Exploração das propriedades de *hubness* para detecção semissupervisionada de *outliers* em dados de alta dimensão

Lucimeire Alves da Silva

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2017

Lucimeire Alves da Silva

Exploração das propriedades de *hubness* para detecção semissupervisionada de *outliers* em dados de alta dimensão

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientadora: Maria Camila Nardini Barioni

Uberlândia
2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

S586e 2017

Silva, Lucimeire Alves da, 1990-

Exploração das propriedades de hubness para detecção semissupervisionada de outliers em dados de alta dimensão / Lucimeire Alves da Silva. - 2017.

88 f. : il.

Orientador: Maria Camila Nardini Barioni.

Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Ciência da Computação.

Disponível em: http://dx.doi.org/10.14393/ufu.di.2017.51

Inclui bibliografia.

1. Computação - Teses. 2. Mineração de dados (Computação) - Teses. 3. Inteligência artificial - Processamento de dados - Teses. I. Barioni, Maria Camila Nardini. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU: 681.3

Dedico este trabalho ao meu noivo, Luciano Lopes; aos meus pais, Rosemeire e Lucio; a minha irmã, Gabriela; aos meus avós, Maria e Ildefonso e aos meus tios, Vicelma e Edio, por todo apoio, amor e incentivo durante cada desafio.

Agradecimentos

Agradeço...

A Deus, por minha vida e por me dar sabedoria e perseverança para enfrentar todos os desafios e sempre ter forças para persistir no meu crescimento independente dos empecilhos encontrados.

Aos meus pais Rosemeire e Lucio pela minha educação, pela dedicação, apoio, confiança, carinho e amor em todos os momentos.

À minha sempre "irmãzinha" Gabriela que me motiva a ser sempre melhor do que sou, para ser sempre o seu exemplo favorito.

Aos meus avós Maria Izabel e Ildefonso pelo amparo e amor incondicionais em todos os momentos.

Aos meus tios e padrinhos Vicelma e Edio pela confiança no meu potencial e por serem sempre meus exemplos.

Ao meu noivo Luciano, com quem sempre posso contar verdadeiramente, por ter me apoiado tanto durante este projeto, principalmente emocionalmente. Você me deu muita força, tranquilidade, suporte, paz, alegria e amor. Te amo.

Aos meus colegas e amigos que fizeram parte dessa fase da minha vida, que direta e indiretamente contribuíram nesse objetivo da minha vida profissional.

Aos professores e funcionários do PPGCO-UFU, que são responsáveis pela manutenção e crescimento do curso, em especial ao secretário da PPGCO Erisvaldo que está sempre muito próximo aos alunos e de prontidão para nos auxiliar.

À CAPES e ao CNPq pela ajuda financeira durante 12 meses desse projeto.

À Professora Dr[a]. Sandra pelo apoio no meu ingresso no mestrado, por me apresentar uma das áreas que mais me interessa no âmbito da computação e pelo exemplo de dedicação absoluta a profissão.

Em especial, à minha orientadora Professora Dr[a]. Maria Camila pelo apoio, profissionalismo, tranquilidade e orientação em todos os momentos da realização deste trabalho. Acredito que a profissão de professor deve ser respeitada e glorificada, por ser o profissional responsável pelo incentivo e divulgação do conhecimento na sociedade, dessa forma muito obrigada por toda paciência e tempo dedicado em me ajudar nesse projeto e no meu crescimento profissional.

"Eu Acredito, que ás vezes são as pessoas que ninguém espera nada que fazem as coisas que ninguém consegue imaginar."

(Alan Turing)

Resumo

Com o crescente aumento da quantidade de dados armazenados, a área de mineração de dados tornou-se imprescindível para que seja possível manipular e extrair conhecimento a partir desses dados. Grande parte dos trabalhos nessa área focam em encontrar padrão nos dados, porém os dados fora do padrão (anomalias) também podem agregar muito no conhecimento do conjunto de dados em estudo. O estudo, o desenvolvimento e o aprimoramento de técnicas de detecção de *outliers* são objetivos importantes e têm se mostrado útil em diversos cenários, como: detecção de fraudes, detecção de intrusão e monitoramento de condições médicas entre outros. O trabalho apresentado aqui descreve um novo método para detecção semissupervisionada de *outliers* em dados com alta dimensionalidade. Os experimentos realizados com diversos conjuntos de dados reais indicam a superioridade do método proposto em relação aos métodos da literatura selecionados como linha de base.

Palavras-chave: *Outliers*. Detecção semissupervisionada de *Outliers*. Análise de dados em alta dimensão. *Hubness*. Mineração de dados. Aprendizado de Máquina. Semissupervisão.

Abstract

With the increase in the amount of data stored, the area of data mining has become essential for it to be possible to manipulate and extract knowledge from these data. Much of the work in this area focuses on finding patterns in the data, but non-standard data (anomalies) can also add much to the knowledge of the data set under study. The study, development and enhancement of outliers detection techniques are important objectives and have proven useful in several scenarios, such as: fraud detection, intrusion detection and monitoring of medical conditions, among others. The paper presented here describes a novel method for semi-supervised detection of outliers in high dimensional data. Experiments with several real datasets indicate the superiority of the proposed method in relation to the literature methods selected as the baseline.

Keywords: Outliers. Semi-supervised detection of outliers. Data mining. High-dimensional Data Analysis. Hubness.

Lista de ilustrações

Lista de tabelas

Lista de siglas

ABOD	*Angle-Based Outlier Detection*
FRSSOD	*Fuzzy Rough Semi-Supervised Outlier Detection*
SOUTH-N	*Semi-supervised OUTlier detection based on Hubness Neighborhood*
LOF	*Local Outlier Factor*
SSODPU	*Semi-Supervised Outlier Detection with Positive and Unlabeled Data*

Lista de símbolos

$X = \{x_1; ...; x_n\}$	Conjunto de dados
k	Vizinhança para cálculo dos K vizinhos mais próximos
$N_k(x)$	Pontuação *hubness* da instância x considerando a vizinhança K
IC	Intervalo de Confiança

Sumário

Capítulo 1

Introdução

A área de pesquisa em Mineração de Dados é uma área multidisciplinar que integra conceitos e técnicas de várias outras áreas como: princípios de Bancos de Dados, Aprendizado de Máquina, Inteligência Artificial, entre outras. As técnicas desenvolvidas nessa área têm sido usadas para analisar grandes volumes de dados com o objetivo de encontrar correlações ou padrões nesses dados que representem alguma informação ou conhecimento. As técnicas de Mineração de Dados existentes são classificadas de acordo com o tipo de conhecimento a ser extraído em: detecção de agrupamentos, associação, detecção de *outlier*, regressão, detecção de classes e etc (AGGARWAL, 2015).O principal foco do trabalho apresentado aqui foi a detecção semissupervisionada de *outliers* em dados com alta dimensão.

Outliers podem surgir devido a erros humanos, erros instrumentais, mudanças no comportamento ou falhas nos sistemas, assim como desvios naturais em sociedades. De forma técnica, um *outlier* é uma instância discrepante de acordo com um padrão observado no conjunto de dados. Logo, em um mesmo conjunto de dados uma instância pode ou não categorizar um *outlier* de acordo com o padrão em análise (AGGARWAL, 2015).

Na maioria das vezes, quando os *outliers* ocorrem suas consequências podem ser dramáticas e muitas vezes no sentido negativo, como pode se observar nas aplicações mostradas no Capítulo 2. Como ilustração, temos a detecção de falhas de motores em aeronaves (YU; CLEARY; CUDDIHY, 2004), análise do mercado de ações (SONG; CAO, 2012), controle de catástrofes e eventos ambientais (ZHENG et al., 2010), detecção de *outliers* em dados textuais para identificar novos tópicos ou notícias em uma coleção de documentos (SRIVASTAVA; ZANE-ULMAN, 2005). Por isso, a demasiada importância da técnica de detecção de *outliers*.

As abordagens disponíveis para detecção de *outlier* podem ser classificadas em

três categorias de acordo com o tipo de aprendizado empregado: aprendizado supervisionado, semissupervisionado e não supervisionado (CHANDOLA; BANERJEE; KUMAR, 2009). Dentre essas abordagens, as técnicas baseadas nas abordagens não supervisionadas e semissupervisionadas têm mostrado serem as mais relevantes (DANESHPAZHOUH; SAMI, 2014). As técnicas baseadas na abordagem supervisionada apresentam alto custo computacional o que torna o seu uso inviável para a manipulação de grandes volumes de dados em altas dimensões(SINGH; UPADHYAYA, 2012). Porém, durante a revisão bibliográfica notou-se que predominante parte dos trabalhos para detecção de *outliers* baseiam-se na abordagem não supervisionada, método que muitas vezes não garante um resultado satisfatório, devido à baixa precisão. Assim, existe uma lacuna que tem sido preenchida por métodos baseados na abordagem semissupervisionada que necessitam de pouca informação prévia para retornar um resultado mais preciso e satisfatório (DANESHPAZHOUH; SAMI, 2014). As abordagens semissupervisionadas são categorizadas de acordo com a entrada do método, abrangendo técnicas de agrupamento, de classificação e estatísticas, conforme descrito no Capítulo 2 e também nos trabalhos (IENCO; PENSA; MEO, 2017), (DANESHPAZHOUH; SAMI, 2015), (DUONG; HAI, 2015) e (ZHANG; LEE, 2005).

Em contrapartida, como o trabalho descrito aqui focou em dados com alta dimensionalidade, outra prioridade para escolha do método de aprendizado foi garantir que a alta dimensionalidade não comprometia o desempenho dos métodos e a veracidade dos resultados. Atualmente, as bases de dados têm aumentado em número de instâncias e dimensões, as quais muitas vezes são, na sua maioria, igualmente relevantes por representarem caraterísticas pertinentes dos dados em análise. Exemplos típicos desse cenário são os conjuntos de dados de imagens nos quais cada imagem é representada por um vetor de n dimensões que correspondem a características de baixo nível da imagem. Por exemplo, se uma imagem for representada pela quantidade de pixels de cada nível de tom de cinza presente nela, cada instância de dados do conjunto de dados será representada por um vetor de 256 dimensões (ou características). Entretanto, tratar conjuntos de dados compostos por instâncias descritas por um elevado número de dimensões tem sido um desafio na área de mineração de dados, conhecido como problema da maldição da dimensionalidade (SAMET, 2005). Esse problema tem afetado a eficácia e eficiência de diversas técnicas de mineração de dados (FACELI et al., 2011).

Para lidar com essa questão, já foram desenvolvidas várias técnicas para a redução de dimensionalidade baseadas em seleção de atributos relevantes ou agregação (FACELI et al., 2011). Contudo, muitas vezes a redução de dimensionalidade acarreta em perda de conhecimento substancial para o conjunto de dados em observação

(SULIC et al., 2010). Levando essa questão em consideração, para tratar dados com alta dimensionalidade alguns trabalhos recentes têm abordado a utilização do aspecto *hubness* como visto em (RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015), (YAGER; DUNSTONE, 2010), (MARQUES, 2015), (HUBNESS-BASED. . . , 2015), (FLEXER, 2016) e (HEYLEN; PARENTE; SCHEUNDERS, 2017) identificando instâncias *hubs* que auxiliam na detecção de grupos e instâncias *anti-hubs* que qualificam os *outliers* em dados de alta dimensão.

1.1 Objetivos

De forma geral o objetivo do trabalho descrito aqui foi investigar, analisar e aprimorar algoritmos de semissupervisão voltados para a detecção semissupervisionada de *outliers* para conjuntos de dados com alta dimensionalidade. Neste trabalho, o principal desafio foi investigar métodos de detecção semissupervisionada de *outliers* que integrem as boas características de duas categorias de métodos, ou seja, boa acurácia de métodos supervisionados e o bom desempenho de métodos não supervisionados. Além disso, também foi considerado um requisito, o fato de que a alta dimensionalidade presente na maioria das bases de dados independente do contexto não deve interferir negativamente nos resultados.

Para corroborar com o objetivo geral e o desafio em questão, este trabalho foi sustentado nos seguintes objetivos específicos:

1. Investigar alternativas para usar informações da relação entre *hubness* e *outliers* na definição de um método para detecção semissupervisionada de *outliers*;

2. Analisar como conciliar os benefícios das duas abordagens investigadas;

3. Comparar o método proposto com o estado da arte por meio de métricas de avaliação comumente empregadas na literatura científica da área, empregando diferentes conjuntos de dados, a fim de comprovar sua eficácia.

O desenvolvimento deste trabalho corroborou que a detecção de *outliers* baseada no aprendizado semissupervisionado com a inclusão do aspecto *hubness*, proveu uma melhora significativa nos métodos de detecção de *outlier* aplicados no contexto de dados com alta dimensionalidade, estimando a relação de precisão e revocação entre os métodos.

O trabalho descrito apresentou alternativas para usar informações da relação entre *hubness* e *outliers* na definição de um método para detecção semissupervisionada de *outliers*. O método proposto, contribuiu-se para a solução do desafio enfrentado pela detecção de *outliers* no espaço de alta dimensão.

A abordagem empregada no desenvolvimento do método proposto neste trabalho consistiu em adotar uma abordagem semissupervisionada baseada apenas em poucas amostras positivas, ou seja, *outliers* previamente rotulados e informados pelo usuário como entrada para o algoritmo proposto. Essa decisão foi motivada pelo fato de que o especialista consegue rotular com facilidade comportamentos fora do padrão e mais agilidade do que todas as possíveis variações de comportamentos normais em um dado cenário em estudo (DANESHPAZHOUH; SAMI, 2015).

1.2 Organização da dissertação

Esta dissertação está organizada em cinco capítulos e um Apêndice, como mostrado a seguir:

- Capítulo 2. Descreve os principais conceitos básicos empregados no desenvolvimento do trabalho, posiciona o trabalho perante o estado da arte, lista o que tem sido feito na literatura e as limitações existentes em relação à detecção de *outliers* em conjuntos de dados de alta dimensionalidade;
- Capítulo 3. Apresenta o método de detecção de *outliers SOUTH-N* (*Semi-supervised OUTlier detection based on Hubness Neighborhood*), que incorpora o aspecto *hubness* e semissupervisão, elaborado e proposto no trabalho de mestrado descrito aqui;
- Capítulo 4. Apresenta os experimentos realizados com diversos conjuntos de dados e descreve a análise de seus resultados;
- Capítulo 5. Descreve as conclusões gerais deste trabalho e as perspectivas de direções para trabalhos futuros;
- Apêndice A. Apresenta a tabela de valores críticos para a avaliação estatística realizada nos experimentos.

Capítulo 2

Conceitos Fundamentais e Trabalhos Correlatos

O capítulo está organizado da seguinte maneira: a Seção 2.1 descreve a tarefa de detecção de *outlier*; a Seção 2.3 descreve formas semissupervisionadas para detectar *outliers*; a Seção 2.4 descreve o impacto da alta dimensionalidade na tarefa de detecção de *outlier*; a Seção 2.5 descreve a relação do aspecto *hubness* com alta dimensionalidade; a Seção 2.6 apresenta a comparação entre os trabalhos correlatos descritos nas seções anteriores; a Seção 2.7 descreve os métodos de avaliação utilizados para avaliar a qualidade dos métodos selecionados. Por fim, a Seção 2.8 apresenta as considerações finais do capítulo.

2.1 Detecção de *Outliers*

Uma das subáreas de pesquisa dentro da área de Mineração de Dados é a detecção de *outliers*. O termo *outlier*, também conhecido como anomalia, é empregado para indicar instâncias discrepantes das demais instâncias da base de dados. A clássica definição de detecção de *outlier* encontrada em (HAWKINS, 1980) denota que *'um outlier é uma instância, que desvia muito de outras instâncias que despertam suspeitas de que são gerados por um mecanismo diferente'*. Outra definição segundo (BARNETT; LEWIS, 1994) afirma que, *'um outlier é uma instância (ou um subconjunto de instâncias) que parece ser inconsistente comparado ao restante do conjunto de dados'*.

É importante lembrar que um *outlier* é uma instância discrepante de acordo com um padrão observado no conjunto de dados. Logo, uma mesma instância pode ser considerada um *outlier* ou não, de acordo com o conjunto de dados selecionado para análise de determinada base de dados. Para ilustrar a definição de *outlier*, na Figura

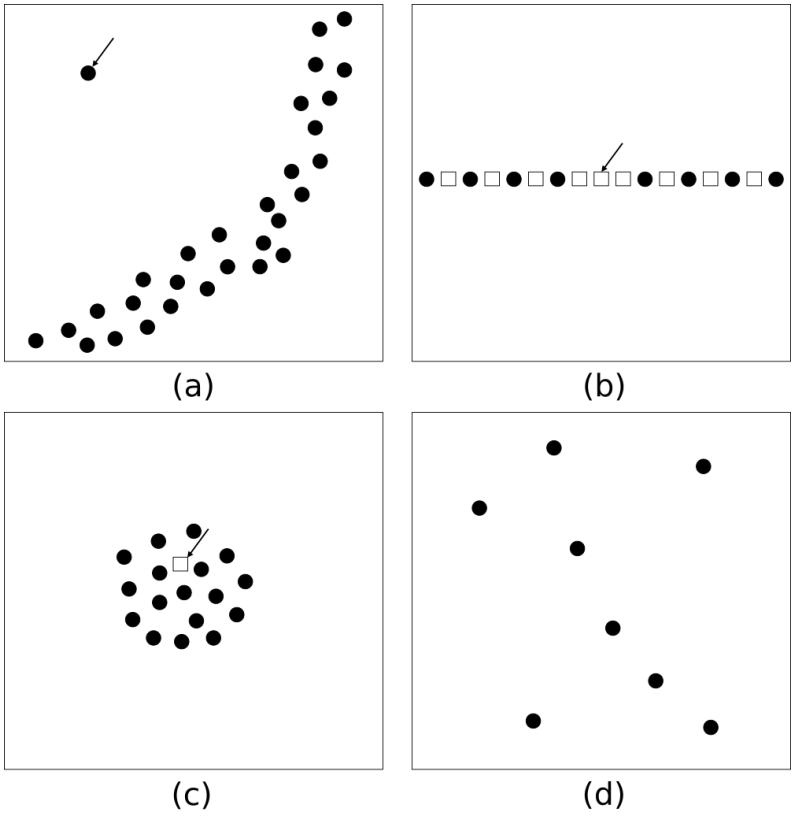

Figura 1 – Exemplos de Detecção de Padrões e Candidatos a *Outliers*. Os *outliers* são identificados pela seta em (a), (b) e (c), para (a) o padrão está na disposição das amostras, (b) o padrão está relacionado a alternancia encontrada de acordo com a forma das amostras e em (c) o padrão esta associado a forma dos elemntos. Porém, em (d) não é possível identificar um padrão e consequentemente a identificação de *outliers* torna-se impraticável.

1 são apresentados quatro diferentes modelos de conjuntos de amostras[1], afim de encontrar um padrão que defina as amostras. Após o padrão ser estabelecido objetiva-se definir possíveis *outliers* que interferem no padrão. Na Figura 1 (a) podemos observar um aglomerado de instâncias que visualmente se assemelha a uma função exponencial, entretanto uma instância destoa das demais, caracterizando um *outlier*. Na Figura 1 (b) temos um conjunto de símbolos • e □ formando uma reta. Neste exemplo podemos notar que o padrão corresponde aos símbolos estarem intercalados entre si. E, entretanto, existe uma falha no padrão encontrado no posicionamento de um símbolo □, conforme indicado pela seta. A Figura 1 (c) mostra outro conjunto formado pelos símbolos • e □, contudo o símbolo □ é nitidamente um *outlier* dentre um grande conjunto formado pelo símbolo •. Já na Figura 1 (d) não é possível notar um padrão nítido no conjunto de amostras, portanto se não há padrão, consequentemente não é possível determinar nenhuma amostra como *outlier* do conjunto.

[1] Os termos 'instâncias' e 'amostras' são usados ao longo do texto como sinônimos

É importante ressaltar que instâncias que se distanciam das demais instâncias de um conjunto de dados de forma insignificante não são considerados *outliers* e sim ruídos. Ruídos não devem ser considerados *outliers*, pois podem prejudicar a precisão dos resultados. Devem ser rotuladas como *outliers* as instâncias que se distanciam substancialmente do padrão das demais instâncias (AGGARWAL; YU, 2001).

Existe uma medida, denominada *outlierness*, definida em (AGGARWAL, 2015) que auxilia nesta distinção entre *outliers* e ruídos. Esta medida, converge para dois resultados, sendo o primeiro voltado a um ranking de *outlier* e o segundo para retornar apenas quais amostras são ou não consideradas.

Existem três tipos de *outlier* definidos na literatura científica da área (CHANDOLA; BANERJEE; KUMAR, 2009). Dado que uma instância individual se distancia do padrão que define todas as demais instâncias em análise, temos o exemplo mais clássico de *outlier*, denominado *outlier* pontual ou global (HAN MICHELINE KAMBER, 2011). Contudo, uma única instância ao ser analisada individualmente pode não representar um *outlier*. Porém, ao ser analisada com um grupo de instâncias ao seu redor é possível identificar que esse conjunto de instâncias não está de acordo com o padrão encontrado nos dados. Esse conjunto de dados pode ser identificado como *outlier* coletivo (NOBLE; COOK, 2003). Por fim, temos os *outliers* contextuais, ou também conhecidos como condicionais, que são designados de acordo com os atributos de comportamento ou de contexto das instâncias em análise. Ou seja, a instância é considerada um *outlier*, se a instância se desvia significantemente em relação a um contexto específico do conjunto de instâncias de dados. Caso este contexto não seja o analisado, essa instância não é considerada *outlier* (MASUD et al., 2013).

Dentre os métodos de detecção de *outlier*, além da entrada do algoritmo, a lógica empregada para a detecção pode também ser variada. Existem métodos baseados em profundidade, nos quais os *outliers* são representados pelas instâncias com profundidade rasa (MONTES, 2014). Por outro lado, em métodos alicerçados em medidas de distância, as instâncias mais distantes são possíveis candidatas a *outliers*. Outro tipo de método existente apoia-se em buscar as instâncias que se desviam das demais de acordo com as características inspecionadas no estudo. E por fim, a abordagem baseada em agrupamentos, define *outliers* como as instâncias isoladas dos demais grupos ou grupos demasiadamente pequenos em relação aos demais (JIANG; YANG, 2009). Nesta última categoria de métodos, uma estratégia para identificar *outliers* seria remover os *clusters*, e categorizar o que sobrar como *outliers* (CHANDOLA; BANERJEE; KUMAR, 2009).

O estudo, o desenvolvimento e o aprimoramento de técnicas de detecção de *ou-*

tliers são objetivos importantes e têm se mostrado útil para diferentes áreas. Como, por exemplo, para identificar fraudes em cartões de créditos (AGRAWAL; KUMAR; MISHRA, 2015); verificar exames médicos para encontrar anomalias que representam doenças, como o câncer (GASPAR; LOPES; FREITAS, 2011); analisar o mercado de ações (SONG; CAO, 2012); controlar catástrofes e eventos ambientais (ZHENG et al., 2010); ou, até mesmo, no processamento de imagens para, por exemplo, caracterizar falhas de vigilância, como em (GULER; TEMIZEL; TEMIZEL, 2013), detectando anomalias na multidão em tempo real.

2.2 Métodos de Detecção de *Outliers*

Na detecção de *outlier*, todos os dados do conjunto em análise são classificados como *outliers* ou *inliers*. Os *outliers* são os dados discrepantes ou anormais e os *inliers* são os dados normais. Para essa identificação dos dados existem várias técnicas descritas na literatura. Essas técnicas podem ser categorizadas em três classes: supervisionadas, não supervisionadas e semissupervisionadas.

2.2.1 Abordagem Supervisionada

As técnicas de detecção supervisionada de *outlier* utilizam um conjunto de dados previamente rotulado por um especialista no assunto em análise para treinar um modelo que depois é utilizado para rotular a base em estudo. Na maioria dos casos os *outliers* são minoria no conjunto de dados o que torna custosa a tarefa de identificação dos mesmos. Assim, a criação de um modelo geral a partir de dados não rotulados para detectar *outliers* se torna uma tarefa complexa. Uma vantagem da utilização de técnicas supervisionadas é que além dos *outliers*, se pode utilizar as amostras *inliers* (que existem em abundância) para criar um modelo que retorne amostras que sejam *outliers* (AGGARWAL, 2013).

Em aplicações que possuem uma quantidade massiva de dados, não é possível rotular todos os exemplos de amostras de *outliers*. Logo, o conjunto de dados pode conter amostras de *outliers* que não foram rotuladas de forma correta. Essa característica implica na redução da acurácia na detecção de *outliers*, pois o dado fornecido pelo especialista já possui rótulos inconsistentes (AGGARWAL, 2013).

Algoritmos de classificação podem ser utilizados na definição de estratégias para a detecção supervisionada de *outliers*. De uma maneira geral, a base de treinamento

define a classe *inlier* e a classe *outlier*. A partir desses dados, um classificador, por exemplo, o SVM gera uma superfície de decisão, que será o modelo para detectar *outliers*. Assim, amostras classificadas como não *inliers* serão consideras como *outliers* (HAN MICHELINE KAMBER, 2011).Esse tipo de abordagem só é adequada para o cenário, no qual é possível e viável rotular as amostras como normais ou *outliers*.

2.2.2 Abordagem Não-Supervisionada

Na detecção não supervisionada de *outlier*, o algoritmo não possui nenhum conhecimento prévio sobre os dados, portanto, ele não possui nenhuma informação que distingue *outliers* e *inliers*. Nessa abordagem, as técnicas supõem que a maioria dos dados são normais e tentam rastrear os dados que mais se diferem (HAN MICHELINE KAMBER, 2011). Como essa abordagem não necessita de conhecimento vindo de um especialista, ela pode ser aplicada em uma gama maior de problemas.

Ao fazer o rastreamento dos dados para verificar quais fogem do padrão, os ruídos podem ser classificados de forma errônea como *outliers*. Além disso, na abordagem não supervisionada, o custo de processamento é maior uma vez que todos os dados devem ser processados para no final selecionar as amostras que mais se diferem. Além disso, na maioria dos casos essas amostras correspondem a um conjunto bem pequeno quando comparado ao conjunto de dados como um todo (HAN MICHELINE KAMBER, 2011). Por fim, outra característica negativa desta abordagem é que a medida de '*outlierness*' é menos precisa que nas demais abordagens (MARQUES, 2015).

Vários métodos de agrupamento podem ser adaptados para detecção não supervisionada de *outlier*. Uma estratégia tradicional consiste em utilizar o *k-means* para encontrar os agrupamentos e depois calcular a pontuação dos *outliers* utilizando uma medida baseada na distância de cada amostra em relação ao centro do agrupamento mais próximo x e a média das distâncias entre a amostra e os elementos do agrupamento x. Outra estratégia consiste em utilizar um algoritmo de agrupamento que já resulte na atribuição de rótulos para os elementos. A execução do algoritmo *DBSCAN*, por exemplo, pode resultar em amostras que não pertençam a nenhum agrupamento (HAN MICHELINE KAMBER, 2011).

2.2.3 Abordagem Semissupervisionada

Nesta categoria, a detecção dos *outliers* é parcialmente supervisionada, isto é, o conhecimento a respeito de algumas instâncias previamente rotuladas é empregado no

processo de detecção de *outliers*. Dentre os trabalhos existentes é mais comum empregar algumas amostras rotuladas como *inliers* na construção do modelo para a detecção de *outliers* (DANESHPAZHOUH; SAMI, 2013). Contudo, também pode ser utilizado um conjunto de amostras rotuladas como *outliers* (DASGUPTA; MAJUMDAR, 2002) ou um conjunto pequeno com amostras que referenciam *inliers* e *outliers* (XUE; SHANG; FENG, 2010). A proposta de trabalho descrita aqui baseia-se na abordagem semissupervisionada. Assim, ela será apresentada em maiores detalhes na próxima Seção.

2.3 Detecção Semissupervisionada de *Outliers*

Atualmente, a maioria dos trabalhos de detecção de *outliers* se baseiam em abordagens não supervisionadas, pela praticidade de conseguir resultados sem informações prévias. Entretanto, esta abordagem apesar de prática, apresenta altos índices de alarmes falsos e baixa taxa de detecção de *outliers* (XUE; SHANG; FENG, 2010). Isso torna a aplicação desses trabalhos inviável e muitas vezes impraticável na maior parte dos cenários reais em que a detecção de *outliers* é necessária e utilizada, como por exemplo a detecção de anomalias em exames médicos. Outra abordagem empregada para detecção de *outliers* consiste na utilização de métodos supervisionados, que necessitam de uma grande quantidade de dados rotulados para treinamento. Essa abordagem gera bons resultados, entretanto, ela demanda muito esforço humano o que a torna mais susceptível a resultados errôneos devido a um modelo de treinamento incorreto. Além disso, a determinação de um exemplo para todos os possíveis casos de desvio do padrão do conjunto de dados em análise é uma tarefa difícil(CHANDOLA; BANERJEE; KUMAR, 2009). Por esses motivos e por fornecer bons resultados em diferentes contextos, a abordagem semissupervisionada tem se mostrado promissora e, atualmente, está sendo mais estudada.

Como citado na Seção 2.2.3, a abordagem de detecção semissupervisionada de *outliers* considera informações prévias fornecidas sobre algumas amostras no processo de descoberta de amostras que caracterizam *outliers*. Em alguns cenários – como detecção de fraude, análise de mercado financeiro e análise de diagnósticos médicos – é possível e factível a definição de alguns exemplos para dar suporte aos algoritmos semissupervisionados. Esses exemplos fornecem uma base de conhecimento ao algoritmo, o que, consequentemente, resulta em uma precisão mais acurada, como constatado nos experimentos vistos em (DANESHPAZHOUH; SAMI, 2014). Por isso, a abordagem semissupervisionada apresenta vantagens em relação a não supervisionada, por apresentar uma melhor performance com resultados mais corretos e precisos. E, ainda, apresenta uma superioridade em relação ao custo de adquirir exemplos de amostras

rotuladas quando comparada a métodos supervisionados, pois reduz a necessidade de uma alta quantidade de dados rotulados (GAO; CHENG; TAN, 2006).

É possível categorizar as técnicas que empregam a abordagem semissupervisionada de detecção de *outliers* em três categorias de acordo com a entrada do algoritmo. A primeira categoria, emprega exemplos de *outliers* e *inliers* no seu aprendizado (FASSETTI; ANGIULLI, 2010). A segunda categoria utiliza como exemplo apenas amostras de *outliers*. Nessa categoria os exemplos de *outliers* são chamados de amostras positivas e os exemplos de *inliers* são denominados amostras negativas. A terceira categoria baseia-se apenas na entrada de dados rotulados como *inliers* para a detecção de *outliers*. *Outliers* são raros e difíceis de rotular com precisão, de forma que cubra todas as possibilidades de desvios do padrão (AGGARWAL, 2013). E em muitos casos, o conjunto de amostras negativas (*inliers*) não está disponível devido ao fato de que rotular instâncias negativas é dispendioso, como no cenário de detecção de fraudes, na qual, é possível rotular com facilidade exemplos de fraudes (DANESHPAZHOUH; SAMI, 2015). Por isso, a segunda categoria torna-se mais viável que as demais, por exigir menos esforço humano.

Assim, os métodos com maior possibilidade de aplicações são fundamentados apenas em receber amostras positivas e os dados não rotulados. Dentro desse contexto existem três diferentes estratégias que podem ser utilizadas. Na primeira, apenas as amostras positivas são consideradas para rotular o restante dos dados (MANEVITZ; YOUSEF, 2002). A segunda estratégia proposta constitui de duas fases, sendo a primeira responsável por extrair prováveis amostras negativas de dados não rotulados e a segunda responsável por utilizar métodos tradicionais de classificação nos dados não rotulados (DANESHPAZHOUH; SAMI, 2014). Por fim, a última estratégia é baseada em conceitos probabilísticos. Nessa estratégia não é necessário extrair amostras negativas, entretanto, vários parâmetros devem ser estimados, o que pode afetar diretamente o desempenho, por tornar o resultado suscetível aos parâmetros usados (ZHANG; LEE, 2005).

Como exemplo de detecção semissupervisionada de *outliers* baseada no aprendizado a partir de amostras positivas e não rotuladas, podemos destacar na literatura, em (DANESHPAZHOUH; SAMI, 2015), o trabalho fundamentado em dois passos utilizados para encontrar *outliers*, sendo eles:

1. Extração de amostras negativas confiáveis a partir de amostras positivas (*outliers*) em dados não rotulados utilizando o K-NN (MITCHELL, 1997);

2. Detecção de *outliers* utilizando as novas amostras negativas rotuladas no passo

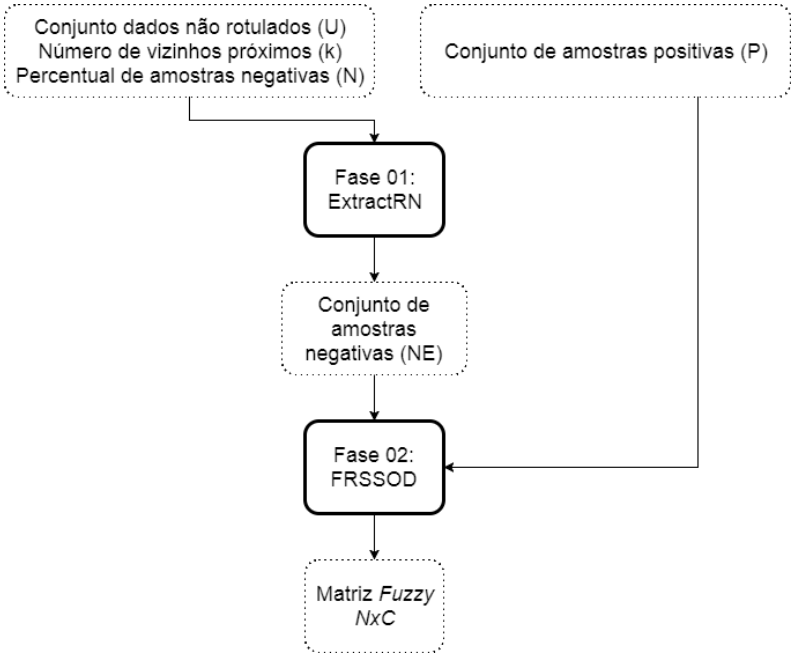

Figura 2 – Etapas da abordagem proposta em (DANESHPAZHOUH; SAMI, 2015).

anterior.

Como visto na Figura 2, o trabalho *Semi-Supervised Outlier Detection with Positive and Unlabeled Data* (SSOPDU) proposto em (DANESHPAZHOUH; SAMI, 2015) possui quatro entradas (P, U, K, N): conjunto de amostras positivas (P), dados não rotulados (U), número de vizinhos (K) e percentual de amostras negativas (N). Na primeira fase *ExtractRN* o objetivo é encontrar o percentual N de amostras negativas em U a seguinte estratégia é adotada: é realizado o somatório das distâncias de cada dado não rotulado em relação aos K vizinhos encontrados para cada amostra positiva. Depois, essa lista de somatórios é ranqueada de forma decrescente para cada amostra positiva, dessa maneira no topo de cada lista estão os elementos não rotulados mais distantes em relação a cada amostra positiva e sua vizinhança. Como saída é gerado o conjunto de amostras negativas confiáveis (NE), obtido a partir dos (N) percentual primeiros elementos de cada lista referente a cada amostra positiva. Feito isso, o segundo passo é utilizar o algoritmo *FRSSOD* (*Fuzzy Rough Semi-Supervised Outlier Detection*) apresentado em (XUE; SHANG; FENG, 2010) e descrito em detalhes na Seção 3.1.2 por ser usado no método *SOUTH-N*.

Em (DUONG; HAI, 2016) é apresentada uma abordagem para detecção de *outliers* ou anomalias em pacotes de rede que podem ser indícios de quebra de segurança, questão importante para várias áreas como mercado financeiro, internet entre outros (DUONG; HAI, 2015). Os métodos mais usuais para detecção de intrusão em redes baseiam-se em conhecimentos prévios do que é um comportamento inadequado na

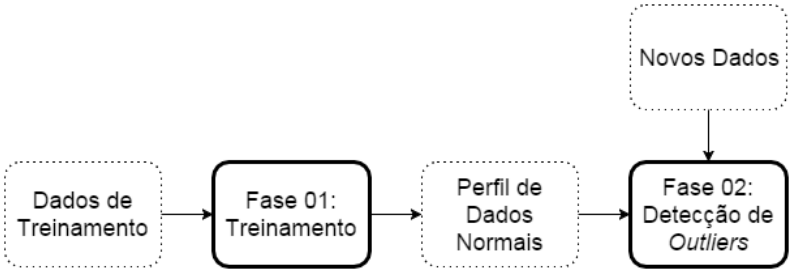

Figura 3 – Etapas da abordagem proposta em (DUONG; HAI, 2016).

rede. Essa abordagem abre brechas para *hackers* inovarem na forma de invasão para serem bem sucedidos em suas intenções ilícitas. Por isso, (DUONG; HAI, 2016) propõe uma estratégia semissupervisionada para detectar *outliers* a partir de dados normais, ou seja, comportamento padrão e correto nas redes.

Assim, foi proposta a estratégia dividida em duas fases, sendo a primeira responsável pelo treinamento dos dados e a segunda pela detecção de *outliers*. Conforme visto na Figura 3, a primeira fase é responsável por projetar o perfil do tráfego normal dos dados a partir do padrão dos dados normais. Para isso, inicialmente é utilizado o algoritmo *k-means* (ARTHUR; MANTHEY; RöGLIN, 2011), para eliminar os ruídos dos dados de entrada. Como os ruídos são bem menores que os dados normais, neste trabalho é assumido que 10% dos dados de entrada são ruídos e estes são eliminados. Após essa redução nos dados, é realizado a normalização dos dados a partir da ponderação entre a média e o desvio padrão dos dados. Depois, o algoritmo *PCA* (SHYU et al., 2003) realiza a análise da significância de cada variável. O algoritmo *PCA* resultará no autovalor e autovetor dos dados normais. E por último, é construída uma função empírica cumulativa de distribuição para determinar os limiares. Assim, é encontrado o perfil de pacotes de rede considerado normal utilizado como entrada na segunda fase, pois esse perfil considerado normal é utilizado como conjunto de treinamento para o classificador *PCA*, adaptado para utilizar a distância de *Mahanalobis*, chamado *M-PCA*, que diferencia pacotes normais (*inliers*) de não-normais (*outliers*). O trabalho se mostrou eficiente utilizando esses dois algoritmos e tem uma performance ainda maior quando o conjunto de dados gerado pelo *K-means* tem um tamanho de até mil amostras.

Em relação a utilização de conceitos probabilísticos, podemos citar na literatura o trabalho (ZHANG; LEE, 2005), que parte da premissa que o conjunto de dados não rotulados e positivos são mais representativos que o conjunto de dados negativos, com isso transforma um modelo probabilístico baseado em amostras positivas e negativas, através de substituições matemáticas em um modelo, que depende apenas da probabilidade da amostra pertencer ao conjunto de amostras não rotulados ou positivos. Neste utiliza-se o algoritmo *PrTFIDF* para estimar essas probabilidades. Essa estraté-

gia probabilística foi chamada de *Biased-PrTFIDF*. A vantagem deste método, é que a implementação simplista apresentou bons resultados em performance e eficiência para os conjuntos de dados escolhidos para testes.

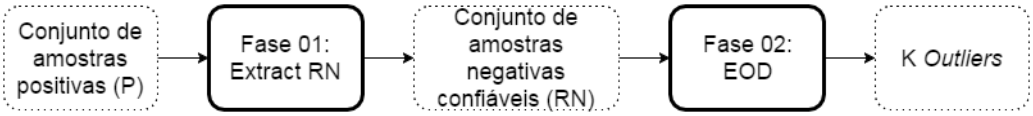

Figura 4 – Etapas da abordagem proposta em (DANESHPAZHOUH; SAMI, 2013).

O trabalho (DANESHPAZHOUH; SAMI, 2014) apresenta uma abordagem para detecção semissupervisionada de *outliers* baseada na estratégia de solucionar a detecção com um processo de duas fases, conforme visto na Figura 4. A primeira fase é responsável por extrair amostras negativas confiáveis a partir de dados não rotulados e poucos exemplos de amostras positivas. Para isso, é utilizado o algoritmo *Extract-RN*, que recebe como entrada o conjunto de dados não rotulados (U), exemplos de amostras positivas (P) e l% de *outliers* a serem reconhecidos. Assim a entropia é calculada para cada instância d_i pertencente a U. Após calcular a entropia de todas as instâncias não rotuladas, as mesmas são ordenadas decrescentemente. Assim, os primeiros l% de amostras são selecionadas como amostras negativas confiáveis (RN).

Para a segunda fase da abordagem, responsável pela detecção dos *outliers*, é utilizado um algoritmo, baseado no algoritmo não supervisionado *LSA* apresentado em (HE; DENG; XU, 2005), (HE; XU; DENG, 2005). O objetivo do algoritmo *LSA* é encontrar os *k outliers* por meio de uma função objetivo baseada em entropia. Este trabalho aproveita a informação prévia já obtida para garantir melhores resultados. Dessa maneira, as entradas necessárias para o algoritmo *Entropy-based Outlier Detection (EOD)* são: o número de *outliers* (k), exemplos de amostras positivas (P), amostras negativas (RN) e o conjunto de dados não rotulados. Inicialmente, *k* amostras negativas são retiradas do conjunto de dados (D). Para isso, é calculada a distância entre cada elemento pertencente a D e as amostras positivas de P, quando a distância for menor que um limiar T pré-definido, esses elementos são retirados do conjunto de dados. Neste novo conjunto de dados, *k-p* primeiros elementos são selecionados como *outliers* e atribuídos ao conjunto O e o restante é rotulado como *inlier* e atribuído no conjunto N. Assim, cada elemento rotulado como não *outlier* é trocado com um elemento do grupo de *outlier* formando uma nova distribuição dos conjuntos de dados e entropia objetiva desta nova distribuição é recalculada, caso a nova entropia seja menor que a atual, essa nova de divisão do conjunto de dados é mantida. Esse processo é executado até a entropia objetiva da divisão atual dos rótulos atribuídos ao conjunto de dados convergir e não variar mais de acordo com as novas combinações entre os conjuntos. Assim, o conjunto

final, possuirá os *k outliers* para retornar ao usuário.

2.4 Detecção de *Outlier* em Alta Dimensão

Com o passar dos tempos não apenas a quantidade de dados aumentou como também a dimensão desses dados, ou seja, cada instância de um conjunto de dados passou a ser caracterizada por dezenas ou até centenas de características ou atributos. Além disso, conjuntos de dados com essa característica passaram a ocorrer em muitos domínios de aplicação, por exemplo, no domínio de reconhecimento de imagens, em que cada imagem pode ser representada por vetores, que por sua vez estão relacionados a diversas características (ZAKI; JR, 2014).

Entretanto com o aumento da dimensionalidade, surge o efeito conhecido como maldição da dimensionalidade. Esse efeito afeta várias tarefas de mineração de dados, principalmente as baseadas em cálculos de distância. Uma das consequenciais desse efeito está relacionada com a distribuição dos dados, ou seja, em cenários de alta dimensão ocorre a concentração das distâncias, fazendo com que a distância entre os dados pareça imperceptível (CLARKE; ZHANG, 2009), assim como a pouca variação no desvio padrão. Essa concentração das distâncias resulta em uma pobre discriminação dos dados (AGGARWAL; HINNEBURG; KEIM, 2001).

Na literatura científica da área as principais propostas para lidar com a detecção de *outliers* em dados com alta dimensão são segmentadas em duas dominantes categorias. Na primeira o objetivo é transformar os dados de alta dimensionalidade em dados com uma menor dimensionalidade utilizando técnicas de redução de dimensionalidade (UMA. . . , 2011). Alguns exemplos de técnicas de redução de dimensionalidade tradicionais são: análise de componentes principais (*PCA*), análise de componentes independentes (*ICA*) e a decomposição singular do valor (*SVD*) (VINAY et al., 2005). Na segunda categoria estão as técnicas em que os algoritmos de detecção de *outlier* relacionam a precisão e a proximidade entre as instâncias no conjunto de dados de alta dimensionalidade(ZHANG, 2013).

Um claro exemplo deste cenário pode ser visualizado na Figura 5 apresentada em (AGGARWAL, 2013), na qual quatro visões bidimensionais de um conjunto de dados hipotético são apresentadas. Nessa Figura cada visão corresponde a conjuntos disjuntos das dimensões do domínio de dados. Nela podemos perceber que na primeira visão, Figura 5(a), a instância A pode ser considerada um *outlier*. Já na quarta visão, Figura 5(d), a instância B é que representa um *outlier*. Contudo, na segunda e na terceira

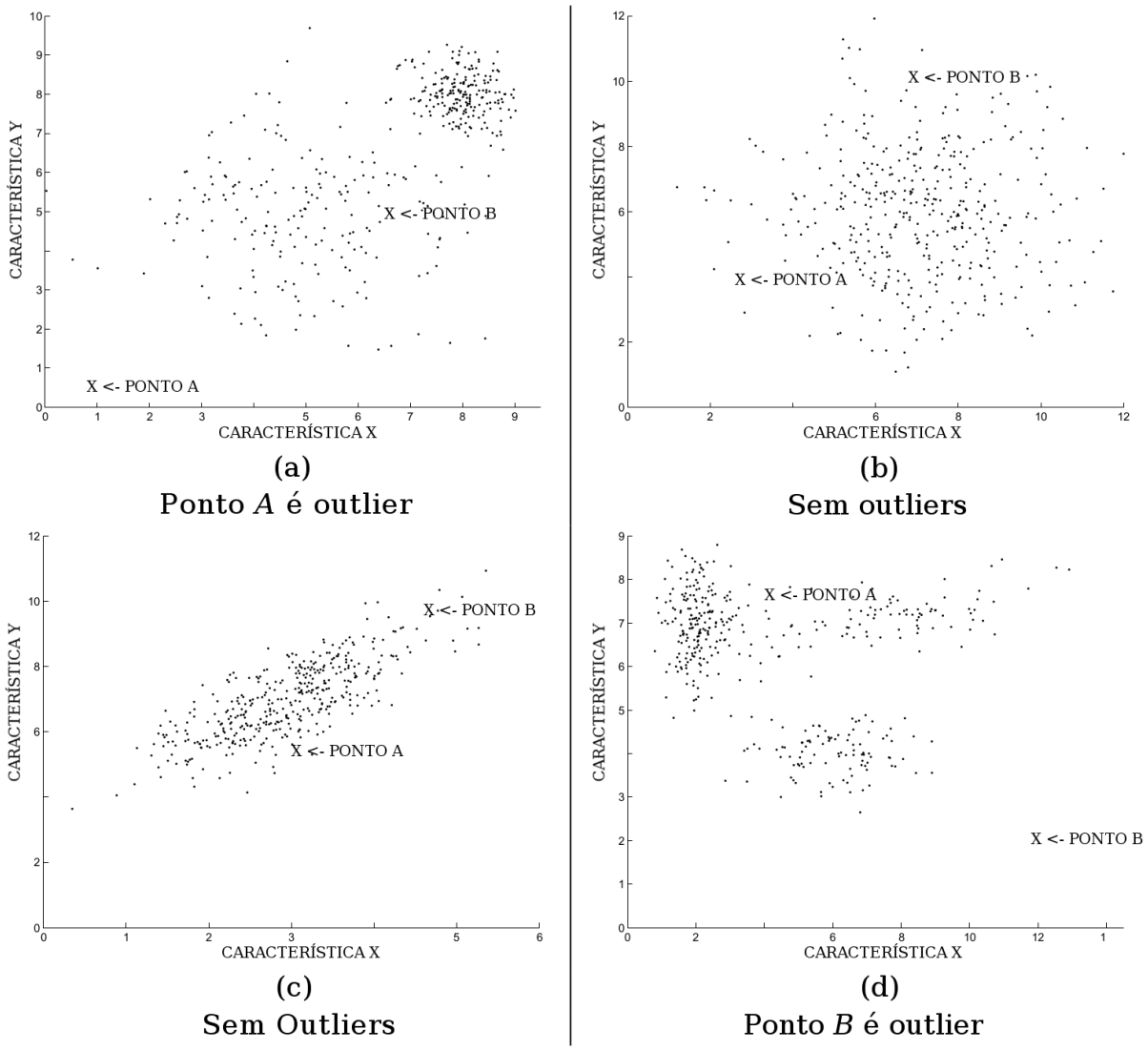

Figura 5 – O *outlier* pode ser perdido na maioria dos subespaços escolhidos aleatoriamente em casos com alta dimensionalidade (adaptado de (AGGARWAL, 2013)).Neste caso, as instâncias A e B mudam de classificação a partir das perspectivas (a), (b), (c) e (d).

visões, Figura 5(b) e (d), nenhuma das duas instâncias são caracterizadas como *outliers* e não são detectados *outliers* nessas visões. Por isso, um método de detecção de *outliers* eficiente deve considerar as instâncias e as dimensões de forma integrada para detectar os *outliers*, pois diferentes subconjuntos de dimensionalidades podem ser relevantes para diferentes *outliers*.

Um exemplo de método bastante utilizado na detecção de *outliers* em alta dimensão,

que leva em consideração o conjunto de dados como um todo, considera métodos de medidas de distância para determinar a similaridade entre as amostras. No entanto, a eficiência desta abordagem diminui à medida que a dimensão aumenta, pois os cálculos se tornam mais onerosos, além de gerar uma pobre discriminação das amostras devido a concentração dos dados (HOULE et al., 2010).

Apesar dos constantes avanços da área de pesquisa em detecção de *outlier* em alta dimensão, esta tarefa continua sendo um desafio, devido a maldição da dimensionalidade citada em 2.4. No entanto, há fatores que a maldição da dimensionalidade geram, como a concentração da distância, que podem ser utilizados a favor não só da detecção de *outliers* como também de outras tarefas de mineração de dados que sofrem os efeitos dessa maldição. Um desses fatores é denominado aspecto *hubness* e será abordado na Seção 2.5 a seguir.

2.5 Conceito *hubness*

Conforme mencionado em 2.1, o aspecto *hubness* é um fenômeno inerente dos conjuntos de dados em alta-dimensão. Esse aspecto representa a tendência dos dados de alta-dimensão conterem instâncias (chamadas de *hubs*) que ocorrem com frequência na listagem dos k-vizinhos mais próximos de outras instâncias. Trabalhos recentes têm buscado tirar proveito desse fenômeno na proposta de métodos de agrupamento (HUBNESS-BASED. . . , 2015), de classificação (TOMASEV; BUZA, 2015), de busca de k-vizinhos mais próximos (RADOVANOVIć; NANOPOULOS; IVANOVIć, 2010) e de detecção não-supervisionada de *outliers* (RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015) mais eficazes no processamento de dados em alta dimensão.

A medida que a dimensionalidade intrínseca dos dados aumenta a distribuição das k-ocorrências na lista de vizinhos mais próximos de cada instância de dados torna-se distorcida e com maior variância. Assim, algumas instâncias de dados (denominadas *hubs*) aparecem frequentemente na listagem dos k-vizinhos mais próximos e, ao mesmo tempo, algumas outras instâncias (denominadas anti-*hubs*) tornam-se vizinhos infrequentes (RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015). Formalmente, pontuação *hubness*, *hubs* e anti-*hubs*são definidos segundo (HUBNESS-BASED. . . , 2015) como apresentado nas definições a seguir.

Definição 1 (pontuação *hubness* N_k). Seja D $\subset R^d$, um conjunto de instâncias de dados, $N_k(x)$ denota o número de k-ocorrências de instâncias $x \in D$, isto é, o número de vezes que x ocorre na listagem dos k-vizinhos mais próximos de outras instâncias

pertencentes a D.

Definição 2 (*Hubs*). São instâncias de dados x que aparecem notavelmente em muitas listas de k vizinhos mais próximos das demais instâncias, ou seja, possuem $N_k(x)$ significativamente acima da média..

Definição 3 (*Anti-hubs*). São instâncias de dados x que não aparecem em praticamente nenhuma lista de k vizinhos mais próximos das demais instâncias de dados, isto é, possuem $N_k(x)$ extremamente baixo, ou até mesmo, $N_k(x) = 0$.

Com isso, podemos concluir que *hubs* tendem a estar em regiões mais densas, enquanto os anti-*hubs* estão nas regiões mais esparsas (RADOVANOVIć; NANOPOULOS; IVANOVIć, 2010). Para a realização do trabalho descrito aqui, a análise dos anti-*hubs* merece especial atenção uma vez que o surgimento de anti-*hubs* está relacionado com a descoberta de *outliers*.

A proposta de métodos de detecção de *outliers* baseados na análise de anti-*hubs* começou a ser explorada na definição de métodos não supervisionados em (RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015). Os métodos propostos baseiam-se nas seguintes propriedades observadas sobre os anti-*hubs*:

1. Existe uma relação entre a pontuação *hubness* N_k e a probabilidade de uma instância de dados ser um *outlier*. Isso faz com que anti-*hubs* sejam bons candidatos a *outliers* para conjuntos em alta dimensão;

2. O fato da dimensionalidade dominar o número de instâncias em um conjunto de dados, o que resulta em grande esparcidade nos dados, implica necessariamente na ocorrência de grandes níveis de *hubness*;

3. O fato do número de instâncias em um conjunto de dados aumentar não diminui ou elimina a ocorrência do fenômeno *hubness* nos dados.

O primeiro método proposto (chamado AntiHub^1) revisita o método de detecção de *outliers* ODIN (HAUTAMAKI; KARKKAINEN; FRANTI, 2004) incorporando o cálculo da pontuação *hubness* N_k como medida de pontuação para *outlier* (veja o Algoritmo 1). Entretanto, esse método apresentou como ponto fraco a baixa discriminação de pontuação. Para reverter essa questão é necessário utilizar valores altos para o k usado no cálculo do N_k, o que pode resultar na falha de detecção de *outliers* locais e no aumento do custo computacional associado ao cálculo do N_k (especialmente se forem conside-

rados valores de k próximos do número de instâncias do conjunto de dados).

Algoritmo 1: AntiHub^1

Entrada:
distancia: Medida de distância
conjuntodedados: Conjunto de dados com n elementos $(x_1, x_2, ..., x_n)$
k: Quantidade de vizinhos
Saída:
Lista com n elementos, cada i-ésimo elemento é o *outlier* score de *conjuntodedados*$[i]$
Variáveis temporárias:
$t \in \mathbb{R}$

```
1 início
2     para cada i ∈ (1, 2, ..., n) faça
3         t := N_k(x_i);
4         s_i := f(t);
5     retorna s
6 fim
```

Com o intuito de aumentar o poder de discriminação do N_k sem consequentemente aumentar demasiadamente o custo computacional, foi proposto o método denominado AntiHub^2. Esse método é uma heurística que busca refinar a pontuação de *outlier* dada pelo método AntiHub^1 para uma dada instância de dados x considerando uma agregação das pontuações N_k dos vizinhos de x (veja o Algoritmo 2).

Outro trabalho que corrobora a eficiência do aspecto *hubness* na detecção de *outliers* em conjuntos de alta dimensionalidade é retratado em (FLEXER, 2016), no qual é apresentado uma abordagem baseada em *hubness* com base na reconfiguração do espaço de distância através dos métodos previamente elaborados em outros trabalhos: *kNN-reject* ((RAMASWAMY; RASTOGI; SHIM, 2000)), *AH-reject* ((RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015)) e *MP-reject* ((SCHNITZER et al., 2012)).

2.6 Análise Comparativa

Esta Seção visa apresentar trabalhos da literatura correlata a proposta deste trabalho, assim como correlacioná-los. Esta correlação pode ser notada na tabela 1. Conforme pode ser observado, os trabalhos que utilizam o aprendizado semissupervisionado para detecção de *outliers*, como, (DANESHPAZHOUH; SAMI, 2015), (DUONG; HAI, 2015) e (ZHANG; LEE, 2005), não levam em consideração o aspecto *hubness* para aumentar a eficiência em dados com alta dimensionalidade. A detecção de *outliers* usando a abordagem semissupervisionada têm alcançado resultados promissores pelas razões

Algoritmo 2: AntiHub^2

Entrada:
distancia: Medida de distância
conjuntodedados: Conjunto de dados com n elementos $(x_1, x_2, ..., x_n)$
k: Quantidade de vizinhos
p: Proporção de outliers para maximizar a discriminização
passo: parâmetro de busca
Saída:
Lista com n elementos, cada i-ésimo elemento é o *outlier* score de *conjuntodedados*$[i]$
Variáveis temporárias:
a: *anti-hub* scores $\in \mathbb{R}^n$
aNN: Soma dos *anti-hub* scores dos vizinhos mais próximos $\in \mathbb{R}^n$
α: Proporção $\in [0, 1]$
$cdist, dist$: Score discriminante $\in \mathbb{R}$
ct, t: Outlier score bruto $\in \mathbb{R}^n$
Funções:
$discScore(y, p)$: para cada $y \in \mathbb{R}^n$ e $p \in (0, 1]$, retorna o número de elementos únicos entre $\lceil np \rceil$ menores que os elementos de y, dividido por $\lceil np \rceil$

início
  $a := anti - hub(conjuntodedados, k)$;
  para *cada* $i \in (1, 2, ..., n)$ **faça**
    $aNN_i := \sum_{j \in NN_{distancia}(k,i)} a_j$, onde $NN_{distancia}(k, i)$ é o conjunto de índices dos k vizinhos mais próximos de x_i;
  $disc := 0$;
  para *cada* $\alpha \in (0, passo, 2 \cdot passo, ..., 1)$ **faça**
    para *cada* $i \in (1, 2, ..., n)$ **faça**
      $ct_i := (1 - \alpha) \cdot a_i + \alpha \cdot aNN_i$
    $cdisc =: discScore(ct, p)$;
    se $cdisc > disc$ **então**
      $t := ct$, $disc =: cdisc$;
  para *cada* $i \in (1, 2, ..., n)$ **faça**
    $s_i := f(t_i)$, onde $f : \mathbb{R} \Rightarrow \mathbb{R}$ é uma função para normalização
  retorna s
fim

explanadas na Seção 2.3.

Porém, um crescente desafio em mineração de dados é o exponencial aumento das bases de dados, tanto em relação a quantidade de instâncias de dados quanto em relação ao número de dimensões que descrevem cada instância. Dessa maneira, é possível notar uma lacuna na literatura que não estuda essa promissora junção dos conceitos para um persuasivo método de detecção de *outliers*.

Tabela 1 – Comparação entre os trabalhos correlatos a proposta de trabalho.

Referência	Semi-Supervisão	*Hubness*
(DANESHPAZHOUH; SAMI, 2014)	√	
(DANESHPAZHOUH; SAMI, 2015)	√	
(DUONG; HAI, 2015)	√	
(ZHANG; LEE, 2005)	√	
(RADOVANOVIC; NANOPOULOS; IVANOVIC, 2015)		√
(FLEXER, 2016)		√
Proposta de trabalho	√	√

2.7 Medidas de Avaliação

Os métodos de avaliação são empregados com objetivo de comparar o método proposto aos métodos concorrentes. Para realizar tal avaliação a Seção 2.7.1 descreve as medidas utilizadas. Geralmente, para analisar as métricas obtidas considera-se índices estatísticos, que é apresentado na Seção 2.7.2.

2.7.1 Métricas

Essa Seção apresenta as métricas selecionadas para avaliação dos métodos. A Seção está organizada da seguinte maneira: a Seção 2.7.1.1 descreve as métricas Precisão, Revocação e Acurácia; a Seção 2.7.1.2 apresenta a Medida F. Por fim, a Seção 2.7.1.3 descreve a Curva Precisão-Revocação.

2.7.1.1 Precisão, Revocação e Acurácia

Dentre as medidas de qualidade objetivas comumente empregadas na avaliação de métodos de detecção de *outliers* estão a precisão e revocação (DANESHPAZHOUH; SAMI, 2014). A precisão avalia a taxa com que todas as instâncias identificadas como *outliers* são realmente *outliers*. Já a revocação avalia a taxa com que um dado método indica como *outliers* todas as instâncias que são *outliers*. Essa medida mensura o quanto do total real de *outliers* foram detectados. Por fim a acurácia avalia as respostas classificadas corretamente do banco de dados de testes. A definição formal dessas medidas é apresentada nas Equações 1, 2 e 3.

$$Precisão\ (Precision) = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoPositivo}; \quad (1)$$

$$Revocação\ (Recall) = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoNegativo}; \tag{2}$$

$$Acurácia = \frac{VerdadeiroPositivo + VerdadeiroNegativo}{VerdadeiroPositivo + FalsoPositivo + VerdadeiroNegativo + FalsoNegativo}. \tag{3}$$

A descrição dos termos empregados no cálculo das medidas de precisão, revocação e acurácia é apresentada a seguir. E a relação desses termos proveniente para os erros e acertos é detalhada na tabela 2.

- ❏ Verdadeiro Positivo (VP): *Outlier* identificado como *Outlier* pelo modelo em observação;
- ❏ Falso Positivo (FP): *Inlier* identificado como *Outlier* pelo modelo em observação;
- ❏ Verdadeiro Negativo (VN): *Inlier* identificado como *Inlier* pelo modelo em observação;
- ❏ Falso Negativo (FN): *Outlier* identificado como *Inlier* pelo modelo em observação.

Tabela 2 – Relação entre *Outliers* e *Inliers* encontrados comparado com os reais categorizados na base em estudo

		Modelo Proposto	
		Outliers	*Inliers*
Real	*Outliers*	**VP**	**FN**
	Inliers	**FP**	**VN**

2.7.1.2 Medida F

Tanto a medida de precisão quanto a medida de revocação são importantes na análise dos resultados, não sendo possível mensurar qual é mais significativa que a outra. A métrica *F-Measure* (TAN; STEINBACH, 2006), em português Medida F, emprega uma média harmônica entre os valores de precisão (*precision*) e revocação (*recall*) com o objetivo de ponderar uma boa taxa de precisão e de revocação. A definição formal dessa medida é apresentada na Equação 4.

$$\text{Medida F} = 2\left(\frac{\text{Precisão x Revocação}}{\text{Precisão + Revocação}}\right) \tag{4}$$

Outra medida habitualmente utilizada para avaliar conjuntos de dados desbalanceados, característica comumente encontrada em conjuntos de dados com *outliers*, visto a raridade da quantidade de *outliers* em relação aos *inliers*, é a métrica curva ROC (*Receiver Operating Characteristics*). Essa métrica não avalia somente o acerto do algoritmo de aprendizado, mas sim, o seu desempenho geral, avaliando a quantidade de acertos em comparação a quantidade de erros e dispondo o resultado em um gráfico. Essa medida é especialmente útil para avaliar se determinado método de detecção de *outliers* detecta uma alta proporção de *outliers* detectados corretamente ao mesmo tempo que indica apenas poucos não-*outliers* como *outliers*. A quantidade de acertos é calculada pela Equação 5 e a quantidade de erros representada pela Equação 6.

$$Acertos = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoPositivo} \tag{5}$$

$$Erros = \frac{FalsoPositivo}{VerdadeiroNegativo + FalsoNegativo} \tag{6}$$

2.7.1.3 Curva Precisão-Revocação

Para formar a curva Precisão-Revocação os elementos retornados são analisados, se o elemento retornado é relevante, então a precisão e a revocação aumentam, e a curva aumenta para a direita. Caso contrário, ou seja, o elemento não é relevante, a revocação é mesma para os elementos superiores, porém a precisão diminui (HAN; KAMBER, 2006).De forma, que seja possível avaliar a precisão em diversos níveis de revocação para obter o desempenho geral do método analisado.

A curva mostra a variação entre o quão sensível e o quão específico é um método. Logo, qualquer aumento na sensibilidade será acompanhado por uma diminuição da especificidade. Se um método se torna mais sensível a *outliers*, ele irá rejeitar mais deles, mas, ao mesmo tempo, também se tornará menos específico e também rejeitará de forma errônea *inliers*. Consequentemente, quanto mais próxima uma curva segue a borda esquerda e depois a borda superior do espaço, melhor será a performance do método (FLEXER, 2016).

Assim, examinar toda a curva de precisão-revocação é muito útil para ponderar a eficácia dos algoritmos ponderando as métricas de precisão e revocação, mas muitas

vezes é necessário ter uma informação numérica que represente a análise, para isso é apresentado na Seção 2.7.1.3 a métrica AUC.

A métrica área sob a curva (AUC) delimita um valor numérico a partir do cálculo da área sob curva, neste caso, a curva em uso é a curva de precisão-revocação. De maneira, que a comparação e os testes entre as estratégias sejam interpretados a partir de um valor numérico.

2.7.2 Teste Estatístico

A validação da eficácia de uma proposta em relação a outras propostas para diversos conjuntos de dados com o resultado de métricas comumente utilizadas, como as apresentadas na Seção 2.7, não são claras o suficiente para determinar com exatidão a melhoria ou não de uma proposta em relação as demais devido a variabilidade dos resultados de acordo com os conjuntos de dados. Para esse fim, pesquisadores adotam técnicas estatísticas que se adaptem a distribuição dos dados em análise, nas quais, sempre existe a validação de uma hipótese para avaliar os resultados das métricas de avaliação (DEMSAR, 2006).

Para a análise estatística, inicialmente os dados devem ser coletados, assim como a definição do problema a ser examinado. Após as etapas iniciais é necessário a análise da distribuição dos dados para escolha da validação estatística mais adequada. Um dos métodos comumente usados é o *boxplot* ('gráfico de caixa') para validação do fator de normalização dos dados (HOULE et al., 2010).

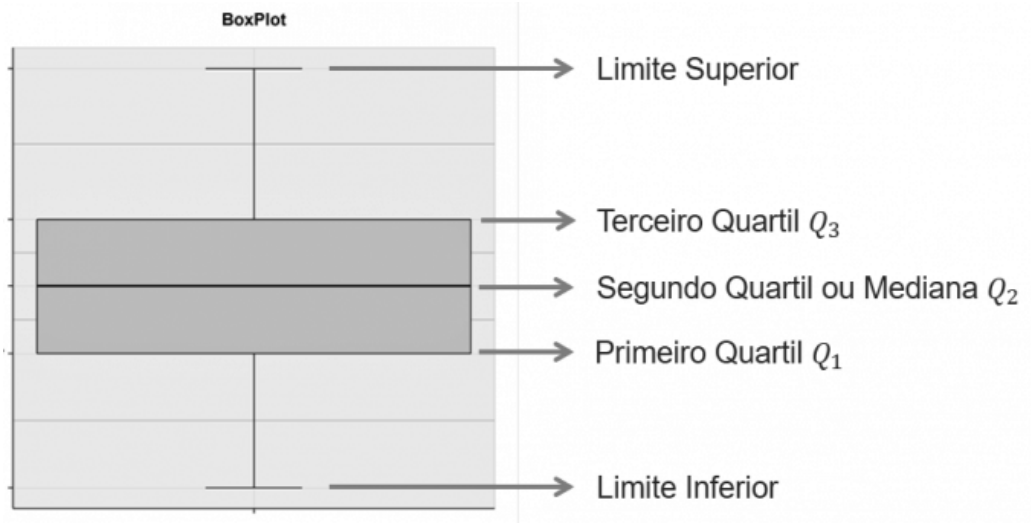

Figura 6 – *Boxplot* e suas informações adaptada de (OTT; LONGNECKER, 2010)

O *boxplot* apresenta cinco valores, o limite inferior ou mínimo, o primeiro quartil (Q1), o segundo quartil ou a mediana, o terceiro quartil (Q3) e o limite superior ou máximo, conforme ilustrado na Figura 6. Assim é possível observar o centro dos dados a partir da mediana, a amplitude dos dados a partir dos limites inferior e superior e por fim a simetria dos dados a partir dos quartils. A informação importante para esse trabalho é a respeito da normalidade dos dados que é adquirido a partir da simetria dos dados. No *boxplot* o limite inferior e superior são respectivamente o menor e o maior valor dos dados e nos casos de distribuições simétricas e, consequentemente, normais, os quartils são divididos a partir da distribuição dos dados de forma que o segundo quartil, Q2, defina o valor que limite 50% de elementos acima e abaixo dele, o primeiro quartil, Q1, é o número que deixa 25% das observações abaixo e 75% acima e o terceiro quartil, Q3, de forma inversa ao primeiro. A partir do posicionamento assimétrico da mediana em relação a caixa e a cauda do *boxplot* é possível constatar que uma distribuição de dados não é normal. Nesse trabalho a distribuição dos dados dos resultados obtidos seguia a normalidade, porém como tratavam-se de amostras não independentes pelas variações dos algoritmos foi adotado o Teste T Pareado para análise (OTT; LONGNECKER, 2010).

2.7.2.1 Teste T Pareado

Alguns testes estatísticos são ideais para situações em que as amostras aleatórias são independentes de duas populações obtidas. Estes métodos não são adequados para estudos ou ensaios em que cada uma das medições em uma amostra é emparelhada com uma medição especial na outra amostra. Neste trabalho, o Teste *T* Pareado foi escolhido e utilizado na experimentação para comparar duas métricas de duas populações e ambas com distribuição normal. Ou seja, é o método adequado para estudos em que cada uma das medidas da primeira amostra deve ser pareada com a medida relativa da segunda amostra. Este método é ideal para o contexto, no qual o resultados das métricas avaliativas obtidas dependem dos dados fornecidos e comparadas especificamente para cada conjunto de dados (MCDONALD, 2009).

São necessários os seguintes passos para realizar o Teste T Pareado:

1. Estabelecer as hipóteses nula e alternativa. As hipóteses do teste de hipóteses que guiam as propostas a serem validadas para concluir se existe diferença entre as amostras, neste caso, dos algoritmos em comparação. As hipóteses são divididas em nula (H_0) e alternativa (H_a), na qual a hipótese nula representa que não existem diferenças entre as amostras e a hipótese alternativa visa provar a diferença existente entre as amostras de cada população.

Para o Teste T Pareado, conforme (OTT; LONGNECKER, 2010), considere dois conjuntos de amostras dependentes, X=$x_1, ..., x_n$ e Y=$y_1, ..., y_n$, de forma que obtêm-se os pares (x_1, y_1), ..., (x_n, y_n). Da diferença dos pares tem se o conjunto $\overline{D}$= $x_1 - y_1, ..., x_n - y_n$= $d_1, ..., d_n$, sendo que, o parâmetro μ será estimado pela média amostral das diferenças das amostras de cada população, ou seja, $\overline{D}$ (DEMSAR, 2006). Após a verificação da disposição dos dados define-se a hipótese dentre as possíveis indicadas abaixo, que difere da hipótese alternativa:

- ❏ H_0: $\mu = 0$ e H_a: $\mu > 0$;
- ❏ H_0: $\mu = 0$ e H_a: $\mu < 0$;
- ❏ H_0: $\mu = 0$ e H_a: $\mu \neq 0$.

2. Fixar o nível de significância α:

   A qualidade de um procedimento estatístico é avaliada de acordo com seu coeficiente de confiança. Ou seja, quando o coeficiente de confiança é 95%, é possível afirmar que a significância estatística de veracidade do teste é de 95%. Dessa forma, α = 1 - (coeficiente de confiança), para o exemplo citado, $\alpha = 0.05$.

3. Determinar a região crítica:

   Com um nível de significância α os pontos críticos são determinados de acordo com a distribuição dos dados e consequentemente sua região crítica. Os pontos críticos são obtidos por $t_{\alpha/2}$ e $-t_{\alpha/2}$ para o caso bilateral (H_0: $\mu = 0$ e H_a: $\mu \neq 0$), t_α para o caso unilateral à direita (H_0: $\mu = 0$ e H_a: $\mu > 0$) e $-t_\alpha$ para o unilateral à esquerda (H_0: $\mu = 0$ e H_a: $\mu < 0$).

4. Calcular T_{Calc} sob a hipótese nula:

   A partir do parâmetro $\overline{D}$ que é obtido pela diferença das médias, o μ_D é o valor da hipótese nula, n é o tamanho das amostras e o cálculo da variância amostral das diferenças que é dado por:

$$s_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}$$

   O teste T pareado é calculado conforme a equação a seguir:

$$T_{\text{Calc}} = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

5. Critério para rejeição ou não de H_0:

   Seja n a quantidade de amostras, o grau de liberdade (df) é dado por $n-1$, e o valor correspondente a região crítica para o df é obtida na tabela *TdeStudent* (A.1). No caso do teste bilateral, se $T_{\text{Calc}} > t_{\alpha/2}$ ou $T_{Calc} < -t_{\alpha/2}$ rejeitamos H_0, caso contrário, não rejeitamos H_0. Para o teste unilateral à direita: se $T_{\text{Calc}} > t_{\alpha}$ rejeitamos H_0, caso contrário, não rejeitamos H_0. E por fim, no teste unilateral à esquerda: se $T_{\text{Calc}} < -t_{\alpha}$ rejeitamos H_0, caso contrário não rejeitamos H_0.

6. O p-valor é dado por:

   - ❏ Teste Bilateral:
     $p - valor = \mathbb{P}[|t| > |T_{\text{Calc}}||H_0] = 2\mathbb{P}[t > |T_{\text{Calc}}||H_0]$.
   - ❏ Teste Unilateral à direita:
     p-valor = $\mathbb{P}[t > T_{\text{Calc}}|H_0]$
   - ❏ Teste Unilateral à esquerda:
     p-valor = $\mathbb{P}[t < T_{\text{Calc}}|H_0]$

   O p-valor corrobora a rejeição ou não definida no item anterior, além de avaliar o quão distante se está da hipótese nula (H_0).

7. O intervalo de confiança é dado por:

   - ❏ Teste Bilateral:
     $IC(\mu_D, 1-\alpha) = \left(\overline{D} - t_{\alpha/2}\frac{s_D}{\sqrt{n}}; \overline{D} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}\right)$
   - ❏ Teste Unilateral à direita:
     $IC(\mu_D, 1-\alpha) = \left(\overline{D} - t_{\alpha}\frac{s_D}{\sqrt{n}}; \infty\right)$
   - ❏ Teste Unilateral à esquerda:
     $IC(\mu_D, 1-\alpha) = \left(-\infty; \overline{D} + t_{\alpha}\frac{s_D}{\sqrt{n}}\right)$

   O intervalo de confiança garante o intervalo de superioridade entre as médias para o nível de significância escolhido.

E o Teste T Pareado pode ser mensurado tanto pelo *p-valor* comparado ao nível de significância α e ao grau de confiança definido para validade do limiar de convicção do teste, e por fim, o intervalo de confiança representa o intervalo da diferença que a hipótese alternativa afirmou de acordo com a distribuição dos dados.

Portanto, para o α escolhido, se o *p-valor* obtido for menor que α, a hipótese nula (H_0) é rejeitada e a hipótese alternativa (H_a) é considerada, caso contrário, a hipótese nula deve ser considerada. E o intervalo de confiança obtido fornece a variação do percentual de vantagem ou não de uma abordagem em relação a outra (OTT; LONGNECKER, 2010).

2.8 Considerações finais

Neste capítulo foram apresentados os conceitos teóricos que sustentaram o desenvolvimento deste trabalho. Assim como algumas das técnicas relatadas na literatura.

Capítulo 3

Semi-supervised OUTlier detection based on Hubness Neighborhood (SOUTH-N)

Este capítulo apresenta o método *Semi-supervised OUTlier detection based on Hubness Neighborhood (SOUTH-N)*. Este método foi desenvolvido com o objetivo de identificar e rotular *outliers* a partir de pouquíssimas informações sobre *outliers* existentes em repositórios de dados estruturados e parcialmente classificados. A dificuldade de falta de *inliers* como exemplos inerentes ao problema é naturalmente tratada pelo método proposto, que, consegue identificar *outliers* em grandes conjuntos com alta dimensionalidade. Essa abordagem foi escolhida devido a facilidade e praticidade em identificar alguns exemplos que se destoam dos demais ao oposto da dificuldade em rotular corretamente todas as variantes dos padrões encontrados nos dados.

O capítulo está organizado da seguinte maneira. A Seção 3.1 descreve as duas principais etapas do método proposto em duas etapas: a Seção 3.1.1 apresenta a estratégia desenvolvida para a obtenção da informação de semissupervisão (amostras positivas e negativas) que é empregada na segunda etapa do método; a Seção 3.1.2 descreve a subrotina responsável pela análise da informação de semissupervisão e pela identificação de *outliers* presentes no conjunto de dados em análise. Por fim, a Seção 3.2 apresenta as considerações finais do capítulo.

3.1 Método proposto

O método proposto fornece uma nova abordagem para detecção de *outliers* que reúne estratégias de semissupervisão, de estimativa de densidade baseada em pon-

tuações *hubness* e de agrupamento não binário (*fuzzy*) com o objetivo de contribuir para a análise de conjuntos de dados de alta dimensão. Tal método, denominado *Semi-supervised OUTlier detection based on Hubness Neighborhood (SOUTH-N)*, baseia-se na estratégia *learning from positive and unlabeled data* (LPU), apresentada mais detalhadamente na seção 2.1, que propõe o aprendizado a partir de poucos dados positivos (P) e não rotulados (U). Assim, na primeira fase, o método obtém a informação de semissupervisão que será usada na segunda fase. Nesta fase as amostras negativas são obtidas a partir da estratégia de análise de vizinhança baseada em informações *hubness* das amostras positivas (*outliers*) existentes no conjunto de dados. Na segunda fase do método, a informação de semissupervisão obtida na primeira fase (amostras positivas e negativas) é empregada para guiar o processo de detecção de *outliers* do conjunto de dados em análise (veja Figura 7). As duas fases do método SOUTH-N são descritas nas Seções 3.1.1 e 3.1.2.

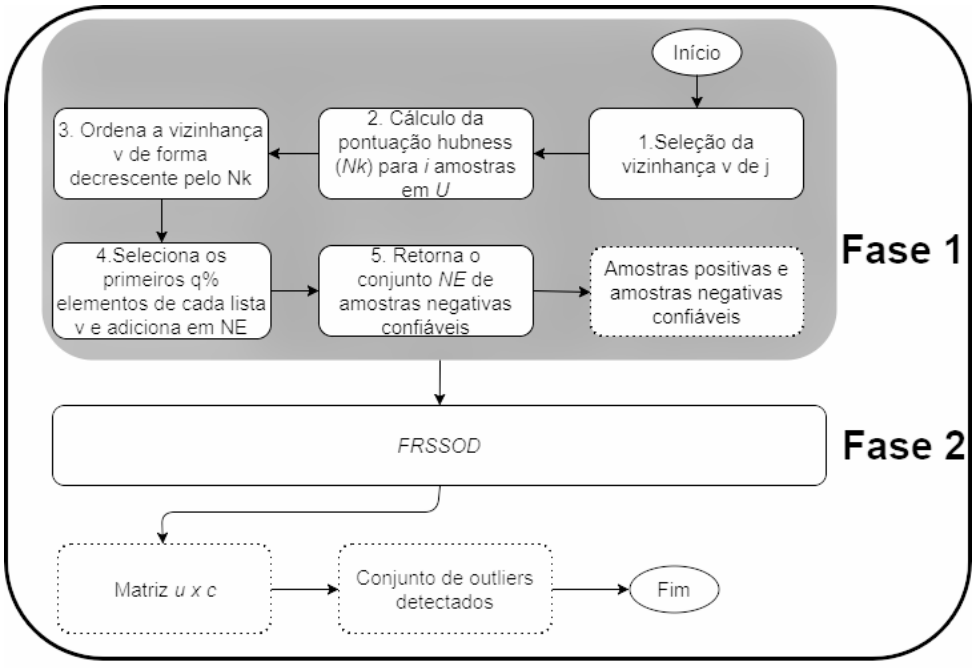

Figura 7 – Fluxograma do processo de detecção semissupervisionada de *outliers*.

3.1.1 Primeira Fase

Na primeira fase são realizadas as etapas indicadas no Algoritmo 3, resumidamente, a partir do conjunto de dados $X = \{x_1; ...; x_n\}$ para o qual é sabido existir um subconjunto $P = \{p_1; ...; p_j\} \subset X$ de amostras positivas (*outliers*) com $j < n$, de um conjunto de amostras não rotuladas U extrai q% de amostras negativas (*inliers*) confiáveis. Essa fase é ilustrada nas etapas 1 a 5 do fluxograma detalhado na Figura 7. Os passos realizados

pela primeira fase são detalhados a seguir.

Algoritmo 3: Primeira Fase

Entrada: Conjunto de amostras positivas P, Conjunto de amostras não rotuladas U, Número de vizinhos próximos k, percentual de exemplos de amostras negativas q

Saída: Conjunto de exemplos de amostras negativas NE

```
1  início
2      V ← {};
3      NK ← {};
4      NE ← {};
5      n_j ← {};
6      n_i ← {};
7      para cada instância p_j ∈ P faça
8          v_j ← Selecionar os k vizinhos mais próximos a p_j em U;
9          Inserir v_j em V;
10     para cada lista de vizinhos v_j ∈ V faça
11         para cada vizinho i ∈ v_j faça
12             n_i ← Calcular N_k(i);
13             Inserir n_i em n_j;
14         Inserir o par (v_j, n_j) em NK;
15     Ordenar de forma decrescente NK;
16     Inserir em NE os q% elementos iniciais v_j de NK;
17     retorna NE
18 fim
```

O primeiro passo do Algoritmo consiste em calcular a vizinhança v_j de cada amostra positiva $p_j \in P$ informada como entrada. Para tanto são selecionados os k vizinhos mais próximos de cada amostra positiva p_j no conjunto de amostras não rotuladas U, sendo $U = X - P$ (linha 7 do Algoritmo 3).

Após a seleção da vizinhança v_j de cada amostra positiva p_j, o próximo passo consiste em realizar o cálculo da pontuação *hubness* (N_k), computado conforme detalhado na Seção 2.5, para cada elemento da lista de vizinhança v_j e armazenar na lista NK o par (v_j, n_j), sendo n_j a pontuação *hubness* de v_j. É importante destacar que cada amostra positiva p_j possui uma lista NK com a pontuação *hubness* dos elementos de sua vizinhança (linha 10 do Algoritmo 3).

Por fim, para a definição da saída do Algoritmo, ou seja a lista NE, as listas NK são

ordenadas de forma decrescente, com relação ao valor da pontuação *hubness*, e uma combinação dos q% primeiros elementos de cada lista *NK* de cada amostra positiva (p_j) são adicionado a *NE*. Ou seja, os q% primeiros elementos não rotulados da vizinhança de cada amostra positiva (p_j) (linhas 15 e 16 do Algoritmo 3).

3.1.2 Segunda Fase

Essa Seção descreve o Algoritmo 4, *Fuzzy rough semi-supervised outlier detection (FRSSOD)* empregado na segunda fase do método *SOUTH-N*. O algoritmo *FRSSOD* foi proposto em (XUE; SHANG; FENG, 2010) e é usado para a detecção de *outliers* dos conjuntos de dados em análise. Esse algoritmo é baseado no algoritmo *FRCM* elaborado a partir da combinação dos algoritmos *FCM* [(DUNN, 1974),(BEZDEK, 1981)] e RCM [(LINGRAS; WEST, 2004), (PETERS, 2005)].

O *FRSSOD* necessita que sejam informados como entrada, exemplos de amostras positivas e negativas, o que dificulta a sua usabilidade no mundo real, uma vez que impõe a necessidade de supervisão para ambos tipos de amostras(DANESHPAZHOUH; SAMI, 2015). Entretanto o método proposto *SOUTH-N* auxilia a mitigar esse problema, pois necessita que sejam informadas como entrada apenas poucas amostras positivas. Os detalhes do Algoritmo *FRSSOD* são apresentados a seguir.

Segundo (XUE; SHANG; FENG, 2010) o algoritmo *FRSSOD* tem como objetivo obter uma matriz $U_{nxc} = \{u_{ik} | 1 \leqslant i \leqslant n, i \leqslant k \leqslant c\}$, na qual os valores para u_{ik} variando entre 0 e 1 indicam uma associação *fuzzy* da i-ésima instância ao k-ésimo grupo, considerando um conjunto de dados $X = \{x_1, ..., x_n\}$ com as primeiras l<n instâncias rotuladas com valores zero ou um, sendo que o valor zero indica se a instância é um *outlier* e o valor um indica o contrário (veja Equação 7).

$$\begin{cases} 0 < \sum_{k=1}^{c} u_{ik} \leqslant 1, & x_i \quad \text{é uma instância normal} \\ \sum_{k=1}^{c} u_{ik} = 0, & x_i \qquad\quad \text{é um } \textit{outlier} \end{cases} \tag{7}$$

Na Equação 7, se o somatório da linha for igual à zero significa que a instância em questão não tem chances de permanecer em nenhum agrupamento, caso contrário, a instância pertence a pelo menos um grupo.

O problema de otimização resolvido pelo FRSSOD é o apresentado na Equação 8.

$$minJ_m(u,v) = \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m d_{ik}^2 + \gamma_1\left(n - \sum_{i=1}^{n}\left(\sum_{k=1}^{c}(u_{ik})\right)^m\right) + \gamma_2\sum_{i=1}^{l}\left(y_i - \sum_{k=1}^{c}(u_{ik})\right)^2 \tag{8}$$

Algoritmo 4: Fuzzy rough semi-supervised outlier detection (FRSSOD)

Entrada: Conjunto de dados parcialmente rotulados X; Número de clusters c; Parâmetros $\zeta, \gamma_1, \gamma_2$; Índice *fuzzy* exponencial m; Critério de parada ϵ

Saída: Matriz u elementos por c clusters

início

  para *cada instância* $z \in \{0, ..., l\}$ **faça**

    para $l = 0$, inicialize os centros $v_k^{(l)} = v_j, k = 1, 2, ..., c$ com instâncias normais parcialmente rotuladas;

    atribua aleatóriamente cada instância de dados a exatamente uma aproximação inferior de acordo com as propriedades definidas em (PETERS, 2006);

    para *cada instância* $x_i \in X$ **faça**

      O proposto no Algoritmo 5;

    para *cada instância na linha* $i \in \{1, ..., n\}$ **faça**

      para *cada instância na coluna* $k \in \{1, ..., c\}$ **faça**

        se $x_i \in \underline{C_k}$ **então**

          $u_{ik} = 1$

        se $x_i \in C_k^B$ **então**

          $//C_k^B$ (Região de borda do agrupamento k) = $\overline{C_k} - \underline{C_k}$;

          $u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ij}}{d_{jk}}\right)^{\frac{2}{m-1}}}$;

    verifique x_i

    caso *não rotulado e localizado na região de borda*

      se $\sum_{k=1}^{c} (u_{ik})^m d_{ik}^2 > \gamma_1 \left(\sum_{k=1}^{c} u_{ik}\right)^m$ **então**

        marque x_i como *outlier*;

        atualize u_{ik} para u'_{ik} $//u_{ik} = 0$;

    caso *rotulado como normal e localizado na região de borda*

      se $\sum_{k=1}^{c} (u_{ik})^m d_{ik}^2 > \gamma_1 \left(\sum_{k=1}^{c} u_{ik}\right)^m + \gamma_2$ **então**

        marque x_i como *outlier*;

        atualize u_{ik} para u'_{ik} $//u_{ik} = 0$;

    caso *rotulado como outlier e localizado na região de borda*

      se $\sum_{k=1}^{c} (u_{ik})^m d_{ik}^2 > \gamma_1 \left(\sum_{k=1}^{c} u_{ik}\right)^m - \gamma_2 \left(\sum_{k=1}^{c} u_{ik}\right)^2$ **então**

        marque x_i como *outlier*;

        atualize u_{ik} para u'_{ik} $//u_{ik} = 0$;

    para *cada instância* $k \in \{1, ..., c\}$ **faça**

      calcule $v_k^{(l+1)} = \frac{\sum_{i=1}^{n} \left(u'_{ik}\right)^m x_i}{\sum_{i=1}^{n} \left(u'_{ik}\right)^m}$;

    verifique a convergência do Algoritmo. Se o Algoritmo convergir, **pare**, senão $l = l + 1$ e recomece;

    retorna NE

fim

O objetivo é encontrar a minimização dos termos. No primeiro termo da Equação 8, o somatório mais interno representa a o somatório das distâncias de cada elemento i do conjunto de dados em relação ao centroide do grupo, ponderado sobre u_{ik}. A minimização desse termo garante um melhor agrupamento, pois ele representa o erro. O segundo termo é responsável por garantir que o resultado não possua um número muito alto de *outliers*, sendo o parâmetro γ_1 responsável por ponderar a quantidade de *outliers*. Assim, para garantir a minimização, o somatório interno deve ser igual a n. Por fim, o terceiro termo, ponderado pelo parâmetro γ_2, garante que as instâncias já rotuladas mantenham os seus rótulos durante o processo. Ou seja, se o elemento y_i era rotulado como *outlier* o somatório da linha da matriz também deve ser zero ou um para as instâncias previamente rotuladas como amostras positivas. Assim, é gerada uma matriz *fuzzy nxc* (Número de amostras X *Cluster*), na qual o elemento E_{ij} da matriz representa o quanto a amostra i (linha) tem a probabilidade de pertencer ao grupo j (coluna).

Algoritmo 5: Algoritmo auxiliar para definição dos agrupamentos

```
1  início
2      para cada representante x_i ∈ X faça
3          para cada representante j ∈ {1, ..., c} faça
4              Calcule distância (x_i, v_j);
5          d_(l) = menordistncia_{1⩽k⩽c}(x_i, v_j);
6          para cada representante j ∈ {1, ..., c} faça
7              A = {∀j; j = 1, 2, ..., c; j ≠ h : d_ij^(l) / d_ih^(l) ⩽ ζ}
8              // O parâmetro de limiar ζ mede a distância relativa de um objeto x_i de
                   um par de clusters com centros v_j^(l) e v_h^(l);
9              se A ≠ ϕ então
10                 x_i ∈ C_h^(l); x_i ∈ C_j^(l); j = 1, 2, ..., c; j ≠ h; x_i ∉ C_k^(l); k = 1, 2, ..., c;
11                 x_i ∈ C_h^(l) e x_i ∈ C_h^(l);
12                 //C é o centro do agrupamento;
13                 // C_h é a aproximação superior do agrupamento h;
14                 // C_h é a aproximação inferior do agrupamento h;
15 fim
```

O primeiro passo do Algoritmo FRSSOD consiste em selecionar aleatoriamente os centros dos agrupamentos (linha 4 do algoritmo 4), que são usados para realizar o agrupamento inicial dos dados, feito conforme descrito no Algoritmo 5 (linha 5 do algoritmo 4). Para o entendimento do Algoritmo 5, é importante considerar as propri-

edades descritas em (LINGRAS; WEST, 2004):

- ❏ Uma instância pode ser um membro de uma aproximação inferior no máximo;
- ❏ Uma instância que é um membro da aproximação inferior de um agrupamento também é membro da aproximação superior do mesmo agrupamento;
- ❏ Uma instância que não pertence a qualquer aproximação inferior é membro de pelo menos duas aproximações superiores.

A partir da Figura 8 pode-se observar a ilustração do conceito de aproximação inferior e superior, assim como a ilustração da região de borda que é representada pela parte da aproximação superior $\overline{C_h}$ que não é coberta pela aproximação inferior $\underline{C_h}$ correspondente, ou seja, $C_h = \overline{C_h} - \underline{C_h}$;

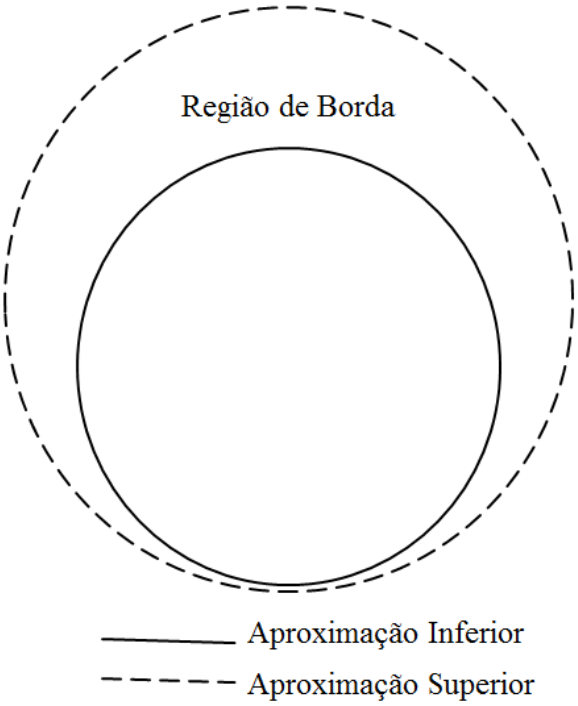

Figura 8 – Representação da região de borda, aproximação inferior e superior adaptada de [(PETERS, 2006)]

Depois da definição do agrupamento, a matriz U_{nxc} é inicializada de acordo com a Equação 9. Como as propriedades apresentadas em (LINGRAS; WEST, 2004) define que se uma instância pertence a uma aproximação inferior ($\underline{C_k}$) com certeza ela faz parte do agrupamento da mesma, a posição correspondente a essa instância (linha) será atribuído 1 no agrupamento correspondente (coluna). Caso contrário, a instância

pertence a região de borda (C_k^B) de um ou mais agrupamentos por isso o peso atribuído a esses agrupamentos serão ponderados (linhas 6 a 13 do Algoritmo 4).

$$u_{ik} = \begin{cases} 1, & x_i \in \underline{C_k} \\ \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ij}}{d_{jk}}\right)^{\frac{2}{m-1}}}, & x_i \in C_k^B \end{cases}, \quad i = 1, 2, ..., n; \quad k = 1, 2, ..., c; \tag{9}$$

Em seguida as instâncias presentes nas regiões de borda são analisadas para verificar a possibilidade dessas instâncias serem *outliers* pela ponderação dos dois últimos termos da Equação 8.

No primeiro caso explorado são verificadas as instâncias na região de borda não rotuladas, ou seja, como não é rotulada ela não vai interfirir no terceiro termo. Caso essa instancia seja classificada como *outlier* o primeiro termo da equação será reduzido, pois apenas os pontos normais são divididos em agrupamentos, de modo que *outliers* não contribuem para a soma de quadrados do erro calculada neste termo. Porém o segundo termo será aumentado por acrescentar mais um *outlier* aos resultados. Desse modo, se o valor diminuído pelo primeiro termo for maior que o valor aumentado pelo segundo termo essa instancia deve ser classificado como *outlier*, atualizando com o valor 0 todas as colunas da linha correspondente a essa instância na matriz (linha 15 a 18 do Algoritmo 4).

A segunda opção revê instâncias na região de borda e rotuladas como normal, nesse caso todos os termos da Equação 8 são afetados e a única maneira dessas instâncias serem classificadas como *outliers* será se a diminuição do primeiro termo for maior do que o aumento somado dos demais termos. Apenas nesse caso o valor 0 será atribuído a todas as colunas da linha correspondente a essa instância, definindo-a como *outlier* na matriz (linha 19 a 22 do Algoritmo 4).

No último caso são verificadas as instâncias na região de borda e rotuladas como normal, o que também interferirá em todos os termos da Equação 8, porém ela só deve ser realmente classificada se a diminuição do primeiro e terceiro termos forem mais significativas que o aumento provocado no segundo termo na equação. Assim, essa instância será classificada como *outlier* e a matriz atualizada (linha 23 a 26 do Algoritmo 4).

Após a ponderação dos elementos em relação aos grupos da matriz (linhas 15 a 26 do algoritmo 4), o centro dos grupos são recalculados de acordo com a média dos elementos em cada grupo (linha 27 do algoritmo 4), após os cálculos a convergência é verificada pelo parâmetro ϵ do critério de parada, que verifica se a cada iteração a

distância de alteração dos novos centros é menor que ϵ (linha 29 do Algoritmo 4). Caso não seja, o algoritmo passa para a próxima iteração ($l + 1$), caso contrário, o algoritmo é encerrado e a partir dessa representação final da matriz, podemos extrair os *outliers* apenas procurando as linhas da matriz em que o seu somatório é igual ao zero (linha 30 do Algoritmo 4).

3.2 Considerações finais

Este capítulo apresentou a descrição do método de detecção semissupervisionada de *outliers* para dados de alta dimensão, chamado *SOUTH-N*, desenvolvido no trabalho descrito aqui. Este algoritmo combina estratégias de semissupervisão, de estimativa de densidade baseada em pontuações *hubness* e técnicas de agrupamento não binário (*fuzzy*) a fim de obter uma eficácia maior na detecção de *outliers*. Essa técnica contribui em vários cenários como detecção de fraudes, intrusão de redes ou sistemas, doenças em exames médicos e etc (AGGARWAL, 2013).

O próximo Capítulo 4, apresenta a validação da proposta deste trabalho, que compara a eficácia do Método proposto SOUTH-N em relação a trabalhos semelhantes da literatura científica da área.

Capítulo 4

Experimentos e Análise dos Resultados

Este capítulo apresenta uma avaliação experimental do método proposto, o *SOUTH-N*. Os experimentos realizados tiveram três objetivos principais: i) comparar a eficácia dos resultados obtidos no método proposto em relação ao método da literatura selecionado como linha de base, o *SSODPU*, a partir de uma análise apurada com diferentes métricas; ii) selecionar uma gama variável de conjunto de dados para corroborar a eficácia do trabalho proposto na detecção de *outliers* e iii) comparar a eficácia do método *SOUTH-N* com diferentes percentuais de entradas de amostras positivas.

O capítulo está organizado da seguinte maneira: a Seção 4.1 descreve o método de avaliação empregado, com a listagem dos conjuntos de dados selecionados, o pré-processamento dos mesmos para os experimentos e a escolha dos parâmetros utilizados; a Seção 4.2 apresenta um experimento minucioso para atestar a veracidade da melhoria do método proposto em relação ao principal trabalho correlato; a Seção 4.3 descreve os resultados obtidos para todas as bases de dados, assim como o teste estatístico; e a Seção 4.4 especifica o comportamento do método SOUTH-N em relação a quantidade de amostras positivas de entrada. Por fim, a Seção 4.5 encerra o capítulo com as considerações finais.

4.1 Descrição do método de avaliação

Os experimentos foram realizados em um desktop Dell XPS-8700 com processador Intel QuadCore i5-4430 CPU@3.00GHz, 8GB de memória RAM e disco rígido SATA-III 1TB 7200 RPM, utilizando o compilador GNU gcc sobre Microsoft Windows 8 64-bits.

Todos os métodos analisados utilizam o parâmetro k que define a quantidade de vizinhos, o k foi alterado de 1 até 65 para cada método e o melhor k foi escolhido com base nos seus resultados. Conforme visto no Capítulo 3 o algoritmo proposto *SOUTH-N* é dividido em duas fases e exige a variação dos seguintes parâmetros de entrada: γ_1, γ_2 e ϵ provenientes do método *FRSSOD* adotado na segunda fase para definição do melhor resultado para cada conjunto de dados. Além disso, como visto no Capítulo 2 o método *SSODPU* proposto em (DANESHPAZHOUH; SAMI, 2015) também utiliza o algoritmo *FRSSOD*. Portanto a variação foi feita para ambos, de forma que tanto o *SOUTH-N*, quanto o *SSODPU* tenham a sua melhor combinação de parâmetros. Para essa ponderação de parâmetros o γ_1 variou de 0.001 até 0.1, incrementado 0.001; o γ_2 de 0.1 até 1, incrementando 0.1 e o ϵ começa de 1 até 3, variando 0.1. Com base nas informações de rótulos dos conjuntos de dados, foram selecionadas 30% das instâncias rotuladas como *outliers* de cada conjunto de dados para serem informadas como amostras positivas para os métodos semissupervisionados *SOUTH-N* e *SSODPU*. As implementações desses algoritmos foram realizadas na linguagem Java.

Levando em consideração o alto volume e dimensão dos dados em análise, para evitar qualquer aleatoriedade durante as execuções dos testes que possa favorecer algum algoritmo em análise, as métricas de avaliação foram calculadas 100 vezes, finalmente considerando apenas a média para análise dos resultados.

4.1.1 Conjuntos de dados

Foram selecionados 15 conjuntos de dados apresentado na Tabela 3, dos quais 1 é composto de dados sintéticos e 14 são compostos de dados reais. O conjunto de dados sintético é gerado pela função "*mvrnorm*"conforme (RIPLEY, 2017) na linguagem R (GENTLEMAN; IHAKA, 1997) pela interface gráfica do utilizador (GUI) chamada RStudio. Os conjuntos de dados reais foram obtidos no Repositório *UCI (University of California Irvine) Machine Learning* (LICHMAN, 2013). Parte desses conjuntos de dados foi escolhida a partir de trabalhos correlatos na tarefa de detecção de *outliers*, outros foram coletados para garantir maior variedade em relação ao números de classes, densidades e dimensionalidades, com o objetivo de avaliar o algoritmo proposto frente aos demais algoritmos em diferentes cenários.

A maioria dos conjuntos de dados possuem atributos com diferentes escalas, ou seja, valores em diferentes ordens de grandeza. Assim, como a função de distância usada foi a distância Euclidiana, foi necessário a normalização dos dados para impedir que a métrica fosse dominada pelo atributo de maior magnitude. Para isso, foi utilizada a fórmula apresentada na Equação 10, que considera o mínimo e o máximo

de cada atributo para reescalar os valores de todos os atributos para cada instância (AGGARWAL; ZHAI, 2012).

$$x_{novo}i = \frac{x(i) - min(i)}{max(i) - min(i)} \tag{10}$$

Tabela 3 – Conjuntos de dados considerados nos experimentos.

Nome do conjunto de Dados	Instâncias	Classes	Dimensões	Negativo	Positivo	Referência
Breast Cancer Wisconsin	699	2	10	458	241	(UCI, 1992)
Cardiotocography- CTG	2.126	10	10	1.950	176	(UCI, 2010)
Ecoli	336	8	8	307	29	(UCI, 1996)
Forest type mapping	326	4	27	289	37	(UCI, 2015)
Glass Identification	214	7	10	163	51	(UCI, 1987a)
Ionosphere	351	2	32	225	126	(UCI, 1989)
New Thyroid	215	3	6	150	65	(UCI, 1987c)
Parkinsons	195	2	23	147	48	(UCI, 2008)
Spambase	4.601	2	57	1.813	2.788	(UCI, 1999)
Statlog (Vehicle Silhouettes)	946	4	18	720	226	(UCI, 1987b)
SPECTF Heart	267	2	44	212	55	(UCI, 2001)
Synthetic	500	3	3	485	15	Autoria própria
Wine	178	3	13	130	48	(UCI, 1991)
Zoo	101	7	17	91	10	(UCI, 1990)

4.1.2 Abordagens Concorrentes

O principal objetivo dos algoritmos baseados na metodologia LPU é apresentar um bom resultado com uma quantidade pequena de instâncias positivas para treinamento. Para avaliar a eficácia do algoritmo proposto para detecção semissupervisionada de *outliers*, o *SOUTH-N*, foi considerado o outro algoritmo da literatura que segue a metodologia LPU e também utiliza uma abordagem semissupervisionada, o *SSODPU* (DANESHPAZHOUH; SAMI, 2015). Além disso, também foram selecionados alguns algoritmos não supervisionados do estado da arte da literatura, conforme descrito a seguir.

Um dos algoritmos não supervisionados escolhido foi o *Angle-based Outlier Detection (ABOD)*. Para cada instância é verificado o ângulo com todos os pares das demais instâncias do conjunto de dados e a partir disso é definido um fator de *outlier* local. De forma, que o ângulo dos pares é ponderado menos se as instâncias correspondentes estiverem distantes da instância em análise e para o valor final é utilizado a ponderação da variância para constituir o fator angular de *outliers* (*ABOF*) de cada instância. Essa estratégia apresenta bons resultados em bases com altas dimensões, porém ela é inviável para grandes conjuntos de dados, uma vez que para cada elemento todos os pares de elementos devem ser considerados (ANGLE-BASED. . . , 2008). A implementação usada desse algoritmo foi obtida em (JIMENEZ, 2015) na biblioteca 'abodOutlier'

Tabela 4 – Conjunto de dados *Wisconsin breast cancer* modificado.

Nome da Classe	Rótulo da Classe	% de Instâncias	Instâncias Conhecidas
Benignas (Classe comum)	Negativa	97,28% (357)	0
Malignas (Classe Rara)	Positiva	2,72% (10)	3

disponível na linguagem *R*.

Outro algoritmo escolhido foi o *Local Outlier Factor (LOF)* (BREUNIG et al., 2000) baseado em distância e densidade local. Sua estratégia consiste em definir um fator de *outlier* local baseado em densidade para cada instância referente a um grau de *outlierness* e comparar a densidade local de cada elemento com as densidades locais dos elementos de sua vizinhança, cuja distância é usada para estimar a densidade. Assim, é possível identificar instâncias que são substancialmente menos densas que seus vizinhos, que são considerados *outliers*. A implementação usada desse algoritmo foi obtida em (HU; SHAN, 2015)na biblioteca '*Rlof*' disponível na linguagem *R*. *ABOD* e *LOF* são citados como estado-da-arte para detecção de *outliers* em dados de alta dimensionalidade.

4.2 Conjunto de dados *Wisconsin breast cancer data* modificado

O primeiro experimento realizado visa corroborar a superioridade do algoritmo proposto *SOUTH-N* frente ao principal algoritmo concorrente da literatura *SSODPU*, por intermédio de diferentes métricas. Para esses experimentos foi adotado o conjunto de dados *Wisconsin Breast Cancer* com dados de câncer de mama da *University of Wisconsin Hospitals* e modificado conforme a referência (DANESHPAZHOUH; SAMI, 2015).

Após essa modificação, o conjunto de dados *Wisconsin Breast Cancer* passou a possuir 367 instâncias, das quais 357 são benignas e 10 são registros de instâncias malignas. Apenas 3 instâncias são passadas como entrada de exemplos positivos para os algoritmos semissupervisionados, e para assegurar a veracidade dos resultados as 3 instâncias de entrada são escolhidas aleatoriamente 100 vezes. O parâmetro k que define a quantidade de vizinhos foi alterado de 1 até 65 para cada algoritmo e o melhor k foi escolhido com base nos seus resultados. O conjunto de dados *Wisconsin Breast Cancer* modificado é apresentado na Tabela 4.

Foram escolhidas três métricas diferentes para analisar os resultados. A primeira métrica representa a curva precisão pela revocação, a segunda métrica apresenta o gráfico da precisão em relação ao tamanho do k passado para definir a quantidades de

vizinhos e por fim a Medida F para comparação dos métodos.

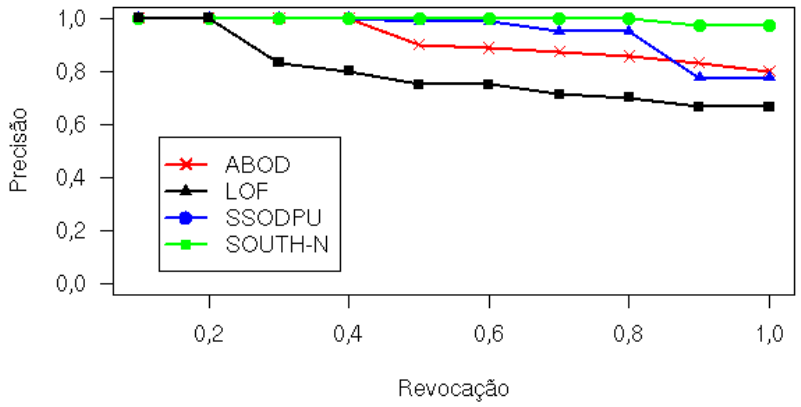

Figura 9 – Representação do gráfico da precisão em relação a revocação

O gráfico de precisão e revocação (calculado de acordo com o apresentado na Seção 2.7.1.3) apresenta no eixo y a precisão e no eixo x a revocação. Para a representação dos métodos no gráfico, foi analisado a precisão e a revocação dos dez *outliers* presentes no banco de dados na saída dos métodos. Analisando os gráficos das Figuras 9 e 10 é possível observar que os métodos semissupervisionados *SOUTH-N* e *SSODPU* apresentam melhor eficácia em relação aos métodos não supervisionados do estado da arte selecionados. Além disso, o método proposto *SOUTH-N* apresenta vantagem em relação ao seu principal concorrente da literatura, conforme pode ser observado na Figura 9.

É importante lembrar que tanto o método *SSODPU* quanto o método *SOUTH-N* utilizam o conceito de vizinhança definido a partir do parâmetro de entrada k. Assim, a precisão é calculada conforme apresentado na Equação 1 da Seção 2.7.1.1 para o k variando de 1 a 65 (veja Figura 10). É importante ressaltar que apesar da variação de k ser passada como parâmetro ambos os métodos não apresentam variação significativa em seus resultados com a alteração do k devido ao algoritmo utilizado na segunda fase (vide Seção 3.1.2).

A Figura 11 apresenta os resultados obtidos para a Medida F (veja a Seção 2.7.1.2). Essa medida faz o cálculo ponderado da precisão e da revocação analisando a eficácia em relação ao resultado dos métodos pelo rótulo dado a todos os *outliers* do banco de dados. Como pode ser visto na Figura 11 os métodos semissupervisionados apresentaram vantagem significativa em relação aos métodos do estado da arte para a detecção não supervisionada de *outliers*. Além disso, o método *SOUTH-N* apresentou superioridade considerável em relação a todos os demais métodos, inclusive em relação ao seu

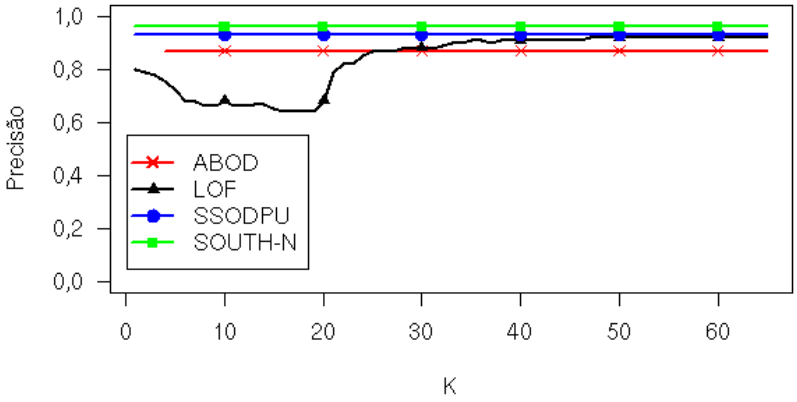

Figura 10 – Representação do gráfico da precisão em relação a variação da vizinhança *k*

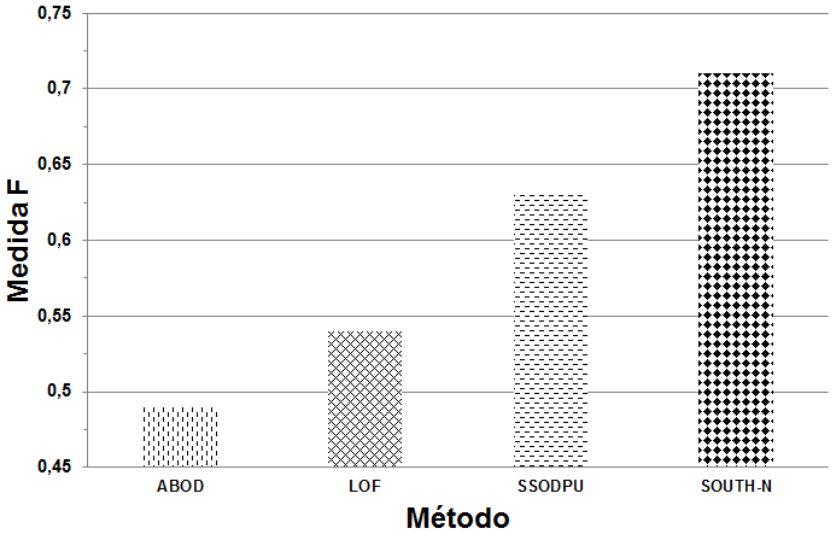

Figura 11 – Representação do gráfico da precisão em relação a revocação

principal concorrente *SSODPU*.

Os experimentos apresentados nessa Seção permitem mostrar que o método proposto neste trabalho permitiu obter melhores resultados que o método *SSODPU*. Além disso é importante ressaltar que o método proposto neste trabalho manteve o custo computacional comparado ao método concorrente *SSODPU*, pois ambos apresentaram tempos similares na execução. Na seção seguinte o objetivo é comprovar a eficácia do método para agregar as pesquisas para detecção semissupervisionada de *outliers*.

4.3 Comparação entre os algoritmos

Esta seção descreve os experimentos que demonstram a qualidade alcançada com a utilização do método *SOUTH-N* em diversos conjuntos de dados. A Seção 4.3.1 apresenta a comparação da métrica *AUC* para todos os métodos em análise. A Seção 4.3.2 apresenta o teste estatístico para corroborar a melhor eficácia do *SOUTH-N* em relação aos métodos selecionados como linha de base.

4.3.1 Área sob a curva

Os resultados experimentais apresentados nesta seção correspondem a análise da AUC obtida pelos métodos *SOUTH-N*, *SSODPU*, *ABOD* e *LOF*. Essa métrica permite avaliar o resultado apresentado por cada método ponderando tanto a quantidade de elementos classificados corretamente como *outliers* quanto a quantidade de elementos que deixaram de ser classificados corretamente como *outliers* pelo método.

Tabela 5 – Resultados dos experimentos, considerando a AUC. Os valores sublinhados destacam os melhores desempenhos.

Conjunto de Dados	SOUTH-N	SSODPU	ABOD	LOF
Breast Cancer Wisconsin	**0,792**	0,709	0,618	0,522
Cardiotocography- CTG	0,5	0,322	0,535	**0,621**
Ecoli	**0,726**	0,544	0,520	0,652
Forest type mapping Data Set	0,611	0,468	**0,739**	0,604
Glass Identification	**0,703**	0,631	0,432	0,586
Ionosphere	0,638	0,596	0,568	**0,642**
Modified Wisconsin	**0,82**	0,71	0,499	0,807
New Thyroid	**0,963**	0,824	0,492	0,669
Parkinsons	**0,774**	0,425	0,639	0,583
Spambase	**0,592**	0,349	0,580	0,512
Statlog (Vehicle Silhouettes)	**0,5**	0,422	0,434	0,486
Synthetic	0,65	0,574	0,551	**0,723**
Wine	**0,623**	0,381	0,559	0,516
Zoo	**0,745**	0,35	0,495	0,711
μ_D (Diferença das médias)	-	0,166	0,141	0,071
s_D (Diferença dos desvios padrão)	-	0,107	0,155	0,118

Os resultados obtidos são mostrados na Tabela 5. Analisando esses resultados é possível perceber que o método proposto *SOUTH-N* apresentou bons resultados para conjuntos de dados com diferentes características, como o *New Thyroid* que possui baixa dimensionalidade e o *Spambase* que possui alta dimensionalidade. De forma geral, considerando todos os conjuntos de dados o *SOUTH-N* foi superior aos demais

métodos em 10 conjuntos de dados do total de 14 conjuntos de dados analisados. Para comprovar a superioridade da eficácia do método *SOUTH-N* em relação aos demais foi realizado o Teste T Pareado descrito na Seção 4.3.2.

4.3.2 Teste estatístico

Para comprovar a melhora significativa entre a eficácia do método proposto *SOUTH-N* com os demais métodos, foi utilizado os teste estatístico Teste T Pareado (OTT; LONGNECKER, 2010) apresentado em detalhe nas Seção 2.7.2.1. Para isso considere as seguintes hipóteses de teste.

- H_0 (Hipótese Nula): $\mu_D = 0$. Os dois métodos apresentaram a mesma eficácia nos experimentos realizados.
- H_a (Hipótese Alternativa): $\mu_D > 0$. O método *SOUTH-N* apresentou eficácia superior aos métodos (*SSODPU*, *ABOD* ou *LOF*) nos experimentos realizados.

Para aplicar o Teste T Pareado considera-se as diferenças ($\overline{D}$), na qual μ_1 representa a média populacional da AUC obtida para todos os conjuntos de dados na execução do algoritmo proposto *SOUTH-N*, μ_2 representa a média populacional dos demais métodos, de acordo com os valores de cada par de medidas AUC da Tabela 5, considerando o nível de significância $\alpha = 0{,}05$.

Como os dados em análise possuem distribuições normais conforme a disposição apresentada no *boxplot* da Figura 12 e as populações de dados são independentes, o Teste T Pareado é indicado para a análise de significância estatística das distribuições.

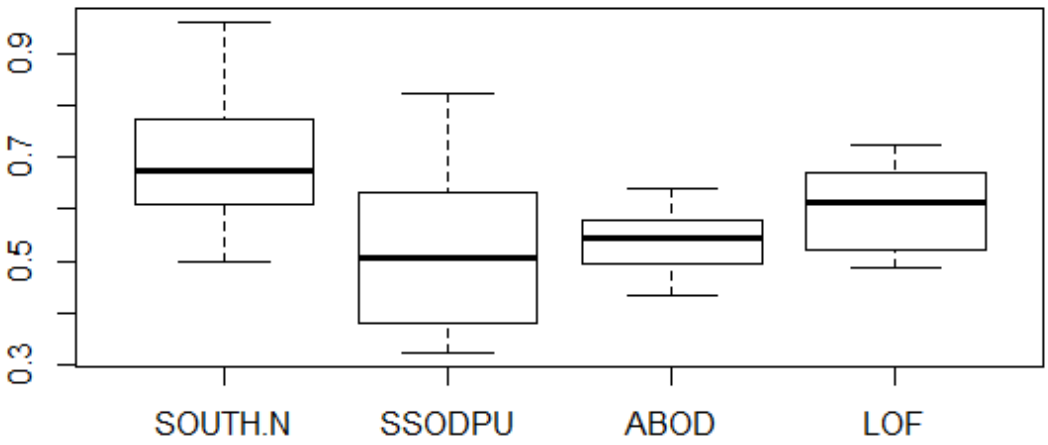

Figura 12 – Representação do boxplot da distribuição dos dados em análise

Para o Teste T Pareado com grau de confiança de 95%, foram considerados os valores dispostos da Tabela 5. A partir dessas informações, o Teste T Pareado foi usado para comparar a eficácia do método proposto *SOUTH-N* em relação aos métodos *SSODPU*, *ABOD* e *LOF*. Os cálculos realizados para o teste estatístico foram executados na linguagem R. Em todos os testes foram considerados o grau de confiança de 95%, ou seja, $\alpha = 0,05$, $n - 1 = 13$ graus de liberdade para a consulta na Tabela Pontos percentuais da distribuição de *Student's t* e $t_{0,05} = 1,771$. A seguir são apresentados os cálculos correspondentes para a comparação com cada método.

Para o método *SSODPU*, temos:

- Diferença das médias: $\mu_D = \overline{D} = 0,1665$;
- Desvio parão: $s_d = 0,1071$;
- $T_{\text{Calc}} = 5,8144$;
- p-valor = $3.015e - 05$;

Como o valor T_{calc} para a comparação do *SOUTH-N* com o *SSODPU* é $(5,8144)$ que é superior ao valor crítico $t_{0,05}$ e o $p - valor = 3.015e - 05$ é inferior a $0,05$ é possível rejeitar a hipótese nula H_0 e concluir com 95% de confiança que existe diferença significativa entre os resultados dos métodos *SOUTH-N* e *SSODPU*.

Para o método *ABOD*, temos:

- Diferença das médias: $\mu_D = \overline{d} = 0,1411$;
- Desvio parão: $s_d = 0,1553$;
- $T_{\text{Calc}} = 3,3993$;
- p-valor = $0,0047$;

Como $T_{\text{Calc}} = 3,3993$ é maior que 1,771, é possível concluir que o p-valor é menor que α ($0,0047 < 0,05$). Logo, na comparação estatística do método proposto *SOUTH-N* com o método *ABOD*, o *SOUTH-N* apresentou melhor eficácia em relação ao *ABOD* com 95% de confiança.

Para o método *LOF*, temos:

- Diferença das médias: $\mu_D = \overline{d} = 0,0716$;
- Desvio parão: $s_d = 0,1184$;
- $T_{\text{Calc}} = 2,2623$;

- p-valor = 0,0414;

Nesse caso, como $T_{\text{Calc}} = 2,2623 > 1,771$ e p-valor é menor que α (0,0414 < 0,05), portanto, o método proposto *SOUTH-N* é estatisticamente superior ao método *LOF* em relação a eficácia do mesmo para 95% de confiança.

Como a hipótese nula foi rejeitada em todos os casos, o Teste T Pareado permite concluir a superioridade do método proposto em relação aos demais empregados nos experimentos.

4.4 Avaliação da variação de parâmetros

Essa seção detalha os experimentos que foram realizados com a variação do percentual da quantidade de amostras positivas disponibilizadas na entrada dos dados com o objetivo de analisar o quanto o valor definido para esse parâmetro pode interferir na qualidade dos resultados.

4.4.1 Variação do percentual de amostras positivas

Uma das principais vantagens do método proposto é a necessidade de que sejam fornecidas poucas amostras positivas rotuladas como entrada. Para corroborar a afirmação de que mesmo com baixa quantidade de amostras positivas rotuladas o *SOUTH-N* mantém seus bons resultados, o experimento descrito aqui considera diferentes porcentagens de amostras positivas. Os resultados são mostrados na Tabela 6.

Tabela 6 – Resultados da acurácia para diferentes quantidades de amostras positivas como entrada.

	SOUTH-N							
	10%		20%		30%		40%	
Conjunto de Dados	Média	s_d	Média	s_d	Média	s_d	Média	s_d
Breast Cancer Wisconsin	0.836	0.003	0.820	0.004	0.805	0.004	0.791	0.005
Cardiotocography- CTG	0.917	0.0	0.917	0.0	0.917	0.0	0.917	0.0
Ecoli	0.870	0.002	0.868	0.003	0.865	0.003	0.862	0.004
Forest type mapping	0.566	0.002	0.556	0.003	0.545	0.029	0.531	0.051
Ionosphere	0.710	0.011	0.703	0.012	0.691	0.014	0.676	0.013
Modified Wisconsin	0.978	0.001	0.977	0.001	0.975	0.002	0.974	0.002
Statlog	0.764	0.0	0.764	0.0	0.764	0.0	0.764	0.0
Synthetic	0.946	0.002	0.946	0.0	0.946	0.0	0.946	0.0

Para essa análise foram considerados os conjuntos de dados: *Breast Cancer Wisconsin, Cardiotocography- CTG, Ecoli, Forest type mapping, Ionosphere, Modified Wisconsin, Statlog e Synthetic*. Para cada percentual considerado o método foi executado 100 vezes. A acurácia média e o desvio padrão (s_d) obtidos são apresentados na Tabela 6. Analisando os dados apresentados na Tabela 6 é possível notar que o método proposto *SOUTH-N* com apenas 10% mantém a qualidade do resultado semelhante ou superior, em alguns casos, aos experimentos com 40% de entrada de amostras positivas. Esses resultados corroboram a afirmação de que o método *SOUTH-N* apresenta bons resultados com pouquíssimos exemplos de amostras positivas como entrada.

4.5 Considerações finais

Neste capítulo verificou-se que o método de detecção semissupervisionada de *outliers* produziu bons resultados, agregando conhecimento a área foco dessa pesquisa. O próximo capítulo trás as considerações finais desta dissertação além das perspectivas de trabalhos futuros.

Capítulo 5

Conclusão

O trabalho apresentado nessa dissertação contribuiu com a criação de um novo método de detecção semissupervisionada de *outliers* compatível com dados de alta dimensionalidade. Dentre os pontos positivos desse método está o fato de que ele lida com a alta dimensionalidade usando o conceito *hubness* e, por isso, evita possíveis perdas de informação por que considera todas as dimensões dos conjuntos de dados em análise. Além disso, sua semissupervisão exige poucos exemplos de treinamento. Os resultados dos experimentos realizados com diferentes conjuntos de dados ajudam a corroborar a afirmação do bom desempenho do método *SOUTH-N*.

O trabalho desenvolvido gerou um artigo, que foi submetido, e está em processo de revisão, ao *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2017)* que é responsável por reunir pesquisadores das áreas de mineração de dados e aprendizado de máquina. O artigo apresenta o método proposto e os resultados encontrados no experimento que compara a *AUC* em relação aos demais métodos.

Como os resultados alcançados no trabalho descrito aqui mostraram a boa eficácia do novo método prosposto, isso motiva novas investigações e pesquisas para o aprimoramento do mesmo. Os trabalhos futuros vislumbrados são listados a seguir:

- ❏ Investigar modificações que aumentem a eficiência computacional do método;
- ❏ Considerar a inclusão do conhecimento de anti-*hubs* para aumentar a eficácia do método;
- ❏ Trabalhar novas estratégias que busquem refinar o método na detecção de *outliers* coletivos;
- ❏ Explorar a variação do parâmetro K no cálculo das pontuações *hubness* independente da variação do parâmetro k da quantidades de vizinhos desejados;

- ❏ Estudar uma maneira de incluir o conhecimento de *hubs* e *anti-hubs* na função de otimização apresentada na Equação 8 da Seção 3.1.2

Referências

AGGARWAL, C. C. **Outlier Analysis**. Springer International Publishing Switzerland, 2013. Disponível em: <http://dx.doi.org/10.1007/978-3-319-47578-3>.

AGGARWAL, C. C. **Data mining: The Textbook**. Springer International Publishing Switzerland, 2015. Disponível em: <http://dx.doi.org/10.1007/978-3-319-14142-8>.

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. On the surprising behavior of distance metrics in high dimensional spaces. In: **Proceedings of the 8th International Conference on Database Theory**. London, UK, UK: Springer-Verlag, 2001. (ICDT '01), p. 420–434. Disponível em: <http://dx.doi.org/10.1007/3-540-44503-x_27>.

AGGARWAL, C. C.; YU, P. S. Outlier detection for high dimensional data. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 30, n. 2, p. 37–46, maio 2001. Disponível em: <http://dx.doi.org/10.1145/375663.375668>.

AGGARWAL, C. C.; ZHAI, C. X. **Mining Text Data**. Springer Publishing Company, Incorporated, 2012. ISBN 1461432227, 9781461432227. Disponível em: <http://dx.doi.org/10.1007/978-1-4614-3223-4>.

AGRAWAL, A.; KUMAR, S.; MISHRA, A. A novel approach for credit card fraud detection. In: **Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on**. [S.l.: s.n.], 2015. p. 8–11.

ANGLE-BASED Outlier Detection in High-dimensional Data. In: PROCEEDINGS of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2008. (KDD '08), p. 444–452. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1401946>.

ARTHUR, D.; MANTHEY, B.; RöGLIN, H. Smoothed analysis of the k-means method. **J. ACM**, ACM, New York, NY, USA, v. 58, n. 5, p. 19:1–19:31, out. 2011. ISSN 0004-5411. Disponível em: <http://dx.doi.org/10.1145/2027216.2027217>.

BARNETT, V.; LEWIS, T. **Data mining and analysis: fundamental concepts and algorithms**. [S.l.]: John Wiley & Sons, 1994.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BREUNIG, M. M. et al. Lof: Identifying density-based local outliers. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 29, n. 2, p. 93–104, maio 2000. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/335191.335388>.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 41, n. 3, p. 15:1–15:58, jul. 2009. ISSN 0360-0300. Disponível em: <http://dx.doi.org/10.1007/978-1-4899-7502-7_912-1>.

CLARKE, E. F. B.; ZHANG, H. H. **Principles and Theory for Data Mining and Machine Learning**. Springer, 2009. Disponível em: <http://dx.doi.org/10.1007/978-0-387-98135-2_7>.

DANESHPAZHOUH, A.; SAMI, A. Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering. In: **Information and Knowledge Technology (IKT), 2013 5th Conference on**. [s.n.], 2013. p. 344–348. Disponível em: <http://dx.doi.org/10.1109/ikt.2013.6620091>.

DANESHPAZHOUH, A.; SAMI, A. Entropy-based outlier detection using semi-supervised approach with few positive examples. **Pattern Recognition Letters**, v. 49, p. 77–84, 2014. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2014.06.012>.

DANESHPAZHOUH, A.; SAMI, A. Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering. **International Journal on Artificial Intelligence Tools**, v. 24, n. 03, 2015. Disponível em: <http://dx.doi.org/10.1109/ikt.2013.6620091>.

DASGUPTA, D.; MAJUMDAR, N. Anomaly detection in multidimensional data using negative selection algorithm. In: **Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on**. [s.n.], 2002. v. 2, p. 1039–1044. Disponível em: <http://dx.doi.org/10.1109/cec.2002.1004386>.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, n. 1, p. 1–30, 2006.

DUNN, J. C. Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems. **Journal of Cybernetics**, v. 4, n. 2, p. 1–15, 1974. Disponível em: <http://dx.doi.org/10.1080/01969727408546062>.

DUONG, N. H.; HAI, H. D. A semi-supervised model for network traffic anomaly detection. In: **Advanced Communication Technology (ICACT), 2015 17th International Conference on**. [s.n.], 2015. p. 70–75. Disponível em: <http://dx.doi.org/10.1109/icact.2015.7224759>.

DUONG, N. H.; HAI, H. D. A model for network traffic anomaly detection. In: **2016 18th International Conference on Advanced Communication Technology (ICACT)**. [s.n.], 2016. p. 644–650. Disponível em: <http://dx.doi.org/10.1109/icact.2016.7423586>.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen-LTC, 2011.

FASSETTI, F.; ANGIULLI, F. Finding distance-based outliers in subspaces through both positive and negative examples. In: **ICAART 2010 - Proceedings of the International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial**

Intelligence, Valencia, Spain, January 22-24, 2010. [s.n.], 2010. p. 5–10. Disponível em: <http://dx.doi.org/10.5220/0002699600050010>.

FLEXER, A. An empirical analysis of hubness in unsupervised distance-based outlier detection. In: **IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain.** [s.n.], 2016. p. 716–723. Disponível em: <http://dx.doi.org/10.1109/icdmw.2016.0106>.

GAO, J.; CHENG, H.; TAN, P.-N. Semi-supervised outlier detection. In: **Proceedings of the 2006 ACM Symposium on Applied Computing**. New York, NY, USA: [s.n.], 2006. p. 635–636. Disponível em: <http://dx.doi.org/10.1145/1141277.1141421>.

GASPAR, J.; LOPES, F.; FREITAS, A. An analysis of hospital coding in portugal: Detection of patterns, errors and outliers in female breast cancer episodes. In: **Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on**. [S.l.: s.n.], 2011. p. 1–6.

GENTLEMAN, R.; IHAKA, R. **The R Project for Statistical Computing**. 1997. <https://www.r-project.org/>. Accessed: May, 2017.

GULER, P.; TEMIZEL, A.; TEMIZEL, T. An unsupervised method for anomaly detection from crowd videos. In: **Signal Processing and Communications Applications Conference (SIU), 2013 21st**. [s.n.], 2013. p. 1–4. Disponível em: <http://dx.doi.org/10.1109/siu.2013.6531292>.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2. ed. [S.l.]: Morgan Kaufmann Publishers Inc., 2006.

HAN MICHELINE KAMBER, J. P. J. **Data mining and analysis: Concepts and algorithms**. Elsevier, 2011. Disponível em: <http://dx.doi.org/10.1016/b978-0-12-381479-1.00010-1>.

HAUTAMAKI, V.; KARKKAINEN, I.; FRANTI, P. Outlier detection using k-nearest neighbour graph. In: **Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03**. Washington, DC, USA: IEEE Computer Society, 2004. (ICPR '04), p. 430–433. ISBN 0-7695-2128-2. Disponível em: <http://dx.doi.org/10.1109/icpr.2004.1334558>.

HAWKINS, D. M. **Identification of outliers**. [S.l.]: Chapman & Hall, 1980.

HE, Z.; DENG, S.; XU, X. An optimization model for outlier detection in categorical data. In: **Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I**. [s.n.], 2005. p. 400–409. Disponível em: <http://dx.doi.org/10.1007/11538059_42>.

HE, Z.; XU, X.; DENG, S. A unified subspace outlier ensemble framework for outlier detection in high dimensional spaces. **CoRR**, abs/cs/0505060, 2005. Disponível em: <http://dx.doi.org/10.1007/11563952_56>.

HEYLEN, R.; PARENTE, M.; SCHEUNDERS, P. Estimation of the intrinsic dimensionality in hyperspectral imagery via the hubness phenomenon. In: **Latent Variable Analysis and Signal Separation - 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings**. [s.n.], 2017. p. 357–366. Disponível em: <http://dx.doi.org/10.1007/978-3-319-53547-0_34>.

HOULE, M. E. et al. Scientific and statistical database management: 22nd international conference, ssdbm 2010, heidelberg, germany, june 30–july 2, 2010. proceedings. In: ____. Berlin, Heidelberg: [s.n.], 2010. cap. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?, p. 482–500. Disponível em: <http://dx.doi.org/10.1007/978-3-642-13818-8_34>.

HU, W. M. Y.; SHAN, Y. **Local Outlier Factor**. 2015. <https://cran.r-project.org/web/packages/Rlof/index.html>. Accessed: Jun, 2017.

HUBNESS-BASED Clustering of High-Dimensional Data. In: PARTITIONAL Clustering Algorithms. Springer International Publishing, 2015. p. 353–386. Disponível em: <http://dx.doi.org/10.1007/978-3-319-09259-1_11>.

IENCO, D.; PENSA, R. G.; MEO, R. A semisupervised approach to the detection and characterization of outliers in categorical data. **IEEE Trans. Neural Netw. Learning Syst.**, v. 28, n. 5, p. 1017–1029, 2017. Disponível em: <http://dx.doi.org/10.1109/tnnls.2016.2526063>.

JIANG, S.-Y.; YANG, A.-m. Framework of clustering-based outlier detection. In: **Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on**. [s.n.], 2009. v. 1, p. 475–479. Disponível em: <http://dx.doi.org/10.1109/fskd.2009.94>.

JIMENEZ, J. **Angle-Based Outlier Detection**. 2015. <https://cran.r-project.org/web/packages/abodOutlier/index.html>. Accessed: Jun, 2017.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. <http://archive.ics.uci.edu/ml>. Accessed: May, 2017.

LINGRAS, P.; WEST, C. Interval set clustering of web users with rough k-means. **J. Intell. Inf. Syst.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 23, n. 1, p. 5–16, jul. 2004. ISSN 0925-9902. Disponível em: <http://dx.doi.org/10.1023/b:jiis.0000029668.88665.1a>.

MANEVITZ, L. M.; YOUSEF, M. One-class svms for document classification. **J. Mach. Learn. Res.**, JMLR.org, v. 2, p. 139–154, mar. 2002. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944790.944808>.

MARQUES, G. C. **Machine learning techniques for music information retrieval.** Tese (Doutorado) — Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática, Lisboa, Portugal, 2015.

MASUD, M. et al. Classification and adaptive novel class detection of feature-evolving data streams. **Knowledge and Data Engineering, IEEE Transactions on**, v. 25, n. 7, p. 1484–1497, jul. 2013. Disponível em: <http://dx.doi.org/10.1109/tkde.2012.109>.

MCDONALD, J. H. **Handbook of Biological Statistics**. Second. Baltimore, Maryland, USA: Sparky House Publishing, 2009. Disponível em: <http://udel.edu/~{}mcdonald/statintro.ht>.

MITCHELL, T. M. **Machine Learning**. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

MONTES, M. C. Depth-based outlier detection algorithm. In: **Hybrid Artificial Intelligence Systems - 9th International Conference, HAIS 2014, Salamanca, Spain, June 11-13, 2014. Proceedings**. [s.n.], 2014. p. 122–132. Disponível em: <http://dx.doi.org/10.1007/978-3-319-07617-1_11>.

NOBLE, C. C.; COOK, D. J. Graph-based anomaly detection. In: **Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2003. (KDD '03), p. 631–636. ISBN 1-58113-737-0. Disponível em: <http://dx.doi.org/10.1145/956804.956831>.

OTT, R. L.; LONGNECKER, M. **An Introduction to Statistical Methods and Data Analysis**. 6. ed. Cengage Learning, 2010. Disponível em: <http://dx.doi.org/10.2307/1269399>.

PETERS, G. Outliers in rough k-means clustering. In: **Proceedings of the First International Conference on Pattern Recognition and Machine Intelligence**. Berlin, Heidelberg: Springer-Verlag, 2005. (PReMI'05), p. 702–707. ISBN 3-540-30506-8, 978-3-540-30506-4. Disponível em: <http://dx.doi.org/10.1007/11590316_113>.

PETERS, G. Some refinements of rough k-means clustering. **Pattern Recogn.**, Elsevier Science Inc., New York, NY, USA, v. 39, n. 8, p. 1481–1491, ago. 2006. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2006.02.002>.

RADOVANOVIć, M.; NANOPOULOS, A.; IVANOVIć, M. Hubs in space: Popular nearest neighbors in high-dimensional data. **J. Mach. Learn. Res.**, JMLR.org, v. 11, p. 2487–2531, dez. 2010. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1756006.1953015>.

RADOVANOVIC, M.; NANOPOULOS, A.; IVANOVIC, M. Reverse nearest neighbors in unsupervised distance-based outlier detection. **Knowledge and Data Engineering, IEEE Transactions on**, v. 27, n. 5, p. 1369–1382, maio 2015. Disponível em: <http://dx.doi.org/10.1109/tkde.2014.2365790>.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 29, n. 2, p. 427–438, maio 2000. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/335191.335437>.

RIPLEY, B. **The R Project for Statistical Computing**. 2017. <https://cran.r-project.org/web/packages/MASS/MASS.pdf>. Accessed: May, 2017. Disponível em: <http://dx.doi.org/10.11120/msor.2001.01010023>.

SAMET, H. **Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0123694469.

SCHNITZER, D. et al. Local and global scaling reduce hubs in space. **J. Mach. Learn. Res.**, JMLR.org, v. 13, n. 1, p. 2871–2902, out. 2012. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=2503308.2503333>.

SHYU, M.-L. et al. A novel anomaly detection scheme based on principal component classifier. In: **IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with ICDM'03**. [s.n.], 2003. p. 171–179. Disponível em: <http://dx.doi.org/10.1007/11539827_18>.

SINGH, K.; UPADHYAYA, D. S. **Outlier Detection: Applications And Techniques**. 2012.

SONG, Y.; CAO, L. Graph-based coupled behavior analysis: A case study on detecting collaborative manipulations in stock markets. In: **Neural Networks (IJCNN), The 2012 International Joint Conference on**. [s.n.], 2012. p. 1–8. Disponível em: <http://dx.doi.org/10.1109/ijcnn.2012.6252762>.

SRIVASTAVA, A. N.; ZANE-ULMAN, B. Discovering recurring anomalies in text reports regarding complex space systems. In: **2005 IEEE Aerospace Conference**. [s.n.], 2005. p. 3853–3862. Disponível em: <http://dx.doi.org/10.1109/aero.2005.1559692>.

SULIC, V. et al. Dimensionality reduction for distributed vision systems using random projection. In: IEEE. **Pattern Recognition (ICPR), 2010 20th International Conference on**. 2010. p. 380–383. Disponível em: <http://dx.doi.org/10.1109/icpr.2010.101>.

TAN, P.-N.; STEINBACH, M. **Vipin Kumar, Introduction to Data Mining**. [S.l.]: Addison Wesley, ISBN 0-321-32136-7, 2006.

TOMASEV, N.; BUZA, K. Hubness-aware knn classification of high-dimensional data in presence of label noise. **Neurocomput.**, Elsevier Science Publishers B. V., v. 160, n. C, p. 157–172, jul. 2015. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2014.10.084>.

UCI, M. L. R. **Glass Identification Data Set** . 1987. <https://archive.ics.uci.edu/ml/datasets/glass+identification>. Accessed: May, 2017.

UCI, M. L. R. **Statlog (Vehicle Silhouettes) Data Set**. 1987. <https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes)>. Accessed: May, 2017.

UCI, M. L. R. **Thyroid Disease Data Set**. 1987. <https://http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/>. Accessed: May, 2017.

UCI, M. L. R. **Ionosphere Data Set**. 1989. <https://https://archive.ics.uci.edu/ml/datasets/Ionosphere>. Accessed: May, 2017.

UCI, M. L. R. **Zoo Data Set**. 1990. <http://archive.ics.uci.edu/ml/datasets/zoo>. Accessed: May, 2017.

UCI, M. L. R. **Wine Data Set**. 1991. <https://archive.ics.uci.edu/ml/datasets/wine>. Accessed: May, 2017.

UCI, M. L. R. **Breast Cancer Wisconsin (Original) Data Set**. 1992. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)>. Accessed: May, 2017.

UCI, M. L. R. **Ecoli Data Set**. 1996. <http://archive.ics.uci.edu/ml/datasets/Ecoli?ref=datanews.io>. Accessed: May, 2017.

UCI, M. L. R. **Spambase Data Set**. 1999. <https://archive.ics.uci.edu/ml/datasets/Spambase>. Accessed: May, 2017.

UCI, M. L. R. **SPECTF Heart Data Set**. 2001. <https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart>. Accessed: May, 2017.

UCI, M. L. R. **Parkinsons Data Set**. 2008. <https://archive.ics.uci.edu/ml/datasets/Parkinsons>. Accessed: May, 2017.

UCI, M. L. R. **Cardiotocography Data Set**. 2010. <https://archive.ics.uci.edu/ml/datasets/Cardiotocography>. Accessed: May, 2017.

UCI, M. L. R. **Forest type mapping Data Set** . 2015. <https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping>. Accessed: May, 2017.

UMA Abordagem de Aprendizado de Máquina. [S.l.]: LTC, 2011.

VINAY, V. et al. A comparison of dimensionality reduction techniques for text retrieval. In: **Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on**. [s.n.], 2005. p. 293–298. Disponível em: <http://dx.doi.org/10.1109/icmla.2005.2>.

XUE, Z.; SHANG, Y.; FENG, A. Semi-supervised outlier detection based on fuzzy rough c-means clustering. **Math. Comput. Simul.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 80, n. 9, p. 1911–1921, maio 2010. ISSN 0378-4754. Disponível em: <http://dx.doi.org/10.1016/j.matcom.2010.02.007>.

YAGER, N.; DUNSTONE, T. The biometric menagerie. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 32, n. 2, p. 220–230, 2010.

YU, L.; CLEARY, D.; CUDDIHY, P. A novel approach to aircraft engine anomaly detection and diagnostics. In: **Aerospace Conference, 2004. Proceedings. 2004 IEEE**. [s.n.], 2004. v. 5, p. 3468–3475. Disponível em: <http://dx.doi.org/10.1109/aero.2004.1368152>.

ZAKI, M. J.; JR, W. M. **Data mining and analysis: fundamental concepts and algorithms**. [S.l.]: Cambridge University Press, 2014.

ZHANG, D.; LEE, W. S. A simple probabilistic approach to learning from positive and unlabeled examples. In: **Proceedings of the 5th Annual UK Workshop on Computational Intell igence (UKCI)**. [S.l.: s.n.], 2005.

ZHANG, J. Advancements of outlier detection: A survey. **EAI Endorsed Transactions on Scalable Information Systems**, ICST, v. 13, n. 1, fev. 2013. Disponível em: <http://dx.doi.org/10.4108/trans.sis.2013.01-03.e2>.

ZHENG, L. et al. Using data mining techniques to address critical information exchange needs in disaster affected public-private networks. In: **Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, 2010. (KDD '10), p. 125–134. Disponível em: <http://dx.doi.org/10.1145/1835804.1835823>.

Apêndices

APÊNDICE **A**

Tabelas Auxiliares

A.1 Pontos percentuais da distribuição de *Student's t*

A Tabela A.1 apresenta os valores críticos de acordo com a *F-Distribution* considerando $\alpha = 0,05$, isto é, com 95% de confiança. Mais informações sobre esses valores podem ser encontradas em (OTT; LONGNECKER, 2010).

STUDENT'S t PERCENTAGE POINTS

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090