(BREUNIG *et al.*, 1999)

DWOF (DWF): É uma técnica baseada em densidade que calcula fatores *outliers* por raio dinâmico, parametrizado por um valor k relacionado a vizinhança. O DWOF usa um raio crescente para cada objeto, com objetivo de calcular uma pontuação não paramétrica. Objetos dentro de um raio pertencerão ao mesmo *cluster* do objeto. Os raios iniciais são calculadas após estimar a densidade circundante de cada objeto, que é inversamente proporcional à distância euclidiana média entre os k-vizinhos mais próximos daquele objeto.(MOMTAZ; MOHSSEN; GOWAYYED, 2013)

LOF: Algoritmo baseado em densidade para calcular fatores *outliers* locais. Originalmente, parâmetro de vizinhança do LOF é *MinPoints*, mas para consistência dentro do ELKI, este parâmetro foi renomeado para k. Mais detalhes sobre o LOF, consultar a Seção 2.2.2.2.2. (BREUNIG *et al.*, 2000)

ALOCI (ALC): É uma algoritmo baseado em densidade que usa o conceito Fator de Desvio de Multi-Granularidade (MDEF). O MDEF lida com as variações de densidade

locais no espaço de dados. o ALOCI identifica candidatos a *outliers*, avaliando se seu valor MDEF se desvia significativamente das médias locais. O MDEF no raio r para um ponto p_i, é o desvio relativo de sua densidade de vizinhança local da densidade média de vizinhança local em sua r-vizinhança. Dessa forma, um objeto cuja densidade de vizinhança corresponde à densidade média de vizinhança local, terá um MDEF de 0. Por outro lado, *outliers* terão MDEFs longe de 0. (PAPADIMITRIOU *et al.*, 2003) (ZHU *et al.*, 2004)

COF: O algoritmo COF (Fator *Outlier* Baseado em Conectividade) diferencia "baixa densidade" da "isolatividade". Enquanto a baixa densidade normalmente se refere ao fato de que o número de objetos em uma vizinhança definida é relativamente pequena, a "isolatividade"está associada ao grau em que um objeto está "conectado" a outros objetos. O COF modifica a estimativa de densidade do LOF para explicar a "conectividade"de uma vizinhança. Isso é possível através de uma árvore geradora mínima (MST), originada no ponto em observação. O MST é reduzida de volta para a raiz p, eliminando progressivamente o incidente do nó da folha até a borda que tem maior comprimento. A computação do comprimento total das arestas para cada MST e, em seguida, a média dos MSTs geram a "distância média do encadeamento"(DME) para o conjunto kNN de p. O COF em p, é a razão do DME de p com a média das DME's dos vizinhos de distância k.(TANG *et al.*, 2002) (CAMPOS *et al.*, 2016)

INFLO (IFL): O algoritmo INFLO calcula os relacionamento simétrico entre as instâncias para detectar *outliers*. Derivado do LOF, usa busca de vizinhos mais próximos (NN) e kNN reverso (RNN). O RNN de um objeto p, são essencialmente objetos que possuem p como um dos seus k-vizinhos mais próximos. Quando consideramos a relação de vizinhança simétrica de NN e RNN, é possível definir o espaço de um objeto que é influenciado por outros objetos, fazendo com que as densidades de sua vizinhança sejam melhor estimadas e, portanto, os *outliers* identificados serão mais significativos. (JIN *et al.*, 2006)

KDEOS (KDS): Este é um algoritmo de DO genérico com estimativas flexíveis de densidade do *kernel*. O KDEOS é uma derivação do LOF, no entanto usa a estimativa de densidade do *kernel* a partir de estatísticas. O KDEOS compara a estimativa de densidade com as densidades nas vizinhanças locais. Apesar dessa ser uma característica originalmente do LOF, o KDEOS usa a estimativa clássica da densidade do *kernel*, ao invés de de experimentar *kernels* não padronizados e sem fundamentação. (SCHUBERT; ZIMEK; KRIEGEL, 2014)

LDOF (LDF): O Fator *Outlier* baseado na Distância Local (LDOF), usa a localização relativa de um objeto com sua vizinhança, determinando assim o grau em que o objeto esta distante. O fator *outlier* é calculado como a razão entre a média das distâncias do ponto p e seus k-vizinhos mais próximos, e a média das distâncias entre pares de um conjunto kNN. O LDOF favorece *outliers* que estão longe de formar um conjunto kNN.

compacto.(ZHANG; HUTTER; JIN, 2009)(CAMPOS *et al.*, 2016)

LOOP (LOP): O LOOP (Probabilidade Outlier Local), é um método híbrido que usa modelo probabilístico do ALOCI e LOF para definir um fator *outlier*, e usa métodos estatísticos para alcançar uma melhor estabilidade de resultados. Ele usa uma estimativa de densidade local mais robusta do que o LOF, baseando-se na distância média quadrática. O valor LOOP será próximo de 0 para pontos dentro de regiões densas, e próximo de 1 para valores *outliers*. (KRIEGEL *et al.*, 2009)

EMOutlier (EM): Este algoritmo de detecção de *outliers* é baseado no *EM Clustering*. Se um objeto não pertence a nenhum *cluster*, é um candidato a *outlier*. Se a probabilidade de um objeto pertencer ao *cluster* mais próximo ser relativamente baixa, esse objeto ainda será um candidato a *outlier*. Mais detalhes sobre o EM, consultar a Seção 2.2.2.1.3 (SCHUBERT *et al.*, 2015)

KMeansOutlier (KMS): O *K-Means* é uma técnica de agrupamento, amplamente discutida na Seção 2.2.2.3. Os *outliers* identificados por esta técnica, estão consideravelmente distantes dos k-centros. (SCHUBERT *et al.*, 2015)

COP: Probabilidade de Correlação Outlier (COP), detecta *outliers* como pontos que não se encaixam em nenhuma correlação local significativa dos dados. O COP é a primeira abordagem a considerar correlações locais dentro do processo de detecção de *outliers*. simultaneamente é identificado, *outliers* em subespaços, que por sua vez são arbitrariamente orientados pelo espaço dos dados, e a correlação de atributos relevantes que precisam ser considerados para detectar o *outlier* são determinados. O parâmetro k é referente ao número de vizinhos a serem considerados. A Seção 2.2.2.1.2 discute com maior detalhe esta técnica. (KRIEGEL *et al.*, 2012)

GaussianModel (GM): É um Método Gaussinao para detectar *outliers*. A pontuação de *outlier* é calculada a partir da densidade de probabilidade da distribuição. A suposição de um modelo Gaussiano para valores *outliers*, compreende em que todos os pontos de dados que estão localizados em uma distribuição de probabilidade com um único centro (ou seja, uma *cluster* Gaussiano único), diferem dos pontos de dados distantes do centro do *cluster*.(SCHUBERT *et al.*, 2015)(AGGARWAL, 2015)

Inter-Quartil(*boxplot*) (INT): O *boxplot* é um técnica estatística comumente usada para identificar *outliers*, amplamente discutida na Seção 2.2.2.1.1. A técnica implementada identifica *outliers* e valores extremos. Para este experimento, consideramos como anomalias somente as instâncias identificadas como *outliers*.

Filtro de Classificação: O filtro de classificação é uma estratégia para identificação e remoção de *outliers*. Instâncias incorretamente rotuladas por um classificador, são potenciais *outliers*, e removê-los é desempenhar uma função de filtro que melhora o desempenho de classificadores que venham a ser induzidos posteriormente (GAMBERGER;

LAVRAC; GROSELJ, 1999). Trata-se de uma estratégia supervisionada para detecção e remoção de *outliers*. É relevante frisar, que os algoritmos usados como filtros de classificação, são distintos dos algoritmos indutores, de outro modo, os modelos de classificação estariam sendo gerados baseando-se em dados de treinamento. Para este experimento foi selecionado três modelos para atuarem como filtro de classificação:

- ***JRipper*(JRP)**: é um algoritmo baseado em regras de decisão, que é uma implicação da forma: se A então B. A parte condicional A é conjunção de condições. Cada condição é definida por uma relação entre um atributo e os valores do domínio (FACELI *et al.*, 2011).
- ***Classificador baseado em k-instâncias*(IBK)**: O IBK é um método de classificação baseado no k-vizinho mais próximo. Como os algoritmos de aprendizado baseados em instâncias, originalmente não mantêm um conjunto de abstrações derivadas de exemplos específicos, utilizar a estratégia de k-vizinho mais próximo aumenta o recurso de armazenamento.(AHA; KIBLER; ALBERT, 1991)
- **Classificador *Fuzzy* Baseado em Regras (FR)**: O Algoritmo de Indução de Regras Fuzzy Não-Ordenadas (FURIA) é um método de classificação baseado em lógica *fuzzy*. O FURIA é um derivação do JRipper, e fornece um conjuntos de regras simples e claras. no entanto, o FURIA aprende regras *fuzzy* em vez das regras convencionais, e o conjuntos de regras não é ordenado.(HÜHN; HÜLLERMEIER, 2009)

4.2.3 Algoritmos Indutores

Os algoritmos indutores irão construir modelos de classificação, partindo dos conjunto de dados antes e depois de pré-processados pelas técnicas de DO. O desempenho de cada técnica, será medido pela diferença da acurácia dos classificadores gerados. Para este experimento, os algoritmos indutores envolvidos fazem parte de métodos baseado em árvore de decisão, em estatística e máquinas de vetores de suporte. O critério para a seleção desses indutores, em parte seguiu o mesmo para os filtros de classificação, serem distintos. Outro fator foi o limitação do recurso computacional, dados que algoritmos com maior complexidade inviabilizaria o experimento computacional.

J48: O J48 é um algoritmo desenvolvido com base na árvore de decisão C4.5. Esse algoritmo constrói uma árvore binária para modelar o processo de classificação. Depois que a árvore é construída, ela é aplicada a cada tupla no banco de dados e resulta em classificação para essa tupla. (DUNHAM, 2006)

***Naive Bayes* (NB)**: O NB é um classificador que calcula um conjunto de probabilidades de acordo com a frequência e as combinações de valores para todas as instâncias

do conjunto. O algoritmo é baseado no teorema das redes Bayesianas, e assume que os atributos são independentes. Essa suposição univariada dificilmente é válida em aplicações reais, no entanto o algoritmo tende a ter bom desempenho para vários problemas de classificação.(PATIL; SHEREKAR, 2013)

***Support Vector Machines* (SVM)**: As SVMs usam fronteiras lineares para a separação de objetos pertencentes a duas ou mais classes (FACELI *et al.*, 2011). A implementação SVM desse experimento faz parte da LibSVM, que é uma biblioteca que reúne implementações SVMs com ampla variedade de parâmetros. (CHANG; LIN, 2011)

4.2.4 Configuração do Experimento

A configuração do experimento está pautada em descrever implementações, plataformas, recurso computacional e parâmetros de algoritmos usados e que possibilitasse o experimento computacional. Todas as técnicas de DO estão inseridas nas ferramentas ELKI e WEKA. O ELKI é uma ferramenta de código aberto para mineração de dados. Destinada a avaliação de algoritmos em bases de alta dimensionalidade, tem ênfase em métodos não supervisionados, em análise de *clusters* e detecção de *outliers* (SCHUBERT *et al.*, 2015). Os Algoritmos de DO implementados no ELKI retornam um *ranking* com o fator *outlier* para cada instância. Esse resultado é usado na fase de pré-processamento da metodologia.

As técnicas de Interquartil e os filtros de classificação, estão inseridos na ferramenta WEKA, bem como a execução do experimento para induzir e estimar os classificadores. O WEKA representa um robusto *software* de aprendizado máquina para tarefas de mineração de dados. O *WEKA* agrega diversos algoritmos e inclui ferramentas para pré-processamento, classificação, agrupamento e visualização dos dados. Por ser uma ferramenta de código aberto, ela também é útil para o desenvolvimento de novos projetos de aprendizado de máquina (HOLMES; DONKIN; WITTEN, 1994).

Para que os resultados gerados pelos algoritmos do ELKI fossem pré-processados, foi necessário realizar implementação na ferramente WEKA. O filtro *RemoveByElki* recebe os resultados dos algoritmos do ELKI como entrada, realiza a padronização dos fatores *outliers* em *Z-Score*, gera o subconjunto de instâncias de acordo com o limiar crítico e então exclui essas instâncias do conjunto de dados selecionado. *RemoveByElki* recebe como entrada o arquivo de texto com a lista de todas instâncias, com seus respectivos fatores *outliers*, e o valor de limiar crítico (λ), que por sua vez está relacionado com o nível de confiança (α) da distribuição normal. Essa implementação constitui uma das contribuições do presente trabalho.

Na Figura 18 é possível demonstrar a relação do nível de confiança e limiar crítico. O Valor de limiar crítico selecionado para o experimento foi de 1,64 unilateral, que representa

92,5% de confiança. Isso significa que instâncias com fator *outlier* distribuído acima desse limiar crítico, é considerado de fato *outlier*.

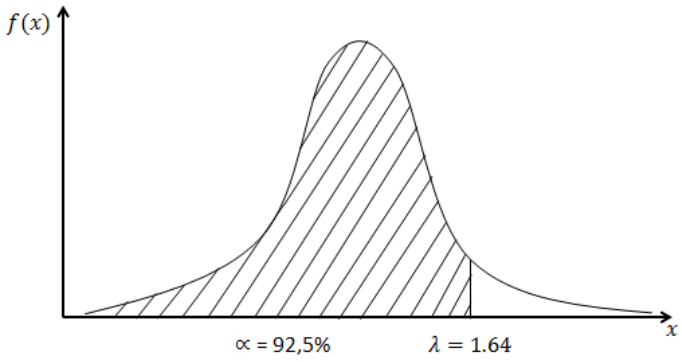


Figura 18 – Parâmetros do limiar crítico para os experimentos

Para executar experimentos computacionais no contexto da classificação de dados, faz-se necessário o uso de recurso computacional adequado. É importante que esse recurso suporte a complexidade dos algoritmos e a dimensão dos conjuntos de dados. Experimentos no âmbito de aprendizado de máquina, atenta-se para os componentes memória RAM e processador. Para este trabalho, foi utilizado três computadores nos experimentos computacionais, a Tabela 4 descreve todo recurso computacional utilizado.

	Processador	**Memória RAM**
Máquina 1	Intel Xeon E5 8 *Core*, 2.6 GHz	32 GB
Máquina 2	AMD A8 4 *Core*, 3.1 GHz	16 GB
Máquina 3	AMD A8 4 *Core*, 3.1 GHz	16 GB

Tabela 4 – Características do recurso computacional utilizado nos experimentos

Um outro aspecto que faz parte da configuração do experimento, é a definição dos parâmetros para cada algoritmo. No ELKI as implementações não estabelecem valores padrão de parâmetro. Por outro lado, dependendo da abordagem que os algoritmos de DO estão inseridos, eles também compartilham dos mesmos parâmetros. Algoritmos derivados de abordagem de distância, o parâmetro está relacionado a quantidade de vizinhos. Algoritmos baseados em densidade, o parâmetro representa a quantidade mínima de vizinhos para estabelecer um grupo. Algoritmos baseados em agrupamento, seu parâmetro é referente a quantidade de grupos que o conjunto irá ser particionado. E algoritmos estatísticos constroem um modelo estocástico sobre os dados e recebem como parâmetro um valor de limiar. O *GaussianModel* não dispõe de parâmetro de entrada e o Interquartil tem parâmetro padrão definido pelo WEKA. A Tabela 5 apresenta os valores dos parâmetros

estabelecidos para cada técnica. Os parâmetros dos filtros de classificação e algoritmos indutores, também foram definidos com os valores padrão do *WEKA*.

Os valores estabelecidos para os parâmetros, para alguns casos, seguiram valores fixos definido de forma empírica, como o caso de técnicas baseadas proximidade, onde o parâmetro k está relacionado ao cálculo de vizinhança. Os valores definidos foram influenciados também pelo recurso computacional disponível. Quanto maior o valor do parâmetro k referente a vizinhança, maior o tempo computacional para gerar os resultados. Em outros casos, também definidos de forma empírica, mas de acordo com a distribuição dos dados, como foi o caso do OPTICOF e o COP. Outro fator influenciador foi a qualidade dos resultados, como exemplo o parâmetro das técnicas DWOF e ALOCI, valores inferiores não geravam resultados satisfatórios, no que tange a identificação de instâncias *outliers*. Por fim, os casos em que as técnicas são baseadas em agrupamento, onde o parâmetro k está relacionado ao particionamento dos dados, para esses o valor definido seguiu o numero de classes que cada conjunto de dados possuía.

BASES	KNN	ODN	OPT	DWF	LOF	ALC	COF	IFL	KDS	LDF	LOP	EM	KMS	COP
Amazon	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
Breast Cancer	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
Creditcard	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=40$
Default Credit	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=25$
EEGEyeState	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
KDDCup99	$k=3$	$k=3$	$MinPts=100$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=23$	$k=23$	$k=20$
Letter	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=26$	$k=26$	$k=20$
MagicGama	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
Mammography	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
Mozilla4	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
PhishingWeb	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=40$
SkinSegmentation	$k=3$	$k=3$	$MinPts=100$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=40$
SpamBase	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$
HRSSAnomalous	$k=3$	$k=3$	$MinPts=5$	$k=10$	$k=3$	$Kmin=20$	$k=3$	$k=3$	$k=3$	$k=3$	$k=3$	$k=2$	$k=2$	$k=20$

Tabela 5 – Valores estabelecidos dos parâmetros do algoritmos de DO.

4.2.5 Métricas de Avaliação

Como já discutido, o desempenho das técnicas de DO, será mensurado pelo efeito do pré-processamento dos dados no comportamento dos classificadores. Para esse fim, é importante avaliar o desempenho antes e depois do pré-processamento. Neste experimento, a medida de avaliação utilizada será somente a Acurácia (Acc), gerado pelos classificadores na dedução ao conjunto de testes. A Acurácia foi amplamente discutida na Seção 2.3.3.1, e ela permitirá avaliar se o pré-processamento realizado pela técnica de DO, teve um efeito positivo ou negativo no desempenho do classificador.

O método selecionado para estimar a precisão do classificador, foi a validação cruzada com 10-*folds*. Amplamente discuta na Seção 2.3.4.1, constitui-se da estratégia comumente usada para medir o desempenho de modelos de classificação. A validação cruzada vai permitir que todos os dados do conjunto, possam ser usados na fase de treinamento (WITTEN *et al.*, 2016).

Como visto na Seção 2.3.5, a diferença de desempenho entre as técnicas, pode estar relacionado ao método de estimativa e não refletir exatamente sua performance. Para evitar esse tipo de situação, e para que exista uma avaliação efetiva das técnicas de DO, é preciso definir testes estatísticos. Para este experimento, foi definido que primeiramente os resultados iriam ser submetidos ao teste não-paramétrico de *Friedman*, discutido amplamente na Seção 2.3.5.2. A intenção desse teste é comparar todos os resultados de $acc0$ (Acurácia dos classificadores antes do pré-processamento) com os resultados de $acc1$ (Acurácia dos classificadores depois do pré-processamento), caso a hipótese nula seja rejeitada, o teste de *Wilcoxon* pareado (Seção 2.3.5.1) é realizado, como forma de pós-teste. O teste de *Wilcoxon* vai identificar diferenças significativas, e o valor absoluto do desempenho vai indicar a superioridade e inferioridade da técnica.

4.3 Considerações Finais

Neste capítulo foi apresentado a metodologia proposta para comparar técnicas de DO. O objetivo é avaliar o desempenho das técnicas selecionadas em cada contexto de dados. Numa perspectiva de pré-processamento dos dados, será possível a avaliar também qual técnica de DO tem melhor atuação, pré-processando *outliers* para induzir determinado classificador. Foi apresentado também, os elementos que irão compor os experimentos. As técnicas de DO, os conjuntos de dados, os algoritmos indutores e as métricas de avaliação dos classificadores constituem os experimento para a metodologia proposta.

O ELKI e WEKA possuem todos os requisitos, ferramentas e implementações necessárias para a condução dos experimentos. O WEKA por ser um *software* de código aberto, foi fundamental para implementar o filtro *RemoveByElki* e realizar os experimentos para estimar os classificadores. O ELKI possui uma gama muito variada de algoritmos de DO implementados, e os dados de saída facilitaram a condução da implementação no WEKA.

Para estimar a precisão dos classificadores, foi estabelecido a utilização do método de validação cruzada como 10-*folds*. É o método mais usado na comunidade, e com ele é possível usar todos os dados na fase de treinamento. Para este experimento, somente a acurácia dos classificadores será considerado para avaliar o desempenho das técnicas de DO. Para comparar os resultados com efetividade, foi estabelecido a realização dos testes estatísticos não-paramétricos *Friedman* e *Wilcoxon* pareado, esse último como pós-teste.

5 Resultados Experimentais

No capítulo anterior foi apresentado a metodologia proposta para comparação das técnicas de DO, definido os elementos e a configuração que constituem os experimentos. Este capítulo é dedicado a apresentação e análise dos resultados experimentais. Aqui será discutido como os resultados são organizados e apresentados, e os fatores que influenciaram o desempenho de cada técnica de DO. Este trabalho limitou-se em observar os resultados de acordo com a configuração geral de parâmetros apresentada na seção 4.2. Notoriamente, um melhor ajuste dos parâmetros para cada algoritmo em cada base provocaria um melhor desempenho. No entanto, o escopo deste trabalho não abrange a otimização de parâmetros.

De acordo com a metodologia proposta, antes do uso de DO, a acurácia dos modelos de classificação (C) são mensurados, essa etapa podemos nomear de Acc_0. No passo posterior é realizado um pré-processamento dos dados, onde as técnicas de DO identificam as instâncias anômalas para então serem removidas dos conjuntos de dados. Em seguida, a acurácia dos classificadores é medida novamente. Essa etapa podemos denominar de Acc_1. Para um melhor entendimento, melhor organização e visualização, a apresentação dos resultados para os algoritmos de DO e os filtros de classificação foram separadas. Esta divisão se dá também pela natureza de cada estratégia. Enquanto os algoritmos de DO fazem uso de estratégias não supervisionadas, os filtros de classificação configuram-se como uma forma supervisionada para detectar *outliers*.

As Tabelas 6 e 7 apresentam a taxa total de instâncias que cada técnica de DO identificou como anômala ($TxOutlier$). Da mesma maneira, as médias por base e por técnica. Compreendendo a fase de pré-processamento, essas instâncias identificadas como *outliers* são removidas. Os valores de $TxOutlier$ são representados em forma de porcentagem. Esta análise levou em consideração as taxas por classe, no entanto para facilitar a leitura, essa informação foi sumarizada em dois possíveis efeitos. Caso o resultado represente toda uma classe identificada como *outlier*, estará seguido de um (*) (Foi considerado para esse caso, taxa por classe $> 99\%$). Caso mais da metade das instâncias sejam identificadas como *outliers*, estará sublinhado (Foi considerado para esse caso, taxa total $> 50\%$).

Os valores dessas tabelas correspondem as taxas de dados *outliers*(%) geradas por cada técnica, seguido de suas médias (μ), onde a horizontal representa a média por base e a vertical, a média por técnica. A Tabela 6 apresenta os resultados gerados pelos algoritmos de DO, em que a média geral de *outliers* identificados é de 6,72%. As técnicas OPTICSOF e COP apresentaram maiores médias, com 18,26% e 17,96% cada. Esses valores destoam dos demais resultados, e quando observado os valores absolutos por

BASES	KNN	ODN	OPT	DWF	LOF	ALC	COF	IFL	KDS	LDF	LOP	EM	KMS	COP	GM	INT	μ
Amazon	5,00	7,83	0,14	0,00	0,14	11,43*	0,10	0,14	0,00	0,23	0,31	0,00	5,69	25,34	9,01	0,00	4,09
Breast Cancer	10,33	8,48	18,22	0,00	1,00	0,28	0,33	0,98	3,63	0,41	3,12	0,74	10,86	21,66	10,52	1,28	5,74
Credicard	2,20	5,46	3,76	4,67	4,05	12,83	4,79	4,04	0,36	4,52	9,04	0,00	0,87	0,00	7,19	5,58	4,33
Default Credit	4,95	11,81	0,70	0,00	0,65	9,63	0,30	0,58	0,12	0,28	0,56	0,00	3,47	0,00	9,24	21,37	3,98
EEGEyeState	0,02	9,23	0,02	0,00	0,02	19,49	0,02	0,02	0,00	0,02	0,03	0,00	0,02	0,00	8,95	7,42	2,83
KDDCup99	0,00	9,58	12,52	0,00	2,06	0,00	0,46	3,05	6,06	0,13	0,01	0,00	0,01	24,20*	0,00	15,98*	4,63
Letter	5,99	9,25	10,93	0,00	10,32	16,70	4,76	10,93	0,83	5,33	9,18	0,00	7,71	0,13	5,52	1,45	6,19
MagicGama	8,20	9,21	7,95	5,94	7,48	14,20	7,74	7,99	0,62	5,09	10,54	0,00	4,68	24,67	10,35	6,25	8,18
Mammography	3,03	29,77	31,11	0,00	1,07	6,68	3,25	1,33	8,21	1,83	9,12	0,01	1,34	2,07	2,44	5,52	6,67
Mozilla4	7,19	6,04	0,91	5,81	1,10	8,88	0,57	1,04	0,01	0,42	1,38	0,00	9,16	14,02	10,18	3,51	4,39
PhishingWeb	8,80	7,07	35,10	0,00	21,37	32,09	2,51	21,00	9,72	5,09	8,58	0,00	11,57	17,13	4,39	0,00	11,53
SkinSegmentation	4,81	8,77	32,05	0,00	8,75	2,04	0,77	8,85	10,31	0,70	8,01	0,00	13,02	38,56	70,34	3,49	13,15
SpamBase	6,86	0,00	97,14	0,00	6,02	0,00	1,60	5,91	2,47	0,38	3,65	0,55	7,15	79,49	0,00	0,00	13,20
HRSSAnomalous	6,75	12,30	5,11	0,54	6,39	10,39	0,27	2,46	2,13	0,67	0,27	0,00	7,10	4,17	7,36	17,89	5,24
μ	5,30	9,63	18,26	1,21	5,03	10,33	1,96	4,88	3,18	1,79	4,56	0,09	5,90	17,96	11,11	6,41	6,72

Tabela 6 – Taxas de *outliers* identificados pelos algoritmos de DO

base, é possível observar uma alta variância. Isso pode indicar uma alta sensibilidade no ajuste dos parâmetros dessas técnicas ou ineficiência da técnica para a metodologia proposta. A técnica OPTICSOF por exemplo, tem como principal parâmetro o $MinPts$, que estabelece o número mínimo de instâncias a ser considerada numa vizinhança. Isso está diretamente relacionado no agrupamento dos dados e identificação de *outliers*. Podemos observar também que alguns algoritmos de DO obtiveram baixa de taxa de identificação de *outliers* (DWOF, COF, LDOF e EM). O efeito desse resultado vai poder ser observado no desempenho pós-processamento, onde as instâncias identificadas como anômalas são removidas. Notoriamente, quanto menor a taxa de remoção de *outliers*, menor irá ser a alteração do estado da base, consequentemente menor a tendência de um efeito positivo ou negativo pós-processamento.

A Tabela 7 apresenta os resultados gerados pelos filtros de classificação, em que a média geral de *outliers* identificados é de 14,60% e o filtro JRP com maior média, 17,39%. É perceptível a diferença das taxas dos filtros de classificação com as taxas dos algoritmos de DO. Isso se dá pela forma que as instâncias são consideradas *outliers* para os filtros de classificação, que está diretamente relacionado com o seu desempenho como classificadores.

BASES	JRP	IBK	FR	μ
Amazon	5,78	6,40	6,33	6,17
Breast Cancer	5,53	3,02	1,91	3,49
Credicard	0,84	0,15	0,13	0,38
Default Credit	18,91	33,74	18,04	23,56
EEGEyeState	52,32	33,10	38,92	41,45
KDDCup99	0,00	0,43*	0,60*	0,34
Letter	50,22	10,31	36,86*	32,46
MagicGama	33,80	26,91	20,29	27,00
Mammography	2,27	1,90	1,65	1,94
Mozilla4	6,79	15,72	6,57	9,69
PhishingWeb	7,21	6,64	6,97	6,94
SkinSegmentation	1,49	0,08	0,28	0,62
SpamBase	34,42	33,79	34,76	34,32
HRSSAnomalous	23,83	5,08	19,34	16,08
μ	17,39	12,66	13,76	14,60

Tabela 7 – Taxas de *outliers* identificados pelos filtros de classificação

Os aspectos negativos que podem ser demonstrados nessa etapa de resultados, resumem-se ao fato de cada técnica ter identificado toda uma classe como *outlier* e/ou ter identificado mais da metade das instâncias como anômalas. Esse último aspecto é negativo porque após o processamento, existe uma maior probabilidade de ocasionar um desbalanceamento na base ou descaracterizar o balanceamento original. Contudo, segundo os resultados, foram poucos casos desse tipo observados. Para todos algoritmos de DO e bases, somente 3/224 casos onde toda uma classe foi identificada como *outlier* e 3/224 casos onde mais da metade das instâncias foi identificada como *outlier*. Para os filtros de classificação, somente 3/42 casos onde toda uma classe foi identificada como *outlier* e 2/42 casos onde mais da metade das instâncias foi identificada como *outlier*.

As Tabelas 9 e 16 apresentam o desempenho dos classificadores antes (Acc_0) e depois (Acc_1) do pré-processamento, realizado pelas técnicas de DO. Os resultados são expressos pela taxa média de acurácia de cada modelo e pelo desvio padrão dos 10-*folds* da validação cruzada. Os valores de *Acc* e desvio padrão, estão aqui representados em forma de porcentagem. Essas tabelas também expressam os resultados dos testes estatísticos descrito na Seção 4.2. Na Tabela 9 os valores de Acc_1 foram comparados com os valores de Acc_0, e aqueles casos que obtiveram hipótese nula rejeitada segundo o teste estatístico, estão destacados. Quando os resultados de Acc_1 tem melhora significativa quando comparado com os valores de Acc_0, este casos estão formatados com fundo cinza. Quando os resultados de Acc_1 tem piora significativa quando comprados com os valores de Acc_0, nestes casos os valores destacam-se com borda simples.

(%)	**Melhorou (42 casos)**	**Piorou (42 casos)**	**Outliers (14 casos)**	**Vitória (28 casos)**
KNN	28,57	7,14	0,00	32,14
ODN	0,00	16,67	7,14	0,00
OPT	4,76	9,52	0,00	3,57
DWF	0,00	0,00	71,43	0,00
LOF	4,76	0,00	0,00	3,57
ALC	21,43	7,14	14,29	25,00
COF	4,76	0,00	0,00	7,14
IFL	7,14	0,00	0,00	7,14
KDS	2,38	0,00	14,29	0,00
LDF	0,00	2,38	0,00	0,00
LOP	9,52	0,00	0,00	7,14
EM	0,00	0,00	78,57	0,00
KMS	11,90	11,90	0,00	14,29
COP	26,19	4,76	21,43	28,57
GM	2,38	19,05	14,29	0,00
INT	16,67	9,52	21,43	3,57

Tabela 8 – Taxa de melhoria,piora e vitórias após a fase de pré-processamento. Taxa de casos em que não foi identificado *outliers* (Coluna "*Outliers*"). Resultado total por algoritmos de DO.

Bases	C	Acc_0(%)	Acc_1(%)														
			KNN	***ODN***	***OPT***	***DWF***	***LOF***	***ALC***	***COF***	***IFL***	***KDS***	***LDF***	***LOP***	***EM***	***KMS***	***COP***	***GM***	***INT***
Amazon	**J48**	94,21±0,01	94,15±0,00	94,16±0,02	94,21±0,02	94,21±0,01	94,21±0,02	100,00±0,01	94,21±0,02	94,21±0,01	94,21±0,01	94,21±0,01	94,20±0,01	94,21±0,01	94,25±0,01	95,40±0,02	93,64±0,02	94,21±0,01
	NB	88,75±0,41	89,03±0,57	88,62±0,70	88,67±0,75	88,75±0,41	88,65±0,74	99,47±0,13	88,63±0,55	88,65±0,61	88,75±0,41	88,73±0,46	88,66±0,42	88,75±0,41	89,04±0,53	91,09±0,51	88,34±0,57	88,75±0,41
	SVM	94,94±0,21	94,92±0,23	95,03±0,22	95,05±0,34	94,94±0,21	94,92±0,28	100,00±0,01	94,98±0,25	94,93±0,24	94,94±0,21	94,91±0,23	95,04±0,43	94,94±0,21	95,01±0,37	96,26±0,33	94,65±0,21	94,94±0,21
Breast Cancer	**J48**	98,17±0,18	98,54±0,25	97,93±0,15	97,78±0,22	98,17±0,18	98,12±0,15	98,19±0,25	98,05±0,25	98,15±0,16	98,23±0,28	98,17±0,25	98,10±0,16	98,26±0,20	98,69±0,21	97,94±0,26	97,97±0,19	98,17±0,20
	NB	95,50±0,19	95,61±0,73	95,39±0,36	95,25±0,27	95,50±0,19	95,46±0,37	95,49±0,17	95,48±0,43	95,46±0,33	95,50±0,36	95,49±0,42	95,44±0,47	95,68±0,32	96,15±0,31	95,51±0,25	95,34±0,33	95,54±0,27
	SVM	97,96±0,21	98,43±0,32	97,78±0,21	97,52±0,31	97,96±0,21	97,95±0,22	97,96±0,19	97,95±0,24	97,95±0,18	97,95±0,34	97,96±0,24	97,92±0,21	97,98±0,23	98,64±0,24	97,72±0,25	97,75±0,25	98,02±0,23
Credicard	**J48**	99,88±0,05	99,89±0,06	99,89±0,07	99,95±0,04	99,87±0,08	99,94±0,03	99,88±0,04	99,95±0,04	99,93±0,05	99,88±0,07	99,91±0,05	99,93±0,04	99,88±0,05	99,88±0,04	99,88±0,05	99,89±0,05	99,93±0,04
	NB	97,35±0,30	97,98±0,23	97,27±0,40	97,43±0,22	97,39±0,23	97,56±0,31	97,77±0,31	97,65±0,24	97,57±0,35	97,32±0,35	97,61±0,38	97,81±0,32	97,35±0,30	97,55±0,54	97,35±0,30	97,33±0,43	97,25±0,29
	SVM	94,70±11,23	94,94±12,66	90,36±17,45	98,63±0,07	86,12±27,04	91,99±20,53	91,11±16,50	98,61±0,13	98,63±0,09	97,81±1,68	90,47±23,84	93,01±17,95	94,70±11,23	74,57±32,19	94,70±11,23	97,68±1,93	98,77±1,58
Default Credit	**J48**	80,33±0,77	79,89±0,63	80,37±0,86	80,19±0,57	80,33±0,77	80,55±0,75	80,23±0,62	80,26±0,45	80,03±0,61	80,60±0,52	80,09±0,66	80,38±0,52	80,33±0,77	80,12±0,88	80,33±0,77	79,70±0,80	78,93±0,48
	NB	69,35±4,92	75,61±0,91	67,96±4,00	70,13±4,55	69,35±4,92	69,71±3,15	74,15±2,13	69,72±3,90	69,72±3,74	69,75±4,15	69,26±4,14	69,98±3,63	69,35±4,92	69,82±3,50	69,35±4,92	63,04±3,37	55,60±1,63
	SVM	60,13±4,65	52,55±7,30	56,47±6,98	57,40±7,30	60,13±4,65	54,60±8,45	52,67±7,97	52,33±7,87	48,97±8,69	55,07±10,04	51,25±6,99	57,52±6,56	60,13±4,65	52,83±6,17	60,13±4,65	52,65±9,33	57,56±9,81
EEGEyeState	**J48**	84,50±1,09	84,35±1,24	84,25±0,69	84,96±0,95	84,50±1,09	84,72±1,24	83,57±0,77	84,91±0,94	84,71±0,55	84,50±1,09	84,60±1,09	84,62±0,67	84,50±1,09	84,76±0,77	84,50±1,09	84,02±0,92	84,11±0,82
	NB	46,77±3,26	49,44±4,26	46,84±3,77	49,44±4,11	46,77±3,26	47,45±3,51	45,89±4,51	49,19±4,06	49,47±4,67	46,77±3,26	47,52±3,80	61,97±1,67	46,77±3,26	49,49±4,38	46,77±3,26	46,59±3,16	45,62±1,65
	SVM	63,91±0,97	63,90±1,03	63,89±0,88	64,13±0,63	63,91±0,97	64,16±1,34	65,05±1,60	64,09±1,29	64,18±1,24	63,91±0,97	63,81±1,16	64,12±0,99	63,91±0,97	64,21±1,37	63,91±0,97	64,15±0,92	64,41±1,57
KDDCup99	**J48**	99,59±0,10	99,60±0,12	99,57±0,11	99,57±0,08	99,59±0,10	99,61±0,07	99,59±0,10	99,60±0,07	99,60±0,07	99,62±0,15	99,59±0,08	99,61±0,11	99,59±0,10	99,59±0,13	99,94±0,05	99,59±0,10	99,68±0,10
	NB	94,22±0,54	94,27±0,49	93,54±0,28	94,65±0,76	94,22±0,54	94,32±0,50	94,22±0,54	94,21±0,48	94,30±0,40	94,75±0,63	94,37±0,41	94,31±0,33	94,22±0,54	94,29±0,45	97,61±0,30	94,22±0,54	95,18±0,41
	SVM	94,16±3,09	97,27±1,78	93,68±5,78	95,47±3,87	94,16±3,09	97,22±2,05	94,16±3,09	95,48±2,27	94,82±4,75	94,92±7,03	94,23±1,88	96,50±3,46	94,16±3,09	94,01±3,62	99,73±0,38	94,16±3,09	98,82±0,74
Letter	**J48**	87,98±0,51	89,33±0,75	86,87±0,59	87,34±0,89	87,98±0,51	87,83±1,01	88,04±0,90	87,63±0,58	87,71±0,81	87,94±0,73	87,82±0,63	88,15±0,96	87,98±0,51	87,81±0,73	87,96±0,58	87,94±0,72	88,32±0,94
	NB	64,11±0,76	67,86±0,94	63,78±0,96	64,65±0,70	64,11±0,76	63,65±1,22	66,82±1,04	63,30±0,79	64,07±1,07	64,04±0,79	63,36±1,63	63,53±0,67	64,11±0,76	64,49±1,10	64,47±1,14	65,07±1,17	65,05±0,82
	SVM	85,10±1,01	87,45±0,83	84,67±0,64	85,05±0,79	85,10±1,01	85,25±1,03	87,73±0,91	85,08±0,59	85,41±0,69	85,29±0,66	85,03±0,97	85,16±0,73	85,10±1,01	85,05±0,80	85,36±0,60	85,32±0,76	85,27±0,98
MagicGamma	**J48**	85,06±0,98	84,51±0,85	84,80±0,58	85,46±0,65	84,97±0,71	85,33±0,85	84,14±0,71	85,59±0,70	85,24±1,18	84,95±1,09	85,24±0,92	85,44±1,19	85,06±0,98	84,64±0,67	86,13±1,14	84,85±0,50	84,19±0,99
	NB	72,69±0,87	75,52±0,85	71,97±0,97	73,54±0,78	72,97±0,39	73,55±0,73	73,86±0,62	73,57±0,84	73,46±1,30	72,77±0,78	72,94±0,72	73,66±1,17	72,69±0,87	72,99±1,04	75,47±1,09	72,01±1,09	74,69±0,99
	SVM	65,05±20,66	76,61±3,92	62,52±20,69	64,02±21,52	65,14±21,06	60,39±23,21	72,32±14,96	47,53±21,32	56,27±20,85	62,43±21,64	72,63±14,12	43,80±18,89	65,05±20,66	52,38±22,68	61,38±26,44	68,96±16,66	66,05±19,01
Mammography	**J48**	98,57±0,26	98,79±0,29	97,96±0,52	98,27±0,51	98,57±0,26	98,60±0,33	98,59±0,30	98,72±0,24	98,66±0,23	98,68±0,33	98,61±0,30	98,62±0,33	98,57±0,34	98,57±0,28	98,64±0,30	98,48±0,28	98,39±0,35
	NB	95,67±0,30	94,87±1,18	95,28±0,59	95,16±0,35	95,67±0,30	95,44±0,66	94,98±0,52	95,52±0,67	95,59±0,90	95,58±0,69	95,56±0,54	95,56±0,63	95,64±0,75	95,23±0,64	95,59±0,52	95,60±0,53	95,06±0,60
	SVM	98,20±0,28	98,20±0,05	97,72±0,27	97,86±0,37	98,20±0,28	98,35±0,23	98,51±0,05	98,40±0,20	98,38±0,17	98,23±0,24	98,24±0,19	98,41±0,20	98,23±0,20	98,09±0,03	98,31±0,16	98,19±0,24	98,33±0,28
Mozilla4	**J48**	94,80±0,60	94,88±0,55	94,64±0,64	94,72±0,36	94,56±0,76	94,73±0,53	94,42±0,81	94,81±0,47	94,75±0,55	94,64±0,64	94,70±0,34	94,76±0,39	94,80±0,60	94,33±0,64	94,84±0,64	94,42±0,72	94,57±0,49
	NB	68,64±1,50	68,33±0,69	68,57±0,57	68,60±0,96	68,60±0,70	68,64±0,95	69,52±1,33	68,51±0,60	68,58±1,25	68,55±1,09	68,49±1,07	68,49±0,88	68,64±1,50	67,88±1,56	69,92±1,45	68,37±1,16	68,03±1,13
	SVM	80,12±4,16	79,48±7,08	72,10±13,79	75,42±7,88	71,96±9,33	69,06±11,49	72,68±7,43	70,26±12,15	69,15±19,63	73,61±20,31	67,11±19,33	79,54±11,10	80,12±4,16	71,85±16,76	69,73±18,48	79,41±4,91	70,27±17,31
PhishingWeb	**J48**	95,88±0,44	95,77±0,79	95,69±0,48	96,40±0,71	95,88±0,44	96,19±1,05	95,50±0,70	96,17±0,58	96,51±0,46	96,28±0,71	95,89±0,68	96,23±0,94	95,88±0,44	95,84±0,45	96,26±0,81	96,20±0,45	95,88±0,44
	NB	92,98±0,64	93,00±0,70	92,62±1,19	93,46±0,92	92,98±0,64	93,26±0,78	92,07±0,90	92,99±0,72	93,29±0,58	93,14±0,78	92,90±0,88	93,03±1,05	92,98±0,64	92,62±0,71	93,12±0,98	92,67±0,80	92,98±0,64
	SVM	93,81±0,77	93,99±0,68	93,50±1,21	94,37±0,68	93,81±0,77	94,23±0,74	93,53±0,58	93,89±0,62	94,23±0,61	94,06±0,48	93,90±0,89	93,94±0,72	93,81±0,77	93,74±0,46	94,16±0,88	93,74±1,00	93,81±0,77
SkinSegmentation	**J48**	99,83±0,08	99,93±0,06	99,83±0,09	99,89±0,10	99,83±0,08	99,90±0,07	99,80±0,10	99,83±0,09	99,87±0,05	99,86±0,08	99,82±0,11	99,85±0,07	99,83±0,08	99,83±0,07	99,76±0,10	99,61±0,22	99,80±0,08
	NB	79,59±0,59	86,79±0,34	78,88±0,90	80,45±1,11	79,59±0,59	78,80±1,13	78,23±1,08	79,20±0,63	79,27±0,68	83,12±0,60	79,49±0,66	79,26±0,59	79,59±0,59	92,94±0,43	79,82±1,09	71,60±1,79	85,73±0,45
	SVM	65,24±22,83	73,18±27,69	70,55±20,13	64,97±15,79	65,24±22,83	70,33±23,31	68,91±25,98	67,04±24,84	70,53±24,64	70,00±23,61	51,82±21,79	82,51±15,98	65,24±22,83	73,28±27,17	66,18±11,09	69,25±10,97	67,12±22,21
SpamBase	**J48**	66,39±1,25	65,98±0,99	66,39±1,25	60,15±5,13	66,39±1,25	66,00±0,74	66,39±1,25	66,63±0,81	66,02±0,81	66,30±0,87	66,49±0,79	66,28±0,66	66,30±0,77	65,07±0,57	84,45±2,23	66,39±1,25	66,39±1,25
	NB	66,52±1,10	65,94±1,03	66,52±1,10	61,88±5,76	66,52±1,10	66,09±0,82	66,52±1,10	66,70±1,06	66,08±0,80	66,35±0,99	66,65±0,83	66,41±0,62	66,52±0,90	65,22±0,63	83,86±2,90	66,52±1,10	66,52±1,10
	SVM	66,59±1,02	65,92±1,04	66,59±1,02	61,54±4,29	66,59±1,02	65,99±0,75	66,59±1,02	66,82±0,85	66,06±0,79	66,54±1,00	66,67±0,63	66,45±0,59	66,63±0,67	65,20±0,59	85,37±1,98	66,59±1,02	66,59±1,02
HRSSAnomalous	**J48**	97,23±0,30	97,30±0,43	96,71±0,51	97,30±0,30	97,28±0,40	97,50±0,48	97,23±0,40	97,38±0,23	97,22±0,30	97,41±0,43	97,43±0,29	97,49±0,34	97,23±0,30	97,23±0,32	98,03±0,31	97,03±0,30	97,12±0,45
	NB	72,60±0,96	71,02±0,98	72,45±0,71	72,65±0,86	72,49±0,88	72,19±1,14	71,92±1,05	72,50±0,62	72,40±1,03	72,49±1,12	72,53±0,83	72,54±0,69	72,60±0,96	70,97±0,94	71,78±1,21	66,52±1,08	69,26±1,14
	SVM	56,49±5,90	59,12±8,95	59,67±7,24	55,60±7,68	56,83±4,81	58,19±7,45	58,11±9,45	57,51±4,19	59,95±6,91	55,66±7,59	55,16±8,01	57,82±5,56	56,49±5,90	57,37±8,13	60,53±6,50	56,84±6,18	55,38±6,93

Tabela 9 – Desempenho geral dos classificadores antes e após o pré-processamento das instâncias *outliers*, identificadas pelos algoritmos de DO.

Bases	C	Melhorou (16 casos)	Piorou (16 casos)
Amazon	**J48**	18,75	12,50
	NB	12,50	0,00
	SVM	12,50	0,00
Total		14,58	4,17
Breast Cancer	**J48**	12,50	25,00
	NB	6,25	0,00
	SVM	12,50	6,25
Total		10,42	10,42
Credicard	**J48**	31,25	0,00
	NB	18,75	0,00
	SVM	18,75	0,00
Total		22,92	0,00
Default Credit	**J48**	0,00	6,25
	NB	12,50	12,50
	SVM	0,00	37,50
Total		4,17	18,75
EEGEyeState	**J48**	0,00	0,00
	NB	6,25	0,00
	SVM	0,00	0,00
Total		2,08	0,00
KDDCup99	**J48**	12,50	0,00
	NB	12,50	6,25
	SVM	25,00	0,00
Total		16,67	2,08
Letter	**J48**	6,25	6,25
	NB	25,00	0,00
	SVM	12,50	0,00
Total		14,58	2,08
MagicGamma	**J48**	0,00	0,00
	NB	25,00	6,25
	SVM	0,00	0,00
Total		8,33	2,08
Mammography	**J48**	6,25	6,25
	NB	0,00	0,00
	SVM	6,25	6,25
Total		4,17	4,17
Mozilla4	**J48**	0,00	0,00
	NB	0,00	0,00
	SVM	0,00	0,00
Total		0,00	0,00
PhishingWeb	**J48**	6,25	0,00
	NB	6,25	0,00
	SVM	0,00	0,00
Total		4,17	0,00
SkinSegmentation	**J48**	6,25	12,50
	NB	25,00	12,50
	SVM	0,00	0,00
Total		10,42	8,33
SpamBase	**J48**	6,25	12,50
	NB	6,25	6,25
	SVM	6,25	12,50
Total		6,25	10,42
HRSSAnomalous	**J48**	12,50	6,25
	NB	0,00	31,25
	SVM	0,00	0,00
Total		4,17	12,50

Tabela 10 – Taxa de melhoria e piora após a fase de pré-processamento. Correlação de classificadores com bases e taxa total por base.

Para um melhor entendimento dos resultados referente a Tabela 8, podemos sumarizar o desempenho por técnica. Ela apresenta um resumo quantitativo de melhora ou piora significativa por técnica, de um total de 42 casos (3 algoritmos indutores X 14 bases). É possível observar também em quantos casos o algoritmo não identificou *outliers* (Coluna "*Outliers*"), nessa situação são 14 casos (14 Bases). Na coluna "Vitória", está registrado os casos em que a técnica foi melhor estatisticamente que as outras, isso quando houve melhoria de desempenho, essa situação somam-se 28 casos. Todos os resultados de melhora, piora e vitórias, foram computados segundo o teste estatístico, quando a hipótese nula foi rejeitada.

Na observação da Tabela 8, Os algoritmos *KNNoutlier*, *ALOCI* e *COP* destacam-se como as técnicas que mais provocaram efeito positivo no desempenho dos classificadores. De maneira geral, quando comparado com as outras técnicas, eles foram mais efetivos no processo de identificar instâncias *outliers* que afetam a performance dos classificadores. Esse mesmo efeito podemos ver na coluna "Melhor", onde eles também se destacam com maior taxa de vitórias, ou seja, foram melhores que outras técnicas quando houve melhora no desempenho. Por outro lado, os algoritmos *ODIN* e *GaussianModel* evidenciaram um efeito oposto. No que tange ao desempenho por classificador, a Tabela 11 apresenta o classificador NB com maior taxa de melhoria. Podemos ver o reflexo desses resultados na Tabela 12, que demonstra que as técnicas com melhores resultados, provocaram um melhor desempenho justamente no classificador NB.

(%)	**Melhorou (224 casos)**	**Piorou (224 casos)**
J48	8,48	6,25
NB	11,16	5,36
SVM	6,70	4,46

Tabela 11 – Taxa de melhoria e piora após a fase de pré-processamento. Resultado total por classificador.

Melhora (%)	**KNN**	**ODN**	**OPT**	**DWOF**	**LOF**	**ALC**	**COF**	**IFL**	**KDS**	**LDF**	**LOP**	**EM**	**KMS**	**COP**	**GM**	**INTl**
J48	28,57	0,00	7,14	0,00	7,14	7,14	7,14	7,14	0,00	0,00	14,29	0,00	14,29	28,57	0,00	14,29
NB	35,71	0,00	0,00	0,00	0,00	35,71	0,00	7,14	7,14	0,00	14,29	0,00	14,29	28,57	7,14	28,57
SVM	21,43	0,00	7,14	0,00	7,14	21,43	7,14	7,14	0,00	0,00	0,00	0,00	7,14	21,43	0,00	7,14

Tabela 12 – Taxa de melhoria após a fase de pré-processamento. Correlação classificador com algoritmos de DO.

Piora (%)	**KNN**	**ODN**	**OPT**	**DWOF**	**LOF**	**ALC**	**COF**	**IFL**	**KDS**	**LDF**	**LOP**	**EM**	**KMS**	**COP**	**GM**	**INTl**
J48	7,14	35,71	14,29	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	7,14	14,29	21,43	7,14
NB	7,14	7,14	0,00	0,00	0,00	14,29	0,00	0,00	0,00	0,00	0,00	0,00	14,29	0,00	28,57	14,29
SVM	7,14	7,14	14,29	0,00	0,00	7,14	0,00	0,00	0,00	7,14	0,00	0,00	14,29	0,00	7,14	7,14

Tabela 13 – Taxa de piora após a fase de pré-processamento. Correlação classificador com algoritmos de DO.

O algoritmo *KNNOutlier* classifica cada instância baseando-se na distância com seu vizinho mais próximo. O produto desse cálculo é um *ranking*, onde as instâncias com maior pontuação são candidatas a *outliers* (RAMASWAMY; RASTOGI; SHIM, 2000). Se observarmos a Tabela 14, é possível constatar que o *KNNOutlier* foi melhor nas bases *Breast Cancer*, *Letter* e *SkinSegmentation*. Essas bases compartilham algumas características, como um número alto de instâncias (>20000), tipo de dados multivariados e todas elas bases balanceadas. Além disso, a base *Letter* onde ele teve maior rendimento, tem um alto número de rótulos (26 classes), indicando uma certa adaptabilidade do algoritmo no que tange o número de rótulos. O *KNNOutlier* foi também um algoritmo que teve uma média de *outliers* identificados muito baixa (Ver Tabela 6), indicando que seu desempenho também foi eficiente. Os resultados experimentais dos trabalhos de Yang e Huang (2008b) e de Ramaswamy, Rastogi e Shim (2000) que estudaram a aplicação deste algoritmo, indicaram que o *KNNOutlier* em conjunto de dados sintéticos, tem um alto grau de adaptabilidade tanto no tamanho do conjunto, quanto na dimensionalidade. Em grandes bases de dados reais também demonstrou-se ser eficiente.

BASES	KNN	ODN	OPT	DWF	LOF	ALC	COF	IFL	KDS	LDF	LOP	EM	KMS	COP	GM	INT
Amazon	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	100,00	0,00	0,00
Breast Cancer	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00
Credicard	33,33	0,00	66,67	0,00	33,33	33,33	66,67	33,33	0,00	0,00	66,67	0,00	0,00	0,00	0,00	33,33
Default Credit	33,33	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EEGEyeState	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00
KDDCup99	33,33	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00	100,00
Letter	100,00	0,00	0,00	0,00	0,00	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	33,33
MagicGama	33,33	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00	33,33
Mammography	33,33	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mozilla4	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PhishingWeb	0,00	0,00	0,00	0,00	0,00	0,00	0,00	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
SkinSegmentation	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	33,33	0,00	0,00	33,33
SpamBase	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00
HRSSAnomalous	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Tabela 14 – Taxa de melhoria após a fase de pré-processamento. Correlação de algoritmos de DO com bases

BASES	KNN	ODN	OPT	DWF	LOF	ALC	COF	IFL	KDS	LDF	LOP	EM	KMS	COP	GM	INT
Amazon	33,33	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00
Breast Cancer	0,00	33,33	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	33,33	0,00
Credicard	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Default Credit	33,33	0,00	0,00	0,00	0,00	33,33	0,00	0,00	0,00	33,33	0,00	0,00	33,33	0,00	66,67	100,00
EEGEyeState	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
KDDCup99	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Letter	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MagicGama	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00
Mammography	0,00	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mozilla4	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PhishingWeb	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
SkinSegmentation	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	33,33	66,67	0,00
SpamBase	0,00	0,00	66,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00
HRSSAnomalous	33,33	33,33	0,00	0,00	0,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	33,33	0,00	33,33	33,33

Tabela 15 – Taxa de piora após a fase de pré-processamento. Correlação de algoritmos de DO com bases.

O algoritmo ALOCI também destaca-se pelo seu desempenho. Baseado em densi-

dade, é uma derivação do algoritmo LOCI. Ele computa com aproximação a detecção de *outliers*, tornando-o mais rápido que o LOCI. Apesar de também usar medidas de distância em seu cálculo, ele introduz o Fator de Desvio de Multi-Granularidade (MDEF), que lida com as variações de densidade locais no espaço de dados, e detecta valores anômalos isolados como agrupamentos periféricos (PAPADIMITRIOU *et al.*, 2003). Apesar de seu desempenho relevante, essa sua última característica pode ter afetado-o negativamente. A *Amazon*, é uma das bases no qual teve melhor desempenho (Ver Tabela 14), esse resultado vem junto com fato dele ter identificado toda uma classe como *outlier*, transformando este resultado em um falso desempenho. A *Amazon* é uma base altamente desbalanceada, a classe no qual o ALOCI identificou com *outlier* tem um número muito baixo de instâncias, podendo ser visto pelo algoritmo como um agrupamento periférico isolado.

Outro algoritmo com bom desempenho foi o COP. Basicamente, o COP calcula a probabilidade local de uma instância ter sido originada por um mecanismo diferente das demais. A ideia geral que envolve essa abordagem, é a possibilidade de dependências entre diferentes atributos. Diferentes mecanismos geram diferentes dependências entre os atributos. O COP considera correlações locais nos dados para produzir um modelo para cada instância, um sistema de equações calcula se a instância é um candidato a *outlier* (KRIEGEL *et al.*, 2012). Apesar do bom desempenho do COP nas bases *Amazon*,*KDDCup* e *SpamBase*, ele também identificou várias classes da base *KDDCup* como *outlier*, e mais da metade da instâncias da base *SpamBase*. A própria característica do COP teve impacto nesses resultados. A KDDCup é uma base com muitos rótulos, derivados de exemplos de tipos de invasões em redes de computadores, que por sua vez são gerados por mecanismos diferentes. Nessa base, existem classes com poucos exemplos, que possivelmente pode estar distribuído perto de instâncias de outra classes. Outro aspecto é o alto número de atributos das bases *KDDCup* e *SpamBase*, onde essa característica torna a probabilidade de existir diferentes dependências entre atributos maior. Apesar disso, o resultado de instâncias *outliers* identificadas pelo COP na base *SpamBase*, não ocasionou desbalanceamento no pré-processamento, e seu bom desempenho na base *Amazon*, demonstra sua boa performance para bases desbalanceadas.

Os algoritmos ODIN e *GaussianModel* tiveram os piores resultados. O ODIN é um algoritmo baseado em distância, que considera o grafo kNN como um grafo de proximidade direcionado, onde os vetores são vértices do grafo e as arestas são distâncias entre os vetores. Classificamos um vetor como *outlier* baseando-se em seu valor no grafo. Além dessa estratégia, o ODIN classifica todos os vetores com base suas distâncias médias de kNN, no qual um limiar global é definido. Vetores com uma grande distância kNN média são candidatos a *outliers* (ISMO *et al.*, 2004). Seguindo essa lógica, o ODIN poderia ter obtido um desempenho semelhante ao *KNNOutlier*, visto que usa de suas mesmas premissas e parâmetros. O trabalho de Campos *et al.* (2016), que compara diversas técnicas não-supervisionadas para detecção de *outliers*, cita o ODIN com menor eficácia quando

comparado com o COF, INFLO e LOF. Da mesma forma, o *GaussianModel* não teve bom desempenho. Trata-se de uma técnica que ajusta um modelo gaussiano multivariado aos dados, e usa o fator de densidade de probabilidade para calcular uma pontuação *outlier* (SCHUBERT *et al.*, 2015). Vale salientar, que o *GaussianModel* teve seu pior desempenho para o classificador NB e o ODIN para o J48 (Ver Tabela 13).

			$Acc_1(\%)$		
Bases	C	$Acc_0(\%)$	***J48***	***IBK***	***FR***
Amazon	**J48**	94.21±0.01	99.69±0.02	97.50±0.02	99.55±0.02
	NB	88.75±0.41	98.03±0.25	94.48±0.50	97.30±0.20
	SVM	94.94±0.21	100.00±0.01	99.57±0.11	99.99±0.01
Breast Cancer	**J48**	98.17±0.18	99.98±0.03	99.39±0.10	99.66±0.08
	NB	95.50±0.19	96.64±0.36	96.35±0.22	96.11±0.43
	SVM	97.96±0.21	99.66±0.08	99.48±0.10	99.17±0.12
Credicard	**J48**	99.88±0.05	100.00±0.01	99.94±0.04	99.97±0.04
	NB	97.35±0.30	97.58±0.28	97.55±0.42	97.44±0.35
	SVM	94.70±11.23	100.00±0.00	90.33±9.35	88.82±13.03
Default Credit	**J48**	80.33±0.77	99.97±0.03	98.55±0.28	100.00±0.00
	NB	69.35±4.92	94.27±0.83	69.40±0.91	93.18±0.76
	SVM	60.13±4.65	50.93±17.29	61.98±14.87	54.98±12.44
EEGEyeState	**J48**	84.50±1.09	99.93±0.10	90.49±1.30	99.69±0.17
	NB	46.77±3.26	80.55±1.33	46.89±6.59	75.22±1.23
	SVM	63.91±0.97	82.58±1.32	68.52±1.25	78.56±1.40
KDDCup99	**J48**	99.59±0.10	99.59±0.10	99.78±0.05	99.89±0.08
	NB	94.22±0.54	94.22±0.54	95.24±0.53	97.12±0.19
	SVM	94.16±3.09	94.16±3.09	98.01±1.81	99.71±0.30
Letter	**J48**	87.98±0.51	99.19±0.43	91.49±0.66	98.50±0.28
	NB	64.11±0.76	93.57±0.70	70.12±0.98	89.17±1.00
	SVM	85.10±1.01	99.86±0.11	90.54±0.39	99.30±0.24
MagicGamma	**J48**	85.06±0.98	100.00±0.00	97.78±0.31	99.71±0.10
	NB	72.69±0.87	95.38±0.89	85.73±0.77	84.16±1.01
	SVM	65.05±20.66	98.59±0.39	92.66±0.49	88.99±5.01
Mammography	**J48**	98.57±0.26	100.00±0.00	99.69±0.16	99.98±0.04
	NB	95.67±0.30	98.25±0.34	97.24±0.52	97.35±0.52
	SVM	98.20±0.28	99.90±0.12	99.63±0.18	99.67±0.20
Mozilla4	**J48**	94.80±0.60	100.00±0.00	98.46±0.24	99.99±0.03
	NB	68.64±1.50	77.07±0.94	75.49±1.07	83.88±1.36
	SVM	80.12±4.16	74.64±13.37	82.57±9.32	86.72±15.33
PhishingWeb	**J48**	95.88±0.44	99.88±0.09	98.16±0.45	99.78±0.17
	NB	92.98±0.64	94.69±0.53	92.72±0.59	95.08±1.00
	SVM	93.81±0.77	97.29±0.46	95.31±0.58	97.63±0.70
SkinSegmentation	**J48**	99.83±0.08	99.96±0.04	99.87±0.09	99.89±0.05
	NB	79.59±0.59	81.03±0.84	79.75±0.88	80.00±1.14
	SVM	65.24±22.83	76.73±20.64	64.52±25.42	79.97±21.36
SpamBase	**J48**	66.39±1.25	99.92±0.08	99.23±0.50	99.97±0.06
	NB	66.52±1.10	99.74±0.19	99.24±0.28	99.63±0.24
	SVM	66.59±1.02	99.97±0.06	99.59±0.32	99.97±0.06
HRSSAnomalous	**J48**	97.23±0.30	99.97±0.03	98.46±0.26	99.87±0.09
	NB	72.60±0.96	82.78±0.93	72.35±1.02	79.32±0.80
	SVM	56.49±5.90	79.42±9.56	57.11±5.59	84.15±9.69

Tabela 16 – Desempenho geral dos classificadores antes e após o pré-processamento das instâncias *outliers*, identificadas pelos filtros de classificação.

Diferentemente dos algoritmos de DO, os filtros de classificação obtiveram uma maior média de instâncias *outliers* identificadas, isso está diretamente relacionado a sua performance como classificador. Os filtros de classificação consideram como *outliers*, as instâncias incorretamente rotuladas pelo classificador. Aqui, foi importante definir indutores diferentes dos usados para medir o desempenho das técnicas de DO. Ao contrário disso, estaríamos levando a uma situação em que modelos de classificação estariam sendo estimados com dados de treinamento, gerando resultados otimistas. Todos os filtros foram estimados usando validação cruzada de 10-*folds*.

A Tabela 16 apresenta o desempenho geral dos classificadores antes e após o pré-processamento realizado pelos filtros de classificação. Apenas observando o desempenho absoluto, é possível perceber a diferença quando comparado com os resultados dos algoritmos de DO. Os resultados destacados com fundo cinza, indicam que o valor de Acc_1 quando comparado com Acc_0, apresentou melhoria significativa segundo o teste estatístico. Em nenhum caso os filtros de classificação provocaram efeito negativo no desempenho dos classificadores. Ao invés disso, na maioria dos casos houve efeito positivo. A Tabela 17, que sumariza os resultados da tabela supracitada, indica que o filtro de classificação FR teve maior quantidade de casos com efeito positivo significativo, com 88,10%. No entanto, nos casos em que houve melhora do desempenho, o filtro de classificação JRP teve melhor desempenho quando comparado com os outros filtros, obteve 86,84%. Se examinarmos o desempenho por classificador na Tabela 18, o J48 aparece com melhor desempenho, em 95,24 dos casos os filtros de classificação provocaram melhoria em sua acurácia.

(%)	**Melhorou (42 casos)**	**Piorou (42 casos)**	**Outliers (14 casos)**	**Vitória (38 casos)**
JRP	83,33	0,00	7,14	86,84
IBK	73,81	0,00	0,00	0,00
FR	88,10	0,00	0,00	36,84

Tabela 17 – Taxa de melhoria,piora e vitórias após a fase de pré-processamento. Taxa de casos em que não foi identificado *outliers* (Coluna "*Outliers*"). Resultado total por filtros de classificação.

(%)	**Melhorou (42 casos)**	**Piorou (42 casos)**
J48	95,24	0,00
NB	76,19	0,00
SVM	73,81	0,00

Tabela 18 – Taxa de melhoria e piora após a fase de pré-processamento. Resultado total por classificador.

A Tabela 19 demonstra a taxa de melhoria, relacionando os resultados das bases com os filtros de classificação. Observando essas duas tabelas, podemos destacar a base

SkinSegmentation com pior desempenho. A *SkinSegmentation* é uma base com tipo de dados multivariados, originalmente com 29592 instâncias, 4 atributos e 2 possíveis rótulos. O classificador j48 teve melhor desempenho em Acc_0 e também em Acc_1 (Ver Tabelas supracitadas). É possível que a própria característica da base tenha induzido a um melhor desempenho somente para o j48, que por sua vez é baseado em AD. Um fator que pode ocasionar um mau desempenho das AD, são atributos irrelevantes, isso pode ocasionar uma árvore maior que o necessário. Assim, podemos afirmar que quanto maior o número de atributos, maior a probabilidade de algum atributo ser irrelevante para a classificação. Características como baixo número de atributos e rótulos, propicia um ambiente favorável para as AD (TAN; STEINBACH; KUMAR, 2009).

BASES	JRP	IBK	FR
Amazon	100,00	100,00	100,00
Breast Cancer	100,00	100,00	100,00
Credicard	66,67	66,67	66,67
Default Credit	66,67	33,33	66,67
EEGEyeState	100,00	66,67	100,00
KDDCup99	0,00	100,00	100,00
Letter	100,00	100,00	100,00
MagicGama	100,00	100,00	100,00
Mammography	100,00	100,00	100,00
Mozilla4	66,67	66,67	66,67
PhishingWeb	100,00	66,67	100,00
SkinSegmentation	66,67	0,00	33,33
SpamBase	100,00	100,00	100,00
HRSSAnomalous	100,00	33,33	100,00

Tabela 19 – Taxa de melhoria após a fase de pré-processamento. Correlação de filtros de classificação com bases.

Melhora (%)	**JRP**	**IBK**	**FR**
J48	92,86	92,86	100,00
NB	85,71	57,14	85,71
SVM	71,43	71,43	78,57

Tabela 20 – Taxa de melhoria após a fase de pré-processamento. Correlação classificador com filtros de classificação.

De acordo com os resultados, os filtros de classificação JRP e FR alcançaram melhor desempenho. Os dois filtros são baseado em regras de decisão. O *JRipper* (JRP) examina as classes em tamanho crescente e um conjunto inicial de regras para casa classe é gerado. Utilizando esse resultado, o erro é reduzido incrementalmente, tratando todos os exemplos com decisões únicas. Por outro lado o FURIA (FR), é uma generalização do JRP, e ao invés de usar regras convencionais, o FR utiliza regras *fuzzy*. A semelhança dos dois filtros

reflete no bom desempenho dos dois nos resultados. Em contrapartida, nas Tabelas 18 e 20, podemos observar que o classificador com melhor performance foi o J48. É aceitável afirmar que o desempenho dos filtros de classificação (JRP e FR), fundamentam-se no fato de que árvores de decisão e classificadores baseados em regras de decisão, compartilham semelhanças no processo de indução, e os resultados sejam produto de classificadores estimados com conjunto de treinamento.

5.1 Considerações Finais

O objetivo deste capítulo foi apresentar os resultados experimentais, e obter uma visão geral de como a metodologia proposta compara as diversas técnicas de DO. Este experimento seguiu restritamente o que foi definido e discutido amplamente na seção de materiais e métodos 4.2, e todos os valores dos parâmetros dos algoritmos indutores seguiu o estabelecido por padrão no *WEKA*.

Os algoritmos de DO desempenharam um papel positivo na metodologia proposta, apesar de terem obtido um desempenho inferior quando comparado com os filtros de classificação, houveram mais casos com melhora do que efeito negativo pôde ser observado (Ver Tabela 10), e identificaram menor número de instâncias *outliers*, indicando um desempenho eficiente. Da mesma maneira, os filtros de classificação obtiveram desempenho positivo e se consolidam como uma área de pesquisa que ainda precisa ser explorada.

De acordo com os resultados, ficou claro que existe um limiar entre taxa de identificação de *outliers* e o desempenho das técnicas. Os filtros de classificação identificaram mais instâncias anômalas e consequentemente obtiveram melhor desempenho. De outra forma, destacamos aqui o *KNNOutlier* por sua boa performance e baixa taxa de identificação *outliers*, apresentando bons resultado principalmente em bases com alto número de instâncias e rótulos. O ALOCI teve seu desempenho destacado para bases desbalanceadas e o COP para conjunto com alto número de atributos. Podemos identificar também, bases que tiveram mais casos de melhora, como a base *Credicard* com os algoritmos de DO, e com desempenho inferior, como a base *SkinSegmentation* com os filtros de classificação. Do mesmo modo, classificadores com melhores performances, como foi o caso do NB para os algoritmos de DO, e o J48 para os filtros de classificação.

De maneira geral, os resultados experimentais permitiram afirmar, que o estudo comparativo através da metodologia proposta, possibilitou uma analisar o desempenho de diversas técnicas de DO, para diferentes domínios, de forma não pareada e objetiva. Algumas lacunas ainda precisam ser preenchidas por trabalhos futuros, direcionado para o ajuste dos parâmetros das técnicas, limiar da quantidade de remoção de *outliers* e desempenho dos classificadores, e algoritmos indutores que melhor se adaptam como filtros de classificação.

6 Considerações Finais

A identificação e o tratamento de *outliers* é um tema bastante estudado e explorado, e tem uma enorme gama de aplicações. A importância do tema motivou a comunidade científica ao desenvolvimento de diversas técnicas, voltadas para diferentes tipos de contextos. Diante de tantos métodos para DO, não é trivial saber qual técnica tende a ser mais eficaz em determinado domínio. Por outro lado, todas compartilham do mesmo objetivo: Identificar dados anômalos. Se o *outlier* vai ser tratado como uma informação valiosa ou como um dado ruidoso que precisa ser removido, é uma questão de aplicação. Por se tratar de um problema, na maioria dos casos, não supervisionado, o grande desafio é comparar técnicas no que tange a sua capacidade em identificar *outliers*. Partindo disso, uma análise comparativa que venha mensurar o desempenho de diferentes técnicas de DO, tal que proporcione uma avaliação objetiva de cada técnica, configura-se uma importante contribuição para o problema de DO. Este trabalho propõe um estudo comparativo de técnicas de DO inseridas em diferentes abordagens e métodos, utilizando uma metodologia baseada na classificação de dados.

O referencial teórico abordou os principais conceitos que envolvem a proposta deste trabalho, desde a concepção geral de *outliers*, abordando suas causas e os principais métodos para sua detecção, até o problema de classificação, definindo os conceitos gerais e como se estima e avalia diferentes modelos. O Capítulo 4 foi dedicado a apresentação da metodologia utilizada para comparar técnicas de DO, definindo os componentes envolvidos e relacionando-os. Esse capítulo foi destinado também à apresentação dos materiais e métodos que compõem os experimentos, determinando os conjuntos de dados, as técnicas de DO e os classificadores envolvidos, bem como as métricas e ferramentas usadas para a experimentação.

O Capítulo 5 apresentou e discutiu os resultados do experimento realizado. Nele, foi possível observar que a metodologia apresentou-se como uma estratégia válida, não se restringindo somente na análise de desempenho das técnicas de DO, mas também de bases e algoritmos indutores. Diante os objetivos estabelecidos, foi possível quantificar o desempenho das técnicas de DO através da metodologia proposta, de maneira objetiva. Foi identificado características da bases que influenciaram o desempenho das técnicas. Constatou-se também que existe uma relação do total de *outliers* identificados pelas técnicas e removidos na fase de pre-processamento, com a melhoria no desempenho dos classificadores. No caso dos filtros de classificação, identificam um número maior de *outliers*, e por sua vez provocaram um efeito positivo maior na acurácia dos classificadores. Dessa forma, o estudo comparativo permitiu não só, identificar técnicas com maior eficácia, mas também técnicas com maior eficiência. Por fim, possibilitou também identificar o algoritmo

indutor que obteve melhor acurácia após a etapa de pré-processamento.

Por outro lado, não foi possível relacionar a melhor técnica de DO a determinado domínio, ficando assim como um objetivo não alcançado por este trabalho. Logo, essa premissa pode se unir a outras lacunas que ainda precisam ser preenchidas por trabalhos futuros. A eficácia das técnicas de DO são fortemente dependentes dos valores dos parâmetros de entrada do algoritmo, que por sua vez variam de acordo com a distribuição dos dados do conjunto. Um trabalho que explore uma melhor configuração nos parâmetros das técnicas, poderia vir a produzir melhores resultados. Como já discutido, foi possível observar uma influência da quantidade de *outliers* removidos com os resultados obtidos. Um estudo exploratório sobre limiar crítico para identificação de *outliers*, poderia estabelecer um melhor ajuste e definir qual seria o melhor limiar para a metodologia.

Os resultados possibilitaram observar que existe uma tendência, quanto maior a identificação e remoção de *outliers* por parte das técnicas de DO, maior o desempenho dos classificadores. Um possível trabalho futuro, poderia estar designado em identificar como essa relação ocorre, qual é o limite entre remoção de *outliers* e consequentemente melhorar o desempenho dos classificadores, sem que ocasione desfalque de instâncias ou desbalanceamento do conjuntos de dados. Foi visto também que o desempenho dos filtros de classificação como técnicas de DO supervisionadas, está diretamente relacionado com seu desempenho como classificador. É preciso um estudo exploratório para identificar a relação do melhor filtro para determinado algoritmo indutor, podendo vir a se tornar uma valiosa contribuição para a área de classificação de dados.

Métodos generalistas para relacionar o desempenho de técnicas para determinados domínios de aplicação, são escassos ou inexistentes na literatura. Um estudo comparativo que possibilite tal análise, permite não só confrontar diferentes técnicas de DO, mas também estabelecer estratégias para melhorar a qualidade da detecção *outliers*. O processo de detectar e consequentemente remover as instâncias anômalas, afeta o desempenho de classificadores. Quantificar esse efeito como fator de desempenho para as técnicas de DO, permitiu que análises fossem realizadas, examinando a eficácia das técnicas para determinada situação. Podemos afirmar, que a efetividade do estudo comparativo proposto, está na viabilidade de uma avaliação objetiva e uma observação mais clara do desempenho de cada técnica.

Referências

AGGARWAL, C. C. *Data mining: the textbook*. [S.l.]: Springer, 2015. Citado 3 vezes nas páginas 19, 62 e 71.

AGRAWAL, S.; AGRAWAL, J. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, Elsevier, v. 60, p. 708–713, 2015. Citado 2 vezes nas páginas 56 e 57.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine learning*, Springer, v. 6, n. 1, p. 37–66, 1991. Citado na página 72.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2014. Citado 3 vezes nas páginas 20, 46 e 62.

ALTMAN, E. I.; DANOVI, A.; FALINI, A. Z-score models' application to italian companies subject to extraordinary administration. 2015. Citado na página 65.

AMOR, N. B.; BENFERHAT, S.; ELOUEDI, Z. Naive bayes vs decision trees in intrusion detection systems. In: ACM. *Proceedings of the 2004 ACM symposium on Applied computing*. [S.l.], 2004. p. 420–424. Citado 2 vezes nas páginas 56 e 61.

ANSCOMBE, F. J. Rejection of outliers. *Technometrics*, Taylor & Francis Group, v. 2, n. 2, p. 123–146, 1960. Citado na página 32.

ASUNCION, A.; NEWMAN, D. *UCI machine learning repository*. 2007. Citado 2 vezes nas páginas 67 e 68.

AUGUSTEIJN, M.; FOLKERT, B. Neural network classification and novelty detection. *International Journal of Remote Sensing*, Taylor & Francis, v. 23, n. 14, p. 2891–2902, 2002. Citado na página 28.

BAKAR, Z. A. *et al.* A comparative study for outlier detection techniques in data mining. In: IEEE. *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*. [S.l.], 2006. p. 1–6. Citado 2 vezes nas páginas 60 e 61.

BANERJEE, A.; BURLINA, P.; DIEHL, C. A support vector method for anomaly detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, v. 44, n. 8, p. 2282–2291, Aug 2006. ISSN 0196-2892. Citado na página 28.

BARBARá, D.; DOMENICONI, C.; ROGERS, J. P. Detecting outliers using transduction and statistical testing. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2006. (KDD '06), p. 55–64. ISBN 1-59593-339-5. Disponível em: <http://doi.acm.org/10.1145/1150402.1150413>. Citado 2 vezes nas páginas 57 e 61.

BASU, S.; BILENKO, M.; MOONEY, R. J. A probabilistic framework for semi-supervised clustering. In: ACM. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2004. p. 59–68. Citado na página 44.

BENTLEY, P. J. *et al.* Fuzzy darwinian detection of credit card fraud. In: *the 14th Annual Fall Symposium of the Korean Information Processing Society*. [S.l.: s.n.], 2000. v. 14. Citado 2 vezes nas páginas 57 e 61.

BHATT, R. B. *et al.* Efficient skin region segmentation using low complexity fuzzy decision tree model. In: IEEE. *India Conference (INDICON), 2009 Annual IEEE*. [S.l.], 2009. p. 1–4. Citado na página 68.

BHUYAN, M. H.; BHATTACHARYYA, D. K.; KALITA, J. K. Network anomaly detection: methods, systems and tools. *IEEE communications surveys & tutorials*, IEEE, v. 16, n. 1, p. 303–336, 2014. Citado na página 28.

BREUNIG, M. M. *et al.* Optics-of: Identifying local outliers. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 1999. p. 262–270. Citado na página 69.

BREUNIG, M. M. *et al.* Lof: identifying density-based local outliers. In: ACM. *ACM sigmod record*. [S.l.], 2000. v. 29, n. 2, p. 93–104. Citado 7 vezes nas páginas 9, 41, 42, 43, 44, 58 e 69.

CAMPOS, G. O. *et al.* On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, Springer, v. 30, n. 4, p. 891–927, 2016. Citado 5 vezes nas páginas 60, 61, 70, 71 e 84.

CATENI, S.; COLLA, V.; VANNUCCI, M. Outlier detection methods for industrial applications. In: *Advances in Robotics, Automation and Control*. [S.l.]: InTech, 2008. Citado na página 28.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 15, 2009. Citado 22 vezes nas páginas 9, 16, 17, 21, 22, 23, 26, 27, 28, 29, 30, 31, 33, 34, 37, 38, 39, 41, 44, 56, 57 e 60.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm, v. 2, n. 3, p. 27, 2011. Citado na página 73.

COX, E. *A fuzzy system for detecting anomalous behaviors in healthcare provider claims. Intelligent system for finance and business (pp. 111–134)*. [S.l.]: New York, NY: Wiley, 1995. Citado 2 vezes nas páginas 57 e 61.

CROOK, P. A. *et al.* A tale of two filters-on-line novelty detection. In: IEEE. *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*. [S.l.], 2002. v. 4, p. 3894–3899. Citado na página 29.

DANG, T. T.; NGAN, H. Y. T.; LIU, W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In: *2015 IEEE International Conference on Digital Signal Processing (DSP)*. [S.l.: s.n.], 2015. p. 507–510. ISSN 1546-1874. Citado na página 29.

DASGUPTA, D.; FORREST, S. Novelty detection in time series data using ideas from immunology. In: *Proceedings of the international conference on intelligent systems*. [S.l.: s.n.], 1996. p. 82–87. Citado na página 59.

DEBES, K.; KOENIG, A.; GROSS, H.-M. Transfer functions in artificial neural networks a simulation-based tutorial. *Brains, Minds and Media*, v. 2005, n. 1, 2005. Citado 2 vezes nas páginas 9 e 35.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006. Citado 2 vezes nas páginas 53 e 54.

DESFORGES, M.; JACOB, P.; COOPER, J. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, SAGE Publications Sage UK: London, England, v. 212, n. 8, p. 687–703, 1998. Citado na página 33.

DUNHAM, M. H. *Data mining: Introductory and advanced topics*. [S.l.]: Pearson Education India, 2006. Citado na página 72.

ESKIN, E. Anomaly detection over noisy data using learned probability distributions. In: CITESEER. *In Proceedings of the International Conference on Machine Learning*. [S.l.], 2000. Citado na página 33.

FACELI, K. *et al.* Cplf (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina. Grupo Gen-LTC*, 2011. Citado 2 vezes nas páginas 72 e 73.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, JSTOR, v. 11, n. 1, p. 86–92, 1940. Citado na página 54.

GAMBERGER, D.; LAVRAC, N.; GROSELJ, C. Experiments with noise filtering in a medical domain. In: *ICML*. [S.l.: s.n.], 1999. p. 143–151. Citado 2 vezes nas páginas 19 e 72.

GRUBBS, F. E. Procedures for detecting outlying observations in samples. *Technometrics*, Taylor & Francis, v. 11, n. 1, p. 1–21, 1969. Citado na página 34.

GUEVARA, C.; SANTOS, M.; LÓPEZ, V. Training strategy to improve the efficiency of an intelligent detection system. In: WORLD SCIENTIFIC. *Decision Making and Soft Computing: Proceedings of the 11th International FLINS Conference*. [S.l.], 2014. p. 544–549. Citado na página 66.

HAMPEL, F. R. *et al. Robust statistics: the approach based on influence functions*. [S.l.]: John Wiley & Sons, 2011. v. 196. Citado na página 59.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado 26 vezes nas páginas 9, 15, 16, 20, 21, 22, 24, 26, 29, 30, 31, 32, 33, 34, 37, 39, 40, 42, 45, 46, 47, 49, 50, 51, 52 e 62.

HAND, H. M. D.; SMYTH, P. *Principles of Data Mining*. [S.l.]: MIT press, 2001. Citado na página 49.

HAUTAMÄKI, V. *et al.* Improving k-means by outlier removal. In: SPRINGER. *Scandinavian Conference on Image Analysis*. [S.l.], 2005. p. 978–987. Citado na página 69.

HAWKINS, D. M. *Identification of outliers*. [S.l.]: Springer, 1980. v. 11. Citado na página 20.

HE, Y.; CHEN, M. A probabilistic, mechanism-indepedent outlier detection method for online experimentation. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. [S.l.: s.n.], 2017. p. 640–647. Citado na página 29.

HECK, D. *et al. CORSIKA: A Monte Carlo code to simulate extensive air showers*. [S.l.], 1998. Citado na página 67.

HODGE, V.; AUSTIN, J. A survey of outlier detection methodologies. *Artificial intelligence review*, Springer, v. 22, n. 2, p. 85–126, 2004. Citado 3 vezes nas páginas 17, 32 e 59.

HOLMES, G.; DONKIN, A.; WITTEN, I. H. Weka: A machine learning workbench. In: IEEE. *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. [S.l.], 1994. p. 357–361. Citado na página 73.

HRANISAVLJEVIC, N.; NIGGEMANN, O.; MAIER, A. A novel anomaly detection algorithm for hybrid production systems based on deep learning and timed automata. In: *International Workshop on the Principles of Diagnosis (DX)*. [S.l.: s.n.], 2016. Citado na página 68.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. Citado na página 60.

HÜHN, J.; HÜLLERMEIER, E. Furia: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, Springer, v. 19, n. 3, p. 293–319, 2009. Citado na página 72.

ISMO, K. *et al.* Outlier detection using k-nearest neighbour graph. In: IEEE. *null*. [S.l.], 2004. p. 430–433. Citado 2 vezes nas páginas 69 e 84.

JIN, W. *et al.* Ranking outliers using symmetric neighborhood relationship. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2006. p. 577–593. Citado 4 vezes nas páginas 43, 59, 61 e 70.

JR, C. E. K. *et al.* Construction of a bayesian network for mammographic diagnosis of breast cancer. *Computers in biology and medicine*, Pergamon, v. 27, n. 1, p. 19–29, 1997. Citado na página 66.

KNORR, E. M.; NG, R. T. A unified approach for mining outliers. In: IBM PRESS. *Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research*. [S.l.], 1997. p. 11. Citado na página 40.

KORU, A. G.; ZHANG, D.; LIU, H. Modeling the effect of size on defect proneness for open-source software. In: IEEE. *Predictor Models in Software Engineering, 2007. PROMISE'07: ICSE Workshops 2007. International Workshop on*. [S.l.], 2007. p. 10–10. Citado na página 67.

KOU, Y.; LU, C.-T.; CHEN, D. Spatial weighted outlier detection. In: SIAM. *Proceedings of the 2006 SIAM international conference on data mining*. [S.l.], 2006. p. 614–618. Citado na página 29.

KOU, Y. *et al.* Survey of fraud detection techniques. In: IEEE. *Networking, sensing and control, 2004 IEEE international conference on*. [S.l.], 2004. v. 2, p. 749–754. Citado na página 28.

KRIEGEL, H.-P. *et al.* Loop: local outlier probabilities. In: ACM. *Proceedings of the 18th ACM conference on Information and knowledge management*. [S.l.], 2009. p. 1649–1652. Citado 2 vezes nas páginas 43 e 71.

KRIEGEL, H.-P. *et al.* Outlier detection in arbitrarily oriented subspaces. In: IEEE. *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. [S.l.], 2012. p. 379–388. Citado 5 vezes nas páginas 9, 35, 36, 71 e 84.

KUMAR, M.; HANUMANTHAPPA, M.; KUMAR, T. S. Intrusion detection system using decision tree algorithm. In: IEEE. *Communication Technology (ICCT), 2012 IEEE 14th International Conference on*. [S.l.], 2012. p. 629–634. Citado 2 vezes nas páginas 56 e 61.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado 5 vezes nas páginas 9, 48, 49, 50 e 62.

LATECKI, L. J.; LAZAREVIC, A.; POKRAJAC, D. Outlier detection with kernel density functions. In: SPRINGER. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. [S.l.], 2007. p. 61–75. Citado 2 vezes nas páginas 58 e 61.

LEE, J. *et al.* Online process monitoring scheme for fault detection based on independent component analysis (ica) and local outlier factor (lof). In: *The 40th International Conference on Computers Indutrial Engineering*. [S.l.: s.n.], 2010. p. 1–6. Citado 2 vezes nas páginas 58 e 61.

LI, Y. *et al.* Network anomaly detection based on tcm-knn algorithm. In: *Proceedings of the 2Nd ACM Symposium on Information, Computer and Communications Security*. New York, NY, USA: ACM, 2007. (ASIACCS '07), p. 13–19. ISBN 1-59593-574-6. Disponível em: <http://doi.acm.org/10.1145/1229285.1229292>. Citado 2 vezes nas páginas 57 e 61.

LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 4, p. 18–36, 2009. Citado na página 58.

LIPPMANN, R. P. *et al.* Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In: IEEE. *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*. [S.l.], 2000. v. 2, p. 12–26. Citado na página 67.

MANDIGOBINDGARH, S. Survey paper on data mining techniques of intrusion detection. Citado na página 27.

MOHAMMAD, R. M.; THABTAH, F.; MCCLUSKEY, L. An assessment of features related to phishing websites using an automated technique. In: IEEE. *Internet Technology And Secured Transactions, 2012 International Conference for*. [S.l.], 2012. p. 492–497. Citado na página 68.

MOMTAZ, R.; MOHSSEN, N.; GOWAYYED, M. A. Dwof: A robust density-based outlier detection approach. In: SPRINGER. *Iberian Conference on Pattern Recognition and Image Analysis*. [S.l.], 2013. p. 517–525. Citado na página 69.

NIU, Z. *et al.* A survey of outlier detection methodologies and their applications. In: SPRINGER. *International Conference on Artificial Intelligence and Computational Intelligence*. [S.l.], 2011. p. 380–387. Citado na página 16.

PAMULA, R.; DEKA, J. K.; NANDI, S. An outlier detection method based on clustering. In: *2011 Second International Conference on Emerging Applications of Information Technology*. [S.l.: s.n.], 2011. p. 253–256. Citado 2 vezes nas páginas 57 e 61.

PANNU, H. S. *et al.* Anomaly detection survey for information security. In: ACM. *Proceedings of the 10th International Conference on Security of Information and Networks*. [S.l.], 2017. p. 251–258. Citado na página 27.

PAPADIMITRIOU, S. *et al.* Loci: fast outlier detection using the local correlation integral. In: *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*. [S.l.: s.n.], 2003. p. 315–326. Citado 5 vezes nas páginas 43, 58, 61, 70 e 84.

PATHAK, J.; VIDYARTHI, N.; SUMMERS, S. L. A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. *Managerial Auditing Journal*, Emerald Group Publishing Limited, v. 20, n. 6, p. 632–644, 2005. Citado 2 vezes nas páginas 57 e 61.

PATIL, T. R.; SHEREKAR, S. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, v. 6, n. 2, p. 256–261, 2013. Citado na página 73.

PHOHA, V. V. *Internet security dictionary*. [S.l.]: Springer Science & Business Media, 2007. Citado na página 27.

POZZOLO, A. D. *et al.* Calibrating probability with undersampling for unbalanced classification. In: IEEE. *Computational Intelligence, 2015 IEEE Symposium Series on*. [S.l.], 2015. p. 159–166. Citado na página 66.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986. Citado 2 vezes nas páginas 20 e 62.

RAJ, S. B. E.; PORTIA, A. A. Analysis on credit card fraud detection methods. In: *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*. [S.l.: s.n.], 2011. p. 152–156. Citado 2 vezes nas páginas 57 e 61.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. In: ACM. *ACM Sigmod Record*. [S.l.], 2000. v. 29, n. 2, p. 427–438. Citado na página 83.

RANA, P.; PAHUJA, D.; GAUTAM, R. A critical review on outlier detection techniques. *International Journal of Science and Research (IJSR) Volume*, v. 3, 2014. Citado 2 vezes nas páginas 16 e 28.

RIJN, J. N. V. *et al.* Openml: A collaborative science platform. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2013. p. 645–649. Citado na página 68.

ROBERTS, S.; TARASSENKO, L. A probabilistic resource allocating network for novelty detection. *Neural Computation*, MIT Press, v. 6, n. 2, p. 270–284, 1994. Citado 2 vezes nas páginas 59 e 61.

SANTHANAM, T.; PADMAVATHI, M. S. Comparison of k-means clustering and statistical outliers in reducing medical datasets. In: *2014 International Conference on Science Engineering and Management Research (ICSEMR)*. [S.l.: s.n.], 2014. p. 1–6. Citado 2 vezes nas páginas 58 e 61.

SCHUBERT, E. *et al.* A framework for clustering uncertain data. *PVLDB*, v. 8, n. 12, p. 1976–1979, 2015. Disponível em: <http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf>. Citado 3 vezes nas páginas 71, 73 e 85.

SCHUBERT, E.; ZIMEK, A.; KRIEGEL, H.-P. Generalized outlier detection with flexible kernel density estimates. In: SIAM. *Proceedings of the 2014 SIAM International Conference on Data Mining*. [S.l.], 2014. p. 542–550. Citado 2 vezes nas páginas 43 e 70.

SHESKIN, D. J. *Handbook of parametric and nonparametric statistical procedures*. [S.l.]: crc Press, 2003. Citado na página 54.

SINDHU, S. S. S.; GEETHA, S.; KANNAN, A. Decision tree based light weight intrusion detection using a wrapper approach. *Expert Systems with applications*, Elsevier, v. 39, n. 1, p. 129–141, 2012. Citado 2 vezes nas páginas 56 e 61.

SINGH, K.; UPADHYAYA, S. Outlier detection: applications and techniques. *International Journal of Computer Science Issues*, Citeseer, v. 9, n. 1, p. 307–323, 2012. Citado 6 vezes nas páginas 9, 24, 25, 26, 28 e 29.

STEINBACH, M. *et al.* Discovery of climate indices using clustering. In: ACM. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2003. p. 446–455. Citado na página 44.

STOLFO, S. J. *et al. Cost-based modeling for fraud and intrusion detection: Results from the JAM project*. [S.l.], 2000. Citado na página 67.

SU, X.; TSAI, C.-L. Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 1, n. 3, p. 261–268, 2011. Citado na página 22.

SUN, P.; CHAWLA, S.; ARUNASALAM, B. Mining for outliers in sequential databases. In: SIAM. *Proceedings of the 2006 SIAM International Conference on Data Mining*. [S.l.], 2006. p. 94–105. Citado na página 29.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciência Moderna, 2009. Citado 23 vezes nas páginas 9, 10, 15, 16, 20, 23, 28, 30, 31, 32, 34, 39, 41, 44, 45, 46, 47, 49, 51, 52, 53, 62 e 87.

TANG, J. *et al.* Enhancing effectiveness of outlier detections for low density patterns. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2002. p. 535–548. Citado 3 vezes nas páginas 26, 43 e 70.

THOTTAN, M.; JI, C. Anomaly detection in ip networks. *IEEE Transactions on signal processing*, IEEE, v. 51, n. 8, p. 2191–2204, 2003. Citado na página 28.

TING, J. A.; D'SOUZA, A.; SCHAAL, S. Automatic outlier detection: A bayesian approach. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2007. p. 2489–2494. ISSN 1050-4729. Citado na página 29.

TUKEY, J. W. *Exploratory data analysis*. [S.l.]: Reading, Mass., 1977. v. 2. Citado na página 59.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin*, JSTOR, v. 1, n. 6, p. 80–83, 1945. Citado na página 53.

WITTEN, I. H. *et al. Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 3 vezes nas páginas 52, 53 e 75.

WOODS, K. S. *et al.* Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 7, n. 06, p. 1417–1436, 1993. Citado na página 67.

WU, S. X.; BANZHAF, W. The use of computational intelligence in intrusion detection systems: A review. *Applied soft computing*, Elsevier, v. 10, n. 1, p. 1–35, 2010. Citado na página 16.

YANG, P.; HUANG, B. Knn based outlier detection algorithm in large dataset. In: IEEE. *Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on*. [S.l.], 2008. v. 1, p. 611–613. Citado na página 69.

YANG, P.; HUANG, B. Knn based outlier detection algorithm in large dataset. In: *2008 International Workshop on Education Technology and Training 2008 International Workshop on Geoscience and Remote Sensing*. [S.l.: s.n.], 2008. v. 1, p. 611–613. Citado na página 83.

YEH, I.-C.; LIEN, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, Elsevier, v. 36, n. 2, p. 2473–2480, 2009. Citado na página 66.

ZAMONER, F. W. *Técnica de aprendizado semissupervisionado para detecção de outliers*. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado 2 vezes nas páginas 16 e 57.

ZHANG, K.; HUTTER, M.; JIN, H. A new local distance-based outlier detection approach for scattered real-world data. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2009. p. 813–822. Citado 3 vezes nas páginas 43, 58 e 71.

ZHU, C. *et al.* Obe: Outlier by example. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2004. p. 222–234. Citado na página 70.

ZIMEK, A.; SCHUBERT, E.; KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 5, n. 5, p. 363–387, 2012. Citado na página 64.