Avaliação, seleção de modelos e detecção de outliers em espaços e subespaços de dados

Henrique Oliveira Marques

Avaliação, seleção de modelos e detecção de outliers em espaços e subespaços de dados

Henrique Oliveira Marques

Orientador: **Prof. Dr. Ricardo José Gabrielli Barreto Campello**

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

USP – São Carlos
Setembro de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Marques, Henrique Oliveira
M357a Avaliação, seleção de modelos e detecção de outliers em espaços e subespaços de dados / Henrique Oliveira Marques; orientador Ricardo José Gabrielli Barreto Campello. - São Carlos - SP, 2016.
70 p.

Monografia (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. detecção de outliers; avaliação interna; subespaços. I. Campello, Ricardo José Gabrielli Barreto, orient. II. Título.

RESUMO

MARQUES, H. O.. **Avaliação, seleção de modelos e detecção de outliers em espaços e subespaços de dados**. 2016. 70 f. Monografia (Doutorado em em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A área de detecção de *outliers* possui um papel fundamental na descoberta de padrões em dados que podem ser considerados excepcionais sob alguma perspectiva. Detectar tais padrões é relevante de maneira geral porque, em muitas aplicações de mineração de dados, tais padrões representam comportamentos extraordinários que merecem atenção especial. Uma importante distinção se dá entre as técnicas supervisionadas, semissupervisionadas e não supervisionadas de detecção. O presente trabalho enfoca principalmente as técnicas de detecção não supervisionadas. Existem dezenas de algoritmos desta categoria na literatura, porém cada um deles utiliza uma intuição própria do que deve ser considerado um *outlier*, que é naturalmente um conceito subjetivo. Isso dificulta sensivelmente a escolha de um algoritmo em particular e também a escolha de uma configuração adequada para o algoritmo escolhido em uma dada aplicação prática. Isso também torna altamente complexo avaliar a qualidade da solução obtida por um algoritmo/configuração em particular adotados pelo analista, especialmente em função da problemática de se definir uma medida de qualidade que não seja vinculada ao próprio critério utilizado pelo algoritmo. Tais questões estão inter-relacionadas e se referem respectivamente aos problemas de seleção de modelos e avaliação (ou validação) de resultados em aprendizado de máquina não supervisionado. Esses problemas têm sido investigados ao longo de décadas na área de agrupamento de dados, mas apenas recentemente uma medida interna e relativa para avaliação não supervisionada de soluções binárias (top-n) de detecção de *outliers*, chamada IREOS (*Internal, Relative Evaluation of Outlier Solutions*), foi proposta pelo próprio aluno no seu trabalho de mestrado. Ainda que a medida represente um importante avanço no estado-da-arte desta área, o desenvolvimento de medidas para soluções que, ao invés de rótulos binários, fornecem *scorings* para as observações (que é o tipo de solução produzida pela ampla maioria dos algoritmos bem conhecidos de detecção não supervisionada de *outliers*) e para soluções de *outliers* detectados em subespaços (que, devido ao problema da alta dimensionalidade, é uma área que recentemente vem recebendo bastante atenção) continuam como problemas em aberto na literatura. A avaliação de resultados produzidos por algoritmos de detecção não supervisionados, em especial os resultados produzidos por algoritmos que forneçam *scorings* e/ou façam detecção em subespaços dos dados, representa o principal objetivo deste projeto de pesquisa. A extensão de IREOS para a avaliação de resultados produzidos por tais categorias de algoritmos de detecção é uma forma pela qual se pretende atacar o problema. Também, como segundo objetivo, pretende-se investigar se princípios originais utilizados no desenvolvimento do índice IREOS podem também ser adaptados para o desenvolvimento de novos algoritmos de

detecção, em particular no contexto de subespaços. Finalmente, pretende-se ainda investigar a adaptação de métodos não supervisionados para operarem no contexto semissupervisionado.

Palavras-chave: detecção de outliers; avaliação interna; subespaços.

SUMÁRIO

LISTA DE ABREVIAÇÕES

ABOD *Angle-Based Outlier Detection* 24

ALOI *Amsterdam Library of Object Images* 45

AUC *Area Under the Curve* 37

fdp função densidade de probabilidade 27

GLOSH *Global-Local Outlier Scores from Hierarchies* 23

HOS-Miner *High-dimensional Outlying Subspaces* 26

IREOS *Internal, Relative Evaluation of Outlier Solutions* 14

KLR *Kernel Logistic Regression* 35

LOCI *Local Correlation Integral* 22

LOF *Local Outlier Factor* 21

kNNDD *k-Nearest Neighbour Data Description* 29

MDEF *Multi-Granularity Deviation Factor* 22

LP *Linear Programming* 29

OCC *One Class Classification* 27

PW *Parzen Windows* 28

ROC *Receiver Operating Characteristic* 34

SOD *Subspace Outlier Degree* 25

SVDD *Support Vector Data Description* 29

CAPÍTULO

1

INTRODUÇÃO

A tecnologia da informática tem evoluído de forma extraordinária. As capacidades de processamento e armazenamento de sistemas de computadores têm aumentado ordens de magnitude durante as últimas décadas. Concomitantemente, a capacidade de armazenamento tem aumentado ainda mais rápido. Em consequência desses avanços, o custo da tecnologia tem diminuído, facilitando o acesso a smartfones, RFID, sistemas de vigilância, entre outras tecnologias que produzem uma quantidade enorme de dados. Porém, extrair informação útil a partir de dessas grandes quantidades de dados não é uma tarefa trivial. O cenário de superabundância de dados tem aumentado cada vez mais nos últimos anos, na medida em que a capacidade de coletar e armazenar dados já ultrapassou em muito a capacidade humana para analisar e extrair conhecimento a partir deles (FAYYAD *et al.*, 1996). A impossibilidade de analisar esses dados de forma manual proporcionou um ambiente adequado para a aplicação de tecnologias emergentes de informação capazes de sintetizar, processar e transformar dados em conhecimento útil de uma forma inteligente e automatizada. Por esta razão, pesquisadores de diversas áreas têm se empenhado no estudo de métodos para o que tem sido chamado de mineração de dados ou *data mining*.

O termo mineração de dados (FAYYAD *et al.*, 1996; HAN; KAMBER, 2006) refere-se amplamente a técnicas estatísticas, matemáticas e computacionais de análise para extrair conhecimento útil de grandes conjuntos de dados de forma eficiente. Há algumas vertentes principais em mineração de dados que juntas cobrem a maioria dos problemas existentes em aplicações práticas (TAN; STEINBACH; KUMAR, 2006): agrupamento de dados, classificação de padrões, análise de associação, análise de regressão e detecção de *outliers*. Dentre tais vertentes, a área de detecção de *outliers* (ou detecção de anomalias, como também é conhecida) possui um papel fundamental na descoberta de padrões que podem ser considerados excepcionais sob alguma perspectiva (HODGE; AUSTIN, 2004; PATCHA; PARK, 2007; HADI; IMON; WERNER, 2009; CHANDOLA; BANERJEE; KUMAR, 2009; CHANDOLA; BANERJEE; KUMAR, 2012; ZIMEK; SCHUBERT; KRIEGEL, 2012). Detectar tais padrões em dados é

relevante por diversas razões, entre as principais estão: (i) em algumas aplicações, tais padrões representam dados espúrios (e. g., falhas ou ruídos em sensores) que se deseja eliminar em uma etapa de pré-processamento; (ii) existem aplicações em que se deseja disparar um alarme quando um evento anômalo é detectado, como em detecção de ataques virtuais e detecção de falhas em sensores; ou, ainda mais importante, (iii) em muitas aplicações tais padrões representam comportamentos extraordinários que merecem algum tipo de atenção especial. Eles podem revelar, por exemplo, genes associados a determinadas doenças, fraudes em sistemas financeiros, funcionários/clientes com perfis não usuais de produtividade/consumo, etc. Nesse contexto, diz-se que "*one person's noise is another person's signal*" (KNORR; NG, 1998; HAN; KAMBER, 2006).

Existem várias definições de *outlier* na literatura, cada uma sendo mais (ou menos) apropriada dependendo do cenário de aplicação. Entre as definições mais amplamente citadas estão:

> "*An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs*" (GRUBBS, 1969).

> "*An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*" (HAWKINS, 1980).

> "*An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*" (BARNETT; LEWIS, 1994).

Devido ao fato de que não existe apenas uma única definição de *outlier*, dezenas de técnicas têm sido desenvolvidas, cada uma utilizando uma intuição própria do que deve ser considerado um *outlier* ou não. Tais técnicas podem ser categorizadas de diferentes maneiras. Por exemplo, uma distinção usual se dá entre as técnicas globais e locais (RAMASWAMY; RASTOGI; SHIM, 2000; BREUNIG *et al.*, 2000), que se refere ao escopo dos dados a ser considerado quando um método decide se uma determinada observação deve ser interpretada como um *outlier* e com qual grau ou probabilidade. Outra distinção se dá entre as técnicas de rotulação (*labeling*) e pontuação (*scoring* ou *ranking*), que respectivamente indicam se um método irá categorizar binariamente as observações como *outliers* ou não *outliers* (*inliers*) ou se, ao invés, irá pontuar e/ou ordenar as observações de acordo com o grau ou probabilidade com que cada uma delas deve ser categorizada como tal. Outra importante distinção se dá entre as técnicas voltadas para detecção em subespaços ou no espaço completo dos atributos dos dados (KRIEGEL *et al.*, 2009b; ZIMEK; SCHUBERT; KRIEGEL, 2012), que respectivamente indica se a técnica utilizará apenas um subconjunto dos atributos para rotular/pontuar uma determinada observação, ainda que o subconjunto de atributos possa diferir de observação para observação, ou se utilizará todo o conjunto de atributos para isso. Uma quarta distinção está entre

as técnicas supervisionadas, semissupervisionadas e não supervisionadas de detecção (TAN; STEINBACH; KUMAR, 2006). As técnicas supervisionadas pressupõem que se dispõe a priori de um conjunto suficiente e apropriado de observações previamente sabidas serem *inliers* e também um conjunto de observações previamente sabidas serem *outliers* de fato. Isso permite o treinamento de classificadores, caso um conjunto apropriado de atributos que descreva as observações esteja disponível. Devido aos *outliers* tenderem a ser naturalmente raros, em alguns cenários se dispõe apenas de um conjunto suficiente e apropriado de *inliers*; neste cenário, técnicas semissupervisionadas podem ser utilizadas para construir um modelo a partir dos *inliers* e as observações que não se encaixarem neste modelo são classificadas como *outlier*. Quando não se tem qualquer informação a priori sobre quais observações são de fato *outliers* ou *inliers*, faz-se necessário utilizar técnicas não supervisionadas, que não pressupõem qualquer conhecimento prévio sobre observações que de fato são ou não *outliers*. Este trabalho enfoca principalmente os métodos não supervisionados, mas a adaptação de tais métodos para operarem no contexto semissupervisionado constitui uma das hipóteses de pesquisa que esse trabalho pretende investigar.

O campo da estatística surgiu inicialmente como a principal área de pesquisa em detecção não supervisionada de *outliers* (HAWKINS, 1980; BARNETT; LEWIS, 1994). Entretanto, os respectivos métodos possuem ao menos uma das seguintes limitações (KNORR; NG, 1998): i) em sua maioria, os testes estatísticos para detecção de *outliers* se limitam a dados numéricos unidimensionais ou descritos por um número pequeno de atributos. Esta limitação os torna inadequados para a maioria das aplicações reais em bases de dados heterogêneas e multidimensionais; e ii) em sua maioria os testes são paramétricos, ou seja, são baseados na hipótese que os dados seguem uma determinada distribuição (e.g. Normal) cujos parâmetros (e.g. média e variância) podem ou não ser conhecidos, dependendo do teste. Em razão dessas limitações, métodos alternativos de detecção de *outliers* passaram a ser desenvolvidos que i) não se baseiam em hipóteses específicas sobre a distribuição dos dados (métodos não paramétricos) e ii) são computacionalmente viáveis para aplicação em bases de dados de forma mais ampla, possivelmente de grande porte e multidimensionais (métodos *database-oriented*). Esta vertente ganhou impulso a partir do trabalho pioneiro de Knorr e Ng (1998), Knorr, Ng e Tucanov (2000), em que se propõem as primeiras técnicas não paramétricas computacionalmente escaláveis de detecção de *outliers* em grandes bases de dados, que ao mesmo tempo generalizam e unificam abordagens estatísticas paramétricas quando certas hipóteses são satisfeitas. Várias novas abordagens surgiram a partir desses trabalhos (ORAIR *et al.*, 2010), tais como os algoritmos KNNOutlier (RAMASWAMY; RASTOGI; SHIM, 2000), WKNN ou HilOut (ANGIULLI; PIZZUTI, 2002; ANGIULLI; PIZZUTI, 2005), OPTICS-OF (BREUNIG *et al.*, 1999), LOF (BREUNIG *et al.*, 2000), LDOF (ZHANG; HUTTER; JIN, 2009), COF (TANG *et al.*, 2002), INFLO (JIN *et al.*, 2006), LoOP (KRIEGEL *et al.*, 2009a), ABOD (KRIEGEL; SCHUBERT; ZIMEK, 2008), LOCI (PAPADIMITRIOU *et al.*, 2003), dentre outros.

Devido à definição vaga de *outlier*, cada algoritmo utiliza um critério próprio para julgar

de forma quantitativa o nível de aderência de cada observação com o conceito de *outlier*, que é naturalmente subjetivo no contexto não supervisionado. Isso dificulta sensivelmente a escolha de um algoritmo em particular e também a escolha de uma configuração adequada para o algoritmo escolhido em uma dada aplicação prática. Isso também torna altamente complexo avaliar a qualidade da solução obtida por um algoritmo/configuração em particular adotados pelo analista, especialmente em função da problemática de se definir uma medida de qualidade que não seja vinculada ao próprio critério utilizado pelo algoritmo. Tais questões estão inter-relacionadas e se referem respectivamente aos problemas de seleção de modelos e avaliação (ou validação) de resultados em aprendizado não supervisionado. Esses problemas têm sido investigados ao longo de décadas na área de agrupamento não supervisionado de dados (*data clustering*) (JAIN; DUBES, 1988; GAN; MA; WU, 2007), mas são raramente mencionados e estavam praticamente intocados na área de detecção de *outliers* até a publicação do índice IREOS (MARQUES, 2015; MARQUES *et al.*, 2015).

As áreas de agrupamento de dados e detecção de *outliers* estão intimamente relacionadas. De fato, ao se referir a um *outlier* como *"an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"*, assume-se implicitamente a existência de um ou mais mecanismos responsáveis por gerar as observações consideradas não suspeitas. Intuitivamente, grupos de dados (*clusters*) são candidatos naturais para modelar tais mecanismos. Além disso, um *outlier* não é necessariamente sempre considerado como uma observação que desvia muito de todas as demais. De fato, em alguns domínios de aplicação *outliers* são vistos como observações que se assemelham umas às outras por serem possivelmente geradas por um mecanismo comum, embora desviem e sejam geralmente menos frequentes do que as demais observações (por exemplo, certos tipos de fraude ou mutações genéticas seguindo um padrão comum).

Dada a relação supramencionada entre as áreas de agrupamento de dados e detecção não supervisionada de *outliers*, é surpreendente que os problemas de avaliação e seleção de modelos já tenham sido extensivamente investigados na primeira (embora ainda sejam temas desafiadores e atuais de pesquisa) porém ainda estejam tão em aberto nesta última. De fato, o procedimento de avaliação de algoritmos de detecção de *outliers* usualmente adotado na literatura é baseado na utilização de bases de dados previamente rotuladas, nas quais os *outliers* (segundo alguma intuição particular) são previamente conhecidos. No contexto não supervisionado, os rótulos não são utilizados pelo algoritmo em si, mas para avaliar os resultados produzidos por este. Especificamente, a categorização binária ou o *ranking* das observações produzidos pelo algoritmo são comparados com a rotulação considerada correta e previamente conhecida (*ground truth*). O resultado quantitativo desta comparação, por exemplo utilizando *precision-at-n* ou área sob a curva ROC (SCHUBERT *et al.*, 2012), é tomado como medida de qualidade do algoritmo de detecção em questão. Entretanto, a disponibilidade de uma base de dados pré-rotulada é obviamente inconsistente com a hipótese fundamental do aprendizado não supervisionado. Logo, embora o procedimento de avaliação descrito acima seja útil durante o desenvolvimento de novos

algoritmos, ele não é viável em aplicações práticas no contexto não supervisionado.

No contexto de agrupamento de dados, os problemas de avaliação e seleção de modelos quando não se tem qualquer informação sobre quais são de fato os rótulos das observações é usualmente abordado utilizando os chamados índices de validação internos (JAIN; DUBES, 1988). Tais índices são denominados internos pois não fazem uso de nenhuma informação externa (como rótulos) na avaliação. A avaliação é feita baseada apenas na informação dos dados e das soluções a serem avaliadas. A maioria desses índices são também relativos no sentido que eles podem ser utilizados para comparar diferentes soluções e apontar qual delas é melhor em termos relativos. Tais índices têm se mostrado ferramentas efetivas e úteis para avaliação e seleção de modelos em agrupamentos de dados (MILLIGAN; COOPER, 1985; VENDRAMIN; CAMPELLO; HRUSCHKA, 2010).

No contexto de detecção não supervisionada de *outliers*, um índice interno e relativo pioneiro para avaliação de soluções binárias de detecção, nomeado IREOS, foi proposto por Marques *et al.* (2015). O índice, baseado no critério de separabilidade dado por um classificador de máxima margem (e.g. SVM e KLR (HASTIE; TIBSHIRANI; FRIEDMAN, 2013)), pode avaliar e comparar diferentes soluções (top-n, i.e., rotulações binárias) candidatas baseando-se apenas nas informações dos dados e nas próprias soluções a serem avaliadas, e consequentemente, permite selecionar soluções mais promissoras, que correspondem a modelos (algoritmos, parâmetros) mais adequados. Porém IREOS aborda apenas uma pequena parte do problema, conhecido como problema top-n de detecção de *outliers* (ANGIULLI; FASSETTI, 2009; ANGIULLI; PIZZUTI, 2005; BAY; SCHWABACHER, 2003; GHOTING; PARTHASARATHY; OTEY, 2008; JIN; TUNG; HAN, 2001), em que o método de detecção rotula um subconjunto de n observações como *outliers* e o restante das observações são rotuladas como *inliers*. A maioria dos algoritmos de detecção de *outliers*, entretanto, não rotulam as observações como *outlier* ou *inlier*, mas pontuam e/ou ordenam as observações de acordo com seu grau ou probabilidade. Pelo fato do índice não conseguir avaliar internamente *scorings/rankings*, ou então, comparar soluções com diferentes n, de alguma forma um mesmo n deve ser determinado para todas as soluções a serem comparadas, o que por si só é uma tarefa desafiadora.

Outra limitação do índice IREOS é com relação à avaliação de soluções de *outliers* detectados em subespaços dos dados. Devido ao subconjunto de atributos relevantes poder diferir de observação para observação, essas soluções não podem ser diretamente comparadas pelo índice nos seus diferentes subespaços, possivelmente, de diferentes dimensionalidades. Um importante tópico de pesquisa que se pretende investigar neste trabalho é como IREOS poderia ser adaptado e/ou estendido para avaliar soluções dadas por algoritmos que produzem *scorings/rankings* e por algoritmos de detecção em subespaços dos dados.

A extensão e/ou adaptação do índice é importante, pois em aplicações práticas de métodos de detecção de *outliers*, usuários se beneficiariam de uma estimativa de qualidade não supervisionada fornecida pelo índice. Entretanto, o potencial de aplicações para os índices

internos são muito mais diversas, por exemplo: (i) a falta de índices internos em detecção de *outliers* tem sido notada como uma lacuna na literatura com relação ao desenvolvimento de métodos avançados para seleção de *ensembles* (ZIMEK; CAMPELLO; SANDER, 2013); (ii) procedimentos análogos aos utilizados em agrupamento de dados para automaticamente determinar o número de grupos (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010) poderiam ser adotados para automaticamente determinar o número de *outliers*; e (iii) os critérios em que tais índices são baseados poderiam levar a novos paradigmas de algoritmos de detecção de *outliers*. Os tópicos acima mencionados são interessantes questões de pesquisa que também se pretende investigar neste trabalho.

1.1 Hipóteses de Pesquisa

Este trabalho se propõe a investigar as seguintes hipóteses de pesquisa:

- **Hipótese I:** Utilizando o *framework* proposto por Janssens, Flesch e Postma (2009) para adaptação de métodos não supervisionados de detecção de *outliers* para o contexto semissupervisionado, outros métodos não supervisionados podem também ser adaptados e um estudo comparativo mais completo e rigoroso do que aquele descrito na referência acima pode ser realizado. Acredita-se que resultados mais consistentes e conclusões mais confiáveis possam ser obtidas deste estudo;
- **Hipótese II:** Avaliações de *rankings/scorings* de soluções de detecção de *outliers* podem ser diretamente realizadas sem a necessidade da binarização da solução em um problema top-n;
- **Hipótese III:** Avaliações de soluções de detecção de *outliers* em diferentes subespaços dos dados podem ser comparadas tornando a separabilidade das observações em seus diferentes subespaços comensuráveis utilizando procedimentos similares aos utilizados por algoritmos de detecção em subespaços (KRIEGEL *et al.*, 2009b) e *ensembles* (LAZAREVIC; KUMAR, 2005) para garantir comensurabilidade de seus *scorings*;
- **Hipótese IV:** Procedimentos similares aos utilizados em agrupamento de dados para determinar automaticamente o números de grupos através da aplicação de índices internos de validação (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010) podem ser adaptados permitindo a utilização do *Internal, Relative Evaluation of Outlier Solutions* (IREOS) para determinar automaticamente o número de *outliers* (n) presente nos dados;
- **Hipótese V:** Soluções com diferentes números de observações rotuladas como *outliers* podem ser comparadas por meio de testes estatísticos que assumam variância e tamanho da amostra diferentes (WELCH, 1947; MANN; WHITNEY, 1947), dado que a distribuição das avaliações das possíveis soluções segue uma distribuição Normal se certas suposições forem respeitadas.

- **Hipótese VI:** O conceito de separabilidade utilizado por IREOS para avaliação das soluções pode ser adaptado com vistas ao desenvolvimento de um novo paradigma de detecção não supervisionada de *outliers*, particularmente no cenário de subespaços, utilizando técnicas de seleção de atributos, tais como *forward selection* e *backward elimination* (GUYON; ELISSEEFF, 2003), similarmente à abordagem utilizada por Micenkova *et al.* (2013) para pós-processamento da solução de um algoritmo de detecção de *outliers*.

- **Hipótese VII:** IREOS pode ser utilizado para avaliar, segundo a ótica da separabilidade na qual o índice se baseia, rótulos externos de bases de dados de detecção de *outliers*, que em sua maioria são oriundas de outras áreas (CAMPOS *et al.*, 2016; GOLDSTEIN; UCHIDA, 2016), para verificar se tais rotulações externas são coerentes com a disposição espacial do dados no espaço de atributos em que esses estão descritos.

1.2 Objetivos

Este trabalho tem como seu objetivo principal a extensão e aplicação do índice IREOS para avaliação interna, seleção de modelo e detecção não supervisionada de *outliers* em espaços e subespaços dos dados, de forma a validar as hipóteses de pesquisa formuladas na seção anterior. Pretende-se que as técnicas desenvolvidas sejam disponibilizadas publicamente utilizando implementações computacionais eficientes baseadas, por exemplo, em índices espaciais (GUTTMAN, 1984; BECKMANN *et al.*, 1990) e/ou *frameworks* de computação paralela (DEAN; GHEMAWAT, 2008).

1.3 Organização do Trabalho

Este trabalho está organizado da seguinte forma:

No Capítulo 2 são apresentados conceitos básicos sobre detecção de *outliers*. Introduz-se seis representantes das diversas abordagens não supervisionadas, assim como exemplos de algoritmos de cada categoria. Em seguida, são abordados os algoritmos que utilizam o paradigma semissupervisionado, sendo que três categorias de técnicas são apresentadas com exemplos de seus respectivos algoritmos. Por fim, uma breve visão geral dos algoritmos que utilizam o paradigma supervisionado é fornecida.

No Capítulo 3 as principais medidas para avaliação de soluções de detecção de *outliers* são apresentadas. Uma medida externa para avaliação de soluções top-n e outra para avaliação *rankings*, bem como a única medida interna conhecida na literatura (IREOS) são introduzidas.

No Capítulo 4 são discutidos os resultados preliminares e/ou direções a serem exploradas com relação às hipóteses de pesquisa formuladas na seção 1.1.

No Capítulo 5 são listadas as atividades a serem desenvolvidas até a conclusão prevista do doutorado e o cronograma para a realização destas atividades.

CAPÍTULO

2

DETECÇÃO DE *OUTLIERS*

Diferentes técnicas de detecção de *outliers* têm sido propostas no contexto de mineração de dados, tais técnicas podem utilizar os paradigmas de aprendizado supervisionado, semissupervisionado ou não supervisionado. Neste Capítulo são apresentados conceitos básicos sobre detecção de *outliers*. Existem diversas abordagens utilizadas pelos algoritmos de detecção não supervisionada de *outliers*; seis delas são introduzidas na seção 2.1, assim como exemplos de algoritmos que utilizam essas abordagens. Na seção 2.2 são abordados os algoritmos que utilizam o paradigma semissupervisionado; três categorias de técnicas são apresentadas com exemplos de seus respectivos algoritmos. Por fim, uma breve visão geral dos algoritmos que utilizam o paradigma supervisionado é fornecida na seção 2.3.

2.1 Detecção Não Supervisionada de *Outliers*

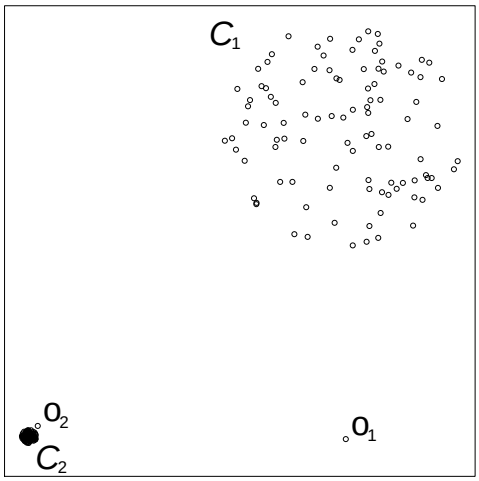
Quando não se tem qualquer conhecimento prévio sobre quais observações são de fato *outliers* ou não, faz-se necessário a utilização de técnicas de aprendizado não supervisionado. Cada técnica, entretanto, utiliza um critério próprio para julgar de forma quantitativa o nível de aderência de cada observação ao conceito de *outlier*, que é naturalmente subjetivo no contexto não supervisionado.

Cada uma dessas técnicas pode adotar diferentes abordagens para determinar se uma observação é um *outlier*. De forma geral, pode-se categorizar tais técnicas em: (i) técnicas de detecção global, que determinam se uma observação é *outlier* com respeito a toda base de dados, ou (ii) técnicas de detecção local, que determinam se uma observação é um *outlier* com respeito à região do espaço dos dados em que a observação está localizada.

A Figura 1 ajuda a ilustrar a diferença entre as técnicas de detecção local e global. A base de dados apresentada na Figura 1 é formada por dois mecanismos diferentes, o grupo C_1 e o grupo C_2, sendo que o grupo C_1 é bem mais esparso que o grupo C_2. Nesta base de

dados também estão presentes dois *outliers*, a observação o_1 considerada um *outlier* global e a observação o_2 considerada um *outlier* local. A ideia central dos algoritmos de detecção local é assumir que a base de dados pode ser gerada por mais de um mecanismo. Logo, no momento de detectar *outliers* nesta base, as técnicas de detecção local irão tratar de forma diferente as observações da vizinhança do grupo C_1 e as observações da vizinhança do grupo C_2. Desta forma, a observação o_2 localizada próxima ao grupo C_2 poderá ser considerada *outlier*, pois ela desvia significantemente em relação às demais observações da sua vizinhança. Entretanto, com a utilização de um algoritmo de detecção global a observação o_2 não seria detectada como *outlier*, pois a mesma desvia tanto quanto as observações do grupo C_1 em relação a toda a base de dados, de forma que, para a observação o_2 ser classificada como *outlier*, muita das observações do grupo C_1 também teriam que ser classificadas como *outliers*.

Figura 1 – Base de dados gerada por dois mecanismos diferentes (inspirado em (BREUNIG *et al.*, 2000))



Uma forma mais específica de se categorizar as diferentes técnicas de detecção não supervisionada de *outliers* é quanto ao paradigma utilizado. As diferentes técnicas existentes podem utilizar diferentes paradigmas para definir se uma observação é um *outlier*, tais como os paradigmas baseados em estatísticas (BARNETT; LEWIS, 1994; HAWKINS, 1980), distância (RAMASWAMY; RASTOGI; SHIM, 2000; ANGIULLI; PIZZUTI, 2002; ZHANG; HUTTER; JIN, 2009), densidade (BREUNIG *et al.*, 2000; JIN *et al.*, 2006), ângulo (KRIEGEL; SCHUBERT; ZIMEK, 2008; PHAM; PAGH, 2012), agrupamento (CAMPELLO *et al.*, 2015a; HE; XU; DENG, 2003), subespaços (KRIEGEL *et al.*, 2009b; ZHANG *et al.*, 2004), dentre outros.

2.1.1 Técnicas Estatísticas

O campo da estatística surgiu inicialmente como a principal área de pesquisa em detecção não supervisionada de *outliers*. Existem dezenas de testes estatísticos para detectar discordância/anomalia nos dados (HAWKINS, 1980; BARNETT; LEWIS, 1994; GRUBBS, 1969). Tais testes diferem na hipótese sobre o tipo de distribuição dos dados, na hipótese sobre o tipo dos dados propriamente ditos, na natureza dos *outliers* esperados, dentre outros. Eles possuem em

comum, entretanto, o fato de necessitarem fazer uma série de suposições. Por exemplo, o teste pode precisar fazer suposição quanto à distribuição dos dados, que geralmente não é conhecida a priori, sendo necessário realizar extensivos testes para encontrar uma distribuição conhecida que descreva bem os dados. Outra suposição comum é quanto aos parâmetros da distribuição (e.g. média e variância), porém os parâmetros da distribuição em geral não são conhecidos e precisam portanto ser estimados a partir dos próprios dados. Tal estimação é geralmente sensível e distorcida pela presença dos próprios *outliers* que se deseja detectar. Essas distorções geram efeitos indesejados conhecidos como *masking*, quando a presença de *outliers* é mascarada por outros *outliers*, e *swamping*, quando *inliers* são incorretamente rotulados como *outliers* devido à distorção causada pelos *outliers* (HADI, 1992). Além da maioria dos testes serem paramétricos, ou seja, assumirem que os dados seguem uma determinada distribuição (e.g. Normal) em que os parâmetros podem ou não ser conhecidos, a maioria deles pode ser aplicado apenas em base de dados numéricas e unidimensionais. Nesse caso, uma regra prática comum é tomar como *outliers* observações que desviam mais do que 3 vezes o desvio padrão a partir da média de uma distribuição Normal (unidimensional); no caso multidimensional limiares equivalentes podem ser estabelecidos e impostos à distância de Mahalanobis, como apresentado em maiores detalhes a seguir.

Distância de Mahalanobis

A distância de Mahalanobis (MAHALANOBIS, 1936) é uma generalização da distância Euclidiana que captura a correlação entre as variáveis. No contexto de detecção de *outliers*, a distância de Mahalanobis é utilizada para calcular a distância de cada observação $\mathbf{x}$ para o centro de seu grupo (μ) baseado na sua matriz de covariância (Σ) (Equação (2.1)). A utilização desta técnica implica em suposições quanto à distribuição dos dados e seus parâmetros. A técnica é aplicada a dados que seguem distribuição Normal quando se sabe seus parâmetros (μ e Σ). Variações da distância de Mahalanobis no contexto de detecção de *outliers* foram propostas para evitar os problemas de *masking* e *swamping* (ROUSSEEUW; ZOMEREN, 1990; HADI, 1992).

$$mahalanobis(\mathbf{x}) = \sqrt{(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)} \tag{2.1}$$

Como resultado o algoritmo retorna um *score* para cada observação. Para definir quais observações deverão ser rotuladas binariamente como *outliers* ou *inliers* comumente é utilizado um resultado estatístico bem conhecido: quando a distância de Mahalanobis é aplicada a dados que seguem uma distribuição Normal d-dimensional, as distâncias resultantes seguem uma distribuição χ^2 com d graus de liberdades (HAIR *et al.*, 2005). A partir da distribuição χ^2 é possível definir o limiar para rotulação das observações. Caso seja de interesse, por exemplo, a rotulação de 2,5% das observações que mais desviam da distribuição teórica, pode-se utilizar o quartil teórico de 0.975 da distribuição χ^2 para rotulação dos *outliers*.

2.1.2 Técnicas Baseadas em Distância

As técnicas baseadas em distância assumem que *outliers* são observações que estão distantes de suas observações mais próximas (tais observações são definidas aqui como *vizinhos mais próximos*). Para isso, é necessário definir uma distância ou medida de similaridade entre duas observações. Existem diferentes distâncias e medidas de similaridades. Enquanto para atributos contínuos a distância Euclidiana é a mais utilizada, para atributos categóricos é comumente utilizado o coeficiente de casamento simples. Dois algoritmos bem conhecidos que utilizam esta abordagem são o KNNOutlier e o HilOut, ambos apresentados a seguir.

KNNOutlier

O algoritmo KNNOutlier é um dos algoritmos baseados em distância existentes. Foi proposto por Ramaswamy, Rastogi e Shim (2000) e é um algoritmo de detecção de *outliers* globais. O KNNOutlier atribui um *score* para cada observação da base de dados para medir o quanto cada uma delas se caracteriza como *outlier*. Para calcular o *score* de uma determinada observação é utilizada simplesmente a distância desta observação ao seu k-ésimo vizinho mais próximo, definida aqui como $k\text{-}distance(\cdot)$ (Equação (2.2)).

$$KNNOutlier_k(\mathbf{x}) = k\text{-}distance(\mathbf{x}) \tag{2.2}$$

Uma variante do KNNOutlier é o HilOut, ou *Weighted k-Nearest Neighbors* (WKNN), que foi proposto por Angiulli e Pizzuti (2002). Diferente do KNNOutlier, que utiliza apenas a distância do k-ésimo vizinho da observação para produzir seu *score*, HilOut utiliza a soma das distâncias aos k vizinhos mais próximos da observação para produzir seu *score* (Equação (2.3)).

$$HilOut_k(\mathbf{x}) = \sum_{i=1}^{k} i\text{-}distance(\mathbf{x}) \tag{2.3}$$

Caso seja de interesse obter uma classificação binária de n observações como *outliers* e as demais como *inliers*, são rotuladas as n observações com maior *score* como *outliers* e as demais como *inliers*.

2.1.3 Técnicas Baseadas em Densidade

Para julgar quantitativamente se uma observação deve ser considerada *outlier*, as técnicas baseadas em densidade estimam a densidade de cada observação $\mathbf{x}$ com respeito à sua vizinhança $V_k(\mathbf{x})$, vizinhança aqui definida como o conjunto de observações cuja distância para a observação $\mathbf{x}$ não seja maior que a $k\text{-}distance(\mathbf{x})$; logo $|V_k(\mathbf{x})| \geq k$ (as observações pertencentes à vizinhança de $\mathbf{x}$ são ditos seus k vizinhos mais próximos, embora possam ser um número maior que k em caso de empates). Essas técnicas então assumem que observações que estejam em vizinhanças de baixa densidade são consideradas *outliers*, enquanto observações localizadas em regiões

densas são consideradas *inliers*. Para estimativa da densidade da vizinhança de uma determinada observação, várias técnicas utilizam as distâncias da observação aos seus k vizinhos mais próximos como base para uma estimativa do inverso da densidade daquela vizinhança. LOF é o algoritmo precursor desta abordagem, a partir dele outras variantes surgiram, como por exemplo o LOCI (PAPADIMITRIOU *et al.*, 2003). LOF e LOCI são apresentados em mais detalhes a seguir.

LOF

Uma das técnicas baseadas em densidade existentes é o *Local Outlier Factor* (LOF) que foi proposto por Breunig *et al.* (2000) e foi o algoritmo pioneiro entre os algoritmos de detecção de *outlier* locais. A partir deste algoritmo surgiram muitos outros (JIN *et al.*, 2006; KRIEGEL *et al.*, 2009a; PAPADIMITRIOU *et al.*, 2003; TANG *et al.*, 2002; ZHANG; HUTTER; JIN, 2009) que de alguma forma seguem a mesma intuição.

Assim como os algoritmos KNNOutlier e HilOut, LOF também retorna um *score* para cada observação. O *score* produzido é baseado na densidade da observação em relação à densidade dos seus k vizinhos mais próximos, quanto maior este *score* maior a evidência de que a observação seja um *outlier*.

Como estimativa de densidade local de uma observação, LOF utiliza o inverso da média das distâncias da observação aos seus k vizinhos mais próximos. Entretanto, para evitar efeitos similares aos de *masking* e *swamping*, Breunig *et al.* (2000) propõem utilizar a distância de alcançabilidade (*reachability distance*) em vez de utilizar simplesmente a distância. A distância de alcançabilidade de uma observação $\mathbf{x}$ em relação à observação $\mathbf{y}$ é definida como $\textit{reach-dist}_k(\mathbf{x},\mathbf{y}) = \text{MAX}\{k\textit{-distance}(\mathbf{y}), d(\mathbf{x},\mathbf{y})\}$, em que $d(\mathbf{x},\mathbf{y})$ é a distância da observação $\mathbf{x}$ para a observação $\mathbf{y}$. Repare que apesar de ser chamada de distância de alcançabilidade, essa medida não é uma distância em sua definição formal, uma vez que ela não é simétrica. A estimativa de densidade local *edl* de uma observação $\mathbf{x}$ é matematicamente apresentada na Equação (2.4).

$$\textit{edl}_k(\mathbf{x}) = \left(\frac{\sum_{\mathbf{y} \in V_k(\mathbf{x})} \textit{reach-dist}_k(\mathbf{x},\mathbf{y})}{|V_k(\mathbf{x})|} \right)^{-1} \tag{2.4}$$

Para cada observação, o *score* produzido pelo LOF é a média da razão entre a estimativa de densidade local dos seus k vizinhos mais próximos e a estimativa de densidade local da própria observação (Equação (2.5)). Sendo assim, se o *score* de determinada observação for igual a 1, significa que a observação possui em média a mesma densidade que seus k vizinhos mais próximos. Em caso do *score* for menor que 1, a observação possui em média uma densidade maior que seus k vizinhos mais próximos. Em ambos os casos a observação pode ser considerada *inlier*. Entretanto, um valor alto de *score* mostra que a observação está em uma região de baixa

densidade comparada aos seus k vizinhos mais próximos, assim quanto maior este *score* maior a evidência de que a observação seja um *outlier*.

$$LOF_k(\mathbf{x}) = \frac{\sum_{\mathbf{y} \in V_k(\mathbf{x})} \frac{edl_k(\mathbf{y})}{edl_k(\mathbf{x})}}{|V_k(\mathbf{x})|} \tag{2.5}$$

LOCI

A escolha do número de vizinhos mais próximos a ser considerado não é uma escolha trivial. Para tentar contornar tal problema, Papadimitriou *et al.* (2003) propuseram o algoritmo *Local Correlation Integral* (LOCI) que tenta lidar com este problema analisando a densidade de uma observação em múltiplas escalas de vizinhanças.

Para calcular a densidade de uma determinada observação $\mathbf{x}$, o algoritmo define um conjunto de raios relevantes R, que representa o conjunto de distâncias em que uma nova observação é incluída em sua vizinhança. LOCI analisa a densidade da observação nos múltiplos raios $r \in R$, definindo a vizinhança-r de $\mathbf{x}$, $V(\mathbf{x}, r)$, como o conjunto de observações contidas dentro do raio r de $\mathbf{x}$, i.e., $V(\mathbf{x}, r) = \{\mathbf{y} | d(\mathbf{x}, \mathbf{y}) \leq r\}$.

O LOCI de uma observação $\mathbf{x}$ é calculado com relação a um parâmetro α entre $(0, 1]$. O parâmetro α é utilizado para definir o número de vizinhos contidos na vizinhança-αr de $\mathbf{x}$, $|V(\mathbf{x}, \alpha r)|$, que funciona como uma estimativa de densidade e será utilizada na comparação feita pelo *Multi-Granularity Deviation Factor* (MDEF). Para o cálculo do MDEF é necessário calcular a média do número de vizinhos da vizinhança-αr entre os vizinhos contidos na vizinhança-r de $\mathbf{x}$ (Equação (2.6)).

$$\hat{\rho}(\mathbf{x}, r, \alpha) = \frac{\sum_{\mathbf{y} \in V(\mathbf{x}, r)} |V(\mathbf{y}, \alpha r)|}{|V(\mathbf{x}, r)|} \tag{2.6}$$

O $MDEF(\mathbf{x}, r, \alpha)$ é calculado como a diferença entre $\hat{\rho}(\mathbf{x}, r, \alpha)$ e $|V(\mathbf{x}, \alpha r)|$ normalizado por $\hat{\rho}(\mathbf{x}, r, \alpha)$, como apresentado na Equação (2.7).

$$MDEF(\mathbf{x}, r, \alpha) = \frac{\hat{\rho}(\mathbf{x}, r, \alpha) - |V(\mathbf{x}, \alpha r)|}{\hat{\rho}(\mathbf{x}, r, \alpha)} = 1 - \frac{|V(\mathbf{x}, \alpha r)|}{\hat{\rho}(\mathbf{x}, r, \alpha)} \tag{2.7}$$

Para a rotulação da observação como *outlier* é então proposto um MDEF normalizado (Equação (2.8)), que consiste na razão entre o desvio padrão do número de vizinhos da vizinhança-αr entre os vizinhos contidos na vizinhança-r de $\mathbf{x}$, e $\hat{\rho}(\mathbf{x}, r, \alpha)$.

$$\sigma_{MDEF(\mathbf{x}, r, \alpha)} = \frac{\sigma_{\rho(\mathbf{x}, r, \alpha)}}{\hat{\rho}(\mathbf{x}, r, \alpha)} \tag{2.8}$$

O LOCI de uma observação é então definido como o valor máximo da razão entre o MDEF da observação e o seu MDEF normalizado entre todos os raios r contidos no conjunto de raios relevantes R, conforme Equação (2.9).

$$LOCI_{\alpha}(\mathbf{x}) = \underset{r \in R}{\text{MAX}} \left\{ \frac{MDEF(\mathbf{x}, r, \alpha)}{\sigma_{MDEF(\mathbf{x},r,\alpha)}} \right\} \tag{2.9}$$

2.1.4 Técnicas Baseadas em Agrupamento

Técnicas baseadas em agrupamento utilizam algoritmos de agrupamentos de dados para detectar *outliers*. A intuição por trás dessas técnicas assemelha-se com as das técnicas estatísticas paramétricas, em que observações que não se encaixem bem nas distribuições dos dados, aqui representadas pelos grupos dos dados, são consideradas *outliers*. Vários algoritmos de agrupamento de dados utilizam um limiar de densidade para agrupar as observações (ESTER *et al.*, 1996; HINNEBURG; HINNEBURG; KEIM, 1998) e observações que não atinjam o limiar de densidade necessário, são rotuladas como *outliers* como subproduto do agrupamento. Uma técnica recente, GLOSH, baseada na hierarquia construída pelo algoritmo de agrupamento HDBSCAN*(CAMPELLO *et al.*, 2015a) é apresentada a seguir.

GLOSH

Global-Local Outlier Scores from Hierarchies (GLOSH) é um algoritmo não supervisionado de detecção de *outliers* baseado na estimativa hierárquica de densidade realizada pelo algoritmo de agrupamento hierárquico HDBSCAN*. Após construir a hierarquia de grupos para toda base de dados, o GLOSH de uma observação $\mathbf{x}$ pode ser computado baseado na diferença entre a densidade da observação $\mathbf{x}$ e a maior densidade no grupo mais próximo de $\mathbf{x}$ na hierarquia do HDBSCAN*, conforme Equação (2.10).

$$GLOSH(\mathbf{x}) = \frac{\lambda_{\text{MAX}}(C_{\mathbf{x}}) - \lambda(\mathbf{x})}{\lambda_{\text{MAX}}(C_{\mathbf{x}})} \tag{2.10}$$

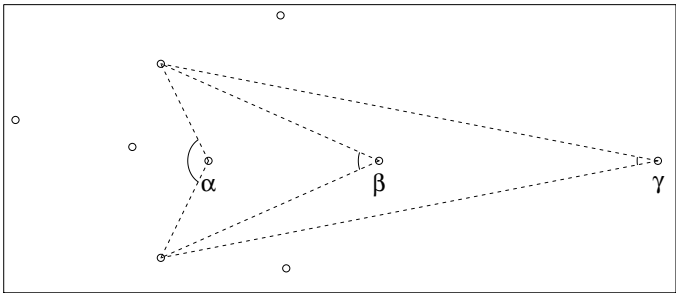
onde $\lambda(\mathbf{x})$ é a densidade de $\mathbf{x}$ e $\lambda_{\text{MAX}}(C_{\mathbf{x}})$ é a densidade da observação de maior densidade pertencente ao "grupo mais próximo" de $\mathbf{x}$ (sob uma perspectiva de conectividade na hierarquia baseada em densidade). As densidades são estimadas pelo HDBSCAN* utilizando um estimador de densidade de k vizinhos mais próximos.

Note que observações pertencentes a grupos que sigam, por exemplo, uma distribuição uniforme, tendem a ter densidades ao menos próximas a densidade da observação mais densa do grupo, e então, o GLOSH das observações serão próximos de 0, sugerindo serem *inliers*. Por outro lado, se o grupo seguir, por exemplo, uma distribuição Gaussiana, será esperado que os valores de GLOSH sejam próximos de 0 para as observações na região do pico da distribuição, e tenderá a 1 conforme as observações se distanciem desta região.

2.1.5 Técnicas Baseadas em Ângulo

As técnicas baseadas em ângulo representam cada observação da base de dados como um ponto no espaço euclidiano, o que restringe estes métodos apenas a bases de dados com atributos numéricos. Para determinar o quanto cada ponto se caracteriza como um *outlier*, elas avaliam a variância do ângulo que cada ponto forma com os demais pares de pontos. Essas técnicas assumem que a maioria dos pontos estarão concentrados em alguma direção a partir de um *outlier*, sendo assim espera-se *outliers* não tenham grande variação de ângulo, como por exemplo o ponto γ da Figura 2. Para um *inlier*, espera-se que os demais pontos estejam distribuídos em todas as direções, tendo grande variação de ângulo, como por exemplo o ponto α da Figura 2. Logo, espera-se que pontos que tenham baixa variação de ângulo sejam *outliers*. Uma grande vantagem desta abordagem em relação às demais é o fato de não utilizar uma medida baseada em distância, uma vez que as distâncias tendem a se deteriorarem com o aumento dimensionalidade. As técnicas que utilizam esta abordagem são mais robustas quando aplicadas a base de dados de alta dimensionalidade.

Figura 2 – Ângulos formados por pontos da base (inspirado em (KRIEGEL; SCHUBERT; ZIMEK, 2008))



ABOD

O algoritmo *Angle-Based Outlier Detection* (ABOD) foi proposto por Kriegel, Schubert e Zimek (2008). A ideia fundamental deste algoritmo está em utilizar também a variância do ângulo formado por um ponto com os demais pares de pontos, em vez de utilizar somente a distância desse ponto para os demais pontos da base de dados. Para computar o *score* de um ponto $\overrightarrow{A}$, $ABOF(\overrightarrow{A})$, ABOD computa a variância ao longo dos ângulos entre o ponto $\overrightarrow{A}$ e todos os demais pontos na base de dados $\mathbf{X}$, sendo os ângulos ponderados pelo inverso das distâncias dos respectivos pontos (Equação (2.11)). Devido à ponderação, observações mais distantes terão menos peso no cálculo do *score*, essa ponderação é importante pelo fato do ângulo entre pares de pontos mais distantes variarem naturalmente mais.

$$ABOF(\overrightarrow{A}) = VAR_{\overrightarrow{B},\overrightarrow{C}\in\mathbf{X}}\left(\frac{\overline{AB}\cdot\overline{AC}}{\|\overline{AB}\|^2\|\overline{AC}\|^2}\right) \tag{2.11}$$

Diferente das outras técnicas apresentadas, é esperado que *outliers* tenham valores baixos de *scorings*, uma vez que é esperada baixa variância de ângulo para os pontos considerados

outliers.

2.1.6 Técnicas de Detecção em Subespaços de Dados

A ideia de definir *outliers* com respeito a um subespaço do espaço original dos dados surgiu em Knorr e Ng (1999). Com o agravamento do problema da alta dimensionalidade das bases de dados encontradas na prática, tais técnicas têm recebido cada vez mais atenção, pois a utilização de muitos atributos irrelevantes pode fazer com que os *outliers* sejam facilmente mascarados, ainda que algumas técnicas de detecção de *outliers* tentem lidar com problema da alta dimensionalidade, como por exemplo as técnicas baseadas em ângulos. Como mostrado na Figura 3, que representa as projeções dos diferentes atributos de uma base de dados, a observação rotulada em vermelho se caracteriza como *outlier* em apenas uma das projeções (Figura 3a), enquanto nas três demais projeções a mesma observação se caracteriza como *inlier*. Se considerado cada atributo independentemente, também não seria possível detectá-la como *outlier*. Uma vez que os atributos mais relevantes são revelados, tais observações tornam-se mais detectáveis. Entretanto, a descoberta destes atributos não é uma tarefa trivial. Neste contexto, diferentes técnicas para detecção de *outliers* em subespaços têm sido propostas. Essas técnicas podem diferir na abordagem de exploração dos subespaços candidatos e/ou na forma em que os *scorings* de *outliers* são produzidos. A seguir são apresentadas duas técnicas de detecção de *outliers* em subespaços, SOD e HOS-Miner, que diferem na abordagem para exploração dos subespaços dos dados.

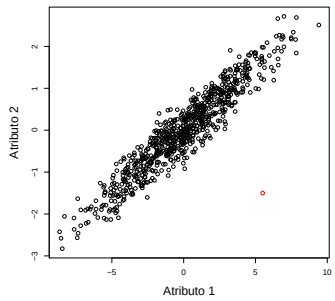
SOD

Subspace Outlier Degree (SOD) é um algoritmo de detecção de *outliers* em subespaços proposto por Kriegel *et al.* (2009b). Para medir o quanto uma observação $\mathbf{x}$ se caracteriza como *outlier* é definida uma vizinhança, chamada pelos autores de conjunto de referência $\mathbf{S}_k(\mathbf{x})$. Entretanto, a definição do conjunto de referência não é uma tarefa trivial no espaço completo de alta dimensionalidade dos dados. Para definir o conjunto de referência, ao invés de utilizarem os vizinhos mais próximos, os autores utilizam o conceito de vizinhos mais próximos compartilhados por serem mais relevantes em dimensões elevadas (HOULE *et al.*, 2010).

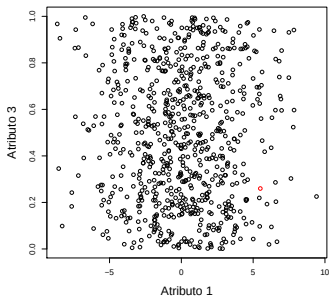
A partir do conjunto de referência $\mathbf{S}_k(\mathbf{x})$ é definido o conjunto dos atributos relevantes $\mathbf{A}(\mathbf{x})$, que consiste no conjunto de atributos em que as observações do conjunto de referência possuem baixa variância. Para definir se a variância de um atributo é baixa, o algoritmo define um parâmetro α entre $(0, 1]$, e considera a variância de um atributo baixa caso ela seja $\frac{\alpha}{d}$ vezes menor que a variância do espaço completo dos dados, sendo d o tamanho do espaço completo dos dados.

Para computar o SOD de uma observação $\mathbf{x}$ é computada a distância, no subespaço de atributos relevantes $\mathbf{A}(\mathbf{x})$, entre a observação $\mathbf{x}$ e a média do conjunto de referência $\mathbf{S}_k(\mathbf{x})$, deno-

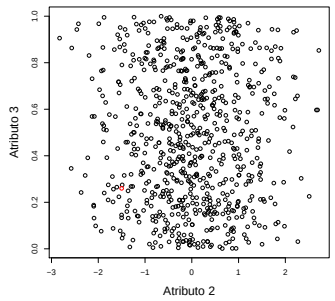
Figura 3 – Projeção dos diferentes atributos da base de dados, mesma observação nas diferentes projeções em vermelho



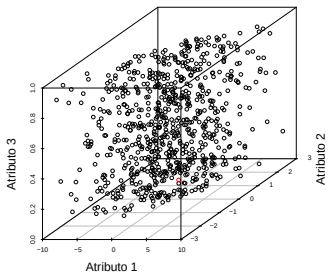
(a) Projeção dos atributos 1 e 2, observação em vermelho se caracteriza como *outlier*



(b) Projeção dos atributos 1 e 3, observação em vermelho não se caracteriza como *outlier*



(c) Projeção dos atributos 2 e 3, observação em vermelho não se caracteriza como *outlier*



(d) Projeção de todos atributos, observação em vermelho não se caracteriza como *outlier*

tada por $dist\left(\mathbf{x}^{\mathbf{A}(\mathbf{x})}, \mu_{\mathbf{S}_k(\mathbf{x})}^{\mathbf{A}(\mathbf{x})}\right)$, normalizado pelo tamanho do subespaço dos atributos relevantes dos dados $\mathbf{A}(\mathbf{x})$, conforme Equação (2.12).

$$SOD_{k,\alpha}(\mathbf{x}) = \frac{dist\left(\mathbf{x}^{\mathbf{A}(\mathbf{x})}, \mu_{\mathbf{S}_k(\mathbf{x})}^{\mathbf{A}(\mathbf{x})}\right)}{|\mathbf{A}(\mathbf{x})|} \tag{2.12}$$

HOS-Miner

O algoritmo *High-dimensional Outlying Subspaces* (HOS-Miner) (ZHANG *et al.*, 2004) difere da abordagem tradicional dos algoritmos de detecção de *outliers* em subespaços, no sentido que ao invés de procurar pelos subespaços relevantes dos dados para então procurar por *outliers*, HOS-Miner procura pelo subespaço em que cada observação mais se caracteriza como *outlier*. Nesta abordagem, o subespaço é determinado a partir da observação, ao invés da abordagem tradicional de primeiro definir o subespaço de atributos relevantes para então avaliar a observação.

HOS-Miner rotula uma observação como *outlier* se a soma das distâncias dos seus k vizinhos mais próximos, em algum subespaço dos dados, seja maior ou igual a ε. Entretanto,

nesta abordagem as distâncias não são normalizadas pelo número de dimensões do subespaço, logo conforme o número de dimensões aumenta, se torna mais provável que a observação seja considerada *outlier* de acordo com o algoritmo.

A procura pelos subespaços dos dados é realizada utilizando uma busca ascendente/-descedente do tipo *Apriori* (AGRAWAL; SRIKANT, 1994), que pode podar o espaço de busca a partir de duas propriedades: i) uma observação não será considerada *outlier* em qualquer subespaço de um subespaço em a observação não é considerada *outlier*; ii) uma observação será considerada *outlier* em todo superespaço de um subespaço em que a observação já é considerada *outlier*.

2.2 Detecção Semissupervisionada de *Outliers*

Na seção anterior foi abordado o problema de detecção não supervisionada de *outliers*, utilizado quando não se tem qualquer conhecimento prévio sobre quais observações são de fato *outliers* ou *inliers*. Em cenários de aplicação em que se dispõe de um conjunto de treinamento, mas que devido aos *outliers* serem normalmente raros não existam observações rotuladas como tais, ou em casos que existam observações rotuladas como *outliers* mas que não sejam suficientes para discriminar entre as classes *outlier* e *inlier* utilizando uma técnica supervisionada, pode-se utilizar técnicas de detecção semissupervisionada de *outliers*. A diferença para o cenário não supervisionado é a informação da classe *inlier* disponível no conjunto de treinamento, logo essa informação deve ser utilizada para se criar o modelo a partir do qual novas observações serão rotuladas de acordo como se encaixem (ou não) no modelo. Diferente do cenário não supervisionado, neste cenário pode-se estimar a função densidade de probabilidade (fdp) ou qualquer outro modelo sem se preocupar com a presença de possíveis *outliers* na base de dados. Este problema também é estudado pela área de classificação de padrões e conhecido como problema de aprendizado de classe única (*One Class Classification* (OCC)) (TAX, 2001b; KHAN; MADDEN, 2010). As técnicas semissupervisionadas podem ser categorizadas em técnicas baseadas em estimativa de densidade, fronteira e reconstrução.

2.2.1 Técnicas Baseadas em Estimativa de Densidade

Métodos baseados em estimativa de densidade utilizam os dados do conjunto de treinamento para ajustar os parâmetros de alguma fdp, e em seguida, novas observações podem ser classificadas usando esta fdp. Uma vez que a fdp é estimada utilizando apenas os *inliers*, não existe qualquer risco de *outliers* distorcerem a distribuição dos dados. Um problema, porém, é a necessidade de uma amostra suficientemente grande de *inliers* para produzir uma boa estimativa de densidade. Dependendo, por exemplo, do número de dimensões do problema, o número de observações necessárias para representar de forma suficiente a distribuição pode tornar-se muito grande e, em alguns cenários de aplicação real, muito custoso para se obter. Métodos baseados

em estimativa de densidade comumente utilizados em OCC são Gaussian e *Parzen Window*, apresentados a seguir.

Gaussian

O classificador Gaussian (TAX, 2001a) é um método muito simples que impõe um modelo de densidade estritamente unimodal e convexo sobre os dados, tentando modelar os dados utilizando uma única fdp Gaussiana (Equação (2.13)).

$$p_{Gauss}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)} \tag{2.13}$$

onde μ é a média, Σ a matriz de covariância e d a dimensionalidade dos dados. Uma nova observação é classificada computando sua probabilidade de pertencer à distribuição estimada.

Parzen Window

Parzen Windows (PW) é baseado no estimador de densidade de Parzen (PARZEN, 1962) que estima a densidade dos dados utilizando uma mistura de *kernels* centrados em cada uma das observações do conjunto de treinamento. *Kernels* Gaussianos podem ser utilizados com a diagonal da matriz de covariância $\Sigma_i = \Lambda I$, onde Λ é o parâmetro que ajusta a largura da Gaussiana, que pode ser otimizado utilizando o método da máxima verossimilhança (CAM, 1990). A probabilidade de uma observação pertencer à classe *inlier* é dada por:

$$PW(\mathbf{x}) = \frac{1}{N}\sum_i p_{Gauss}(\mathbf{x}|\mathbf{x}_i,\Lambda I) \tag{2.14}$$

Diferente de outros métodos OCC baseados em estimativa de densidade, PW é não paramétrico, logo, classificar novas observações pode ser relativamente custoso.

2.2.2 Técnicas Baseadas em Fronteira

Métodos baseados em fronteira não necessitam de um grande número de observações pelo fato desses métodos, ao invés de tentarem estimar a distribuição dos dados, apenas se preocuparem em definir uma fronteira de separação que limita a classe de interesse. Os métodos dessa categoria utilizam a fronteira construída pelo classificador para rotular novas observações, de tal forma que, caso a nova observação esteja dentro da fronteira de decisão, ela é classificada como *inlier*, caso contrário, a observação é classificada como *outlier*. Uma vez que o interesse é apenas em definir a fronteira, não é necessário obter um grande número de observações para representar plenamente a classe *inlier*. A seguir são apresentados alguns dos métodos dessa categoria: SVDD, LP e kNNDD.

SVDD

Support Vector Data Description (SVDD) (TAX; DUIN, 2004) é um método de OCC baseado em fronteira e inspirado no classificador *Support Vector Machines* (SVM) (VLADIMIR; VAPNIK, 1995) usado em problemas tradicionais de classificação. A principal diferença entre SVDD e SVM é que enquanto SVM tenta separar duas ou mais classes utilizando um hiperplano de máxima margem, SVDD tenta envolver a classe de interesse em uma hiperesfera de volume mínimo, minimizando o seguinte erro:

$$\mathscr{E}(R,\mathbf{a},\xi) = R^2 + C\sum_i \xi_i \tag{2.15}$$

s.a.

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall i \tag{2.16}$$

onde R é o raio da hiperesfera, $\mathbf{a}$ o centro da hiperesfera, ξ são variáveis de folga que permitem que observações da classe de interesse violem a fronteira do SVDD, e C é o parâmetro de penalidade/regularização.

Assim como o SVM tradicional, a formulação acima também pode ser estendida para problemas não lineares por meio da utilização de funções *kernels* (SCHOLKOPF; SMOLA, 2001).

LP

Linear Programming (LP) (PEKALSKA; TAX; DUIN, 2003) é um método baseado em fronteira que utiliza uma medida de dissimilaridade para comparar novas observações às observações do conjunto de treinamento (*inliers*). A medida de dissimilaridade pode ser computada diretamente a partir das observações (que podem ser grafos, imagens, textos e etc) sem necessariamente que os atributos sejam explicitados. Essas medidas devem cumprir uma série de critérios definidos pelos autores e podem ser específicas para o contexto de aplicação. Uma das medidas propostas é a função sigmóide. A intuição básica desse método é que se determinada observação pertence a uma classe, ela deveria ser similar às demais observações daquela classe. Sendo assim um limiar é imposto a uma função de proximidade, que determina o quão similar é a observação em relação à classe *inlier*, para então classificá-la.

kNNDD

k-Nearest Neighbour Data Description (kNNDD) (RIDDER; TAX; DUIN, 1998) é semelhante ao LOF e LOCI por aproximar a densidade local das observações do conjunto de treinamento, porém de uma forma mais simples. Uma observação é classificada calculando a

razão entre a distância da observação para o seu k-ésimo vizinho mais próximo $\mathrm{NN_k}(\mathbf{x})$, e a distância do k-ésimo vizinho mais próximo para o seu k-ésimo vizinho mais próximo:

$$\mathrm{kNNDD}_k(\mathbf{x}) = \frac{d(\mathbf{x}, \mathrm{NN_k}(\mathbf{x}))}{d(\mathrm{NN_k}(\mathbf{x}), \mathrm{NN_k}(\mathrm{NN_k}(\mathbf{x})))} \tag{2.17}$$

2.2.3 Técnicas Baseadas em Reconstrução

Por fim, pode-se utilizar os métodos baseados em reconstrução para modelar o conjunto de treinamento. Os métodos baseados em reconstrução escolhem um processo gerador para modelar os dados, afim de obter uma representação compacta preservando a maior parte da informação contida nos dados. O modelo é construído a partir da minimização do erro de reconstrução. Os métodos dessa categoria podem diferir no tipo de erro de reconstrução ou na rotina de otimização utilizada. Novas observações são descritas por meio deste modelo para então classificá-las. A ideia é que *outliers* não deveriam ser satisfatoriamente descritos pelo modelo e, consequentemente, espera-se que tenham um erro de reconstrução alto. Entre os métodos baseados em reconstrução está o Auto-Encoder, apresentado a seguir.

Auto-Encoder

Auto-Encoder (TAX, 2001a) é uma rede neural com unidades sigmóide tangente hiperbólicas, uma única camada escondida, e um parâmetro que define o número de neurônios na camada escondida que serão utilizados no treinamento na classe *inlier*. A fim de classificar novas observações, cada observação a ser classificada é fornecida como entrada para a rede, e a diferença entre a entrada original e da saída da rede em termos de erro quadrático médio é calculada.

2.3 Detecção Supervisionada de *Outliers*

Outra categoria de técnicas, mas que não faz parte do escopo deste trabalho, são as técnicas que utilizam aprendizado supervisionado. Estas técnicas pressupõem que se dispõe de um conjunto apropriado de observações rotuladas como *outlier* e *inlier* para criar um modelo de classificação a partir dos dados rotulados. A detecção supervisionada de *outliers* pode ser vista como um caso especial de classificação. O problema de classificação tem sido amplamente estudado, com uma enorme quantidade de técnicas disponíveis na literatura (TAN; STEINBACH; KUMAR, 2006; HAN; KAMBER; PEI, 2011; DUDA; HART; STORK, 2001). No contexto de detecção de *outliers*, essas técnicas necessitam lidar com o problema do desbalanceamento entre as classes; o fato da classe anômala ser geralmente muito desbalanceada em relação à classe normal dificulta a criação de um modelo que generalize bem a classe de interesse. Para a construção de modelos que melhor generalizem o problema, duas abordagens são tipicamente utilizadas: i) *Cost sensitive learning*: A função objetivo do classificador é modificada para ponde-

rar os erros de classificação cometidos, classes com menos observações possuem pesos maiores (DOMINGOS, 1999; ZADROZNY; ELKAN, 2001; ZADROZNY; LANGFORD; ABE, 2003); ii) *Adaptive re-sampling*: Os dados são re-amostrados de modo a diminuir o desbalanceamento das classes (CHAWLA; JAPKOWICZ; KOTCZ, 2004; CHAN; STOLFO, 1998; DRUMMOND; HOLTE, 2003).

Devido ao contexto supervisionado possuir um conjunto apropriado de observações rotuladas, técnicas bem estabelecidas na área de classificação também podem ser utilizadas na avaliação do modelo. Entretanto, a utilização de tais abordagens de forma trivial pode levar a resultados enganosos. Por exemplo, em casos em que 99% das observações pertencem à classe normal, pode-se conseguir 99% de acurácia utilizando um classificador que apenas rotule todas as observações como normais. A utilização de medidas que consigam lidar com problemas desbalanceados, como por exemplo área sob a curva ROC, medida-F e *precision-at-n*, são mais indicadas para este cenário (TAN; STEINBACH; KUMAR, 2006; HAN; KAMBER; PEI, 2011; CAMPOS *et al.*, 2016).

CAPÍTULO

3

AVALIAÇÃO E SELEÇÃO DE MODELOS EM DETECÇÃO DE *OUTLIERS*

Neste Capítulo são apresentas duas das medidas externas mais utilizadas na literatura de detecção de *outliers*, assim como a única medida interna conhecida na literatura para avaliação de resultados de detecção de *outliers*.Especificamente, na seção 3.1.1 é apresentada uma medida externa para avaliação de soluções top-*n* de detecção de *outliers* e na seção 3.1.2 uma medida externa para avaliação de *rankings*. Na seção 3.2.1 é apresentada uma revisão detalhada do índice interno e relativo IREOS que será utilizado como base para este trabalho.

3.1 Avaliação Externa em Detecção de *Outliers*

O procedimento de avaliação de algoritmos de detecção de *outliers* usualmente adotado na literatura é baseado na utilização de bases de dados previamente rotuladas, nas quais os *outliers* (segundo alguma intuição particular) são previamente conhecidos. A categorização binária ou o *ranking* das observações produzidos pelo algoritmo são comparados com a rotulação considerada correta e previamente conhecida (*ground truth*). Tais procedimentos, conhecidos como avaliação (ou validação) externa, originalmente surgiram nas áreas de classificação de padrões e recuperação de informação onde a suposição de uma base de dados pré-rotulada é válida. Entretanto, em detecção não supervisionada de *outliers*, onde tal suposição é inválida, esses procedimentos de avaliação não são viáveis em aplicações práticas, limitando-se apenas à experimentos controlados.

3.1.1 Precision-at-n

Seja um problema de detecção de *outliers* em que se deseja avaliar uma dada solução **S**, que consiste do conjunto de observações rotuladas como *outliers* por algum método. Considere o *ground truth*, **G**, que consiste do conjunto de observações previamente sabidas serem os corretos

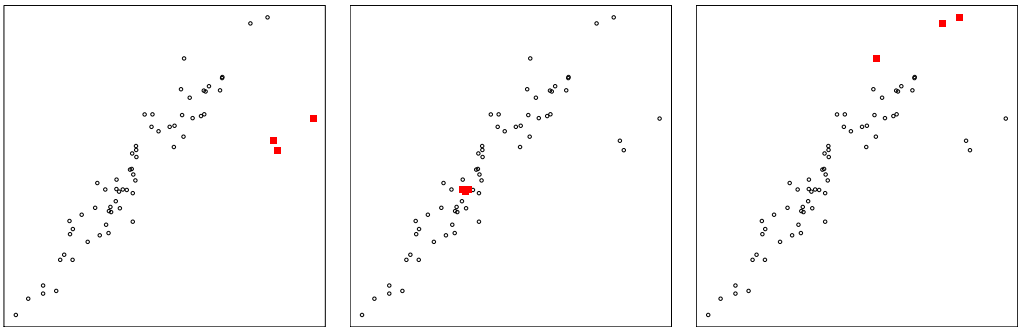
outliers. *Precision-at-n* (*prec@n*) mede a fração de *outliers* no *ground truth* **G** que são detectados como *outliers* na solução **S** que está sendo avaliada (Equação (3.1)).

$$prec@n = \frac{|\mathbf{S} \cap \mathbf{G}|}{|\mathbf{S}|} \tag{3.1}$$

Como a maioria dos algoritmos não produzem uma solução binária e sim um *ranking*, pode-se definir um liminar no *ranking* para rotulação das observações como *outliers*. Tipicamente a escolha desse limiar coincide com o número de *outliers* presente no *ground truth* (*n*), mas também pode ser de interesse medir *prec@2n*, *prec@3n* ou a *precision* em algum outro ponto do *ranking*. A modificação do ponto em que será medido a *precision* ajuda a resolver uma limitação que ocorre quando os *outliers* segundo o *ground truth* são ranqueados logo abaixo das *n* primeiras observações. Por exemplo, um *ranking* de 100 observações em que os dois únicos *outliers* aparecem na terceira e quarta colocação terá *precision-at-n* igual 0, apesar do algoritmo ranquear bem ambos *outliers*.

Uma segunda limitação apresentada por essa medida é o fato dela, por muitas vezes, avaliar soluções muito ruins e soluções boas com a mesma qualidade. Por exemplo, vide a Figura 4, em que a solução da Figura 4b é avaliada como sendo tão boa quanto a solução 4c de acordo com a *precision-at-n*.

Figura 4 – Ambas soluções com *precision-at-n* = 0 de acordo com o *ground truth*



(a) *Ground truth* (b) *Precision-at-n* = 0 (c) *Precision-at-n* = 0

Um terceiro problema é que a medida não é comensurável para diferentes valores de *n*. Para permitir a comparação entre soluções com diferentes valores de *n*, Campos *et al.* (2016) propõem a utilização do clássico *framework* estatístico para ajuste para aleatoriedade (HUBERT; ARABIE, 1985), dando origem a uma chamada *adjusted precision-at-n*.

3.1.2 Área sob a Curva ROC

Área sob a Curva *Receiver Operating Characteristic* (ROC) é a medida mais utilizada na literatura de detecção de *outliers*. Se for dada uma solução que consiste no *ranking* das *N* observações da base de dados a partir do grau com que cada uma delas se caracteriza como *outlier* de acordo com o método utilizado, pode-se plotar a curva ROC ou sumarizá-la pela

sua área para comparar diferentes soluções. A curva ROC avalia o *ranking* de acordo com o *ground truth* binário plotando a taxa de verdadeiros positivos ao longo do eixo y pela taxa de falsos positivos ao longo do eixo x. Cada ponto ao longo da curva corresponde à variação do limiar que definirá quais observações deverão ser rotuladas como *outliers* e *inliers*. Um *ranking* aleatório resultaria em uma curva próxima a diagonal, com área próxima de 0.5, enquanto um *ranking* perfeito, em que todos os *outliers* seriam ranqueados primeiro, resultaria em uma linha vertical no eixo da taxa de falsos positivos e uma linha horizontal no eixo da taxa de verdadeiros positivos, indicando que as taxas de verdadeiros positivos se manteriam constantes em 1 para todas as taxas de falsos positivos maior que 0, resultando em uma área igual a 1.

A curva ROC também possui limitações para avaliar soluções de detecção de *outliers*. Uma vez que esse procedimento avalia *rankings*, as informações dos *scorings* são completamente descartadas. Além disso, para conseguir a área máxima possível é necessário apenas que os *outliers* sejam ranqueados acima dos *inliers*. Porém, existem $n!(N-n)!$ soluções com a área máxima possível em uma base de dados com N observações e n *outliers*. Todas elas são igualmente perfeitas de acordo com esta medida, independentemente de que observações que se caracterizem mais como *outlier* sejam ranqueadas depois de observações que se caracterizem menos.

3.2 Avaliação Interna em Detecção de *Outliers*

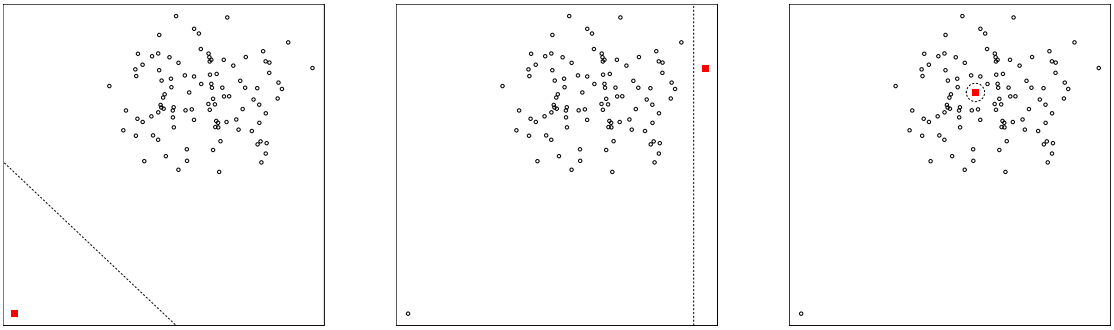
Embora o procedimento de avaliação descrito na seção anterior seja útil no contexto supervisionado, ele não é viável em aplicações práticas do contexto não supervisionado divido à indisponibilidade de bases de dados pré-rotuladas, limitando a aplicação de tais medidas, neste contexto, apenas a experimentos controlados durante o desenvolvimento de novos algoritmos. No cenário não supervisionado, faz-se necessário procedimentos de avaliação que utilizem critérios internos, ou seja, independentes de rótulos pré-conhecidos (externos). Entretanto, apesar de tais índices terem se mostrado úteis em agrupamento de dados, raramente são abordados em detecção de *outliers*. Um índice pioneiro e até então o único em detecção de *outliers* foi recentemente proposto para avaliação interna de soluções top-n, este índice é apresentado em maiores detalhes a seguir.

3.2.1 *IREOS*

A intuição básica por trás do índice *Internal, Relative Evaluation of Outlier Solutions* (IREOS) (MARQUES, 2015; MARQUES *et al.*, 2015) é que um *outlier* é uma observação que de certa forma está mais distante e então pode ser mais facilmente separada das demais observações que um *inlier*. IREOS utiliza classificadores de máxima margem, tais como SVM ou *Kernel Logistic Regression* (KLR) (HASTIE; TIBSHIRANI; FRIEDMAN, 2013; BISHOP, 2006), para quantificar o quão difícil é separar das demais observações cada observação rotulada como

outlier na solução que está sendo avaliada. O índice utiliza a distância de cada observação para a fronteira de decisão como base para uma medida de separabilidade da observação. Esta ideia é ilustrada na Figura 5. As Figuras 5a, 5b e 5c destacam diferentes observações rotuladas como *outlier* (quadrado vermelho) em hipotéticas soluções de detecção de *outliers* a serem avaliadas. Na Figura 5a pode-se ver que a observação destacada é um *outlier* global, que está bem distante do hiperplano (linha tracejada) do classificador que a separa das demais observações. Na Figura 5b, por se tratar de um *outlier* local, a margem do classificador é menor, mas ainda assim é mais larga que a da Figura 5c. Neste último caso, a observação rotulada é sem dúvida um *inlier* e não apenas a margem é bem pequena como também é necessária uma fronteira de decisão complexa (i.e. não linear).

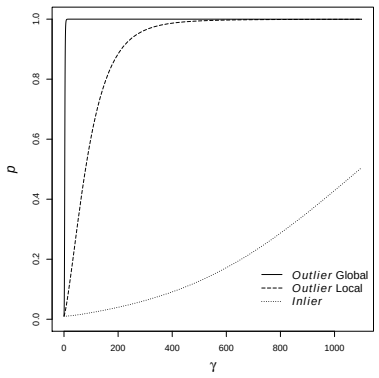
Figura 5 – Dificuldade em separar uma observação das demais



(a) *Outlier* Global (b) *Outlier* Local (c) *Inlier*

Para lidar com possíveis fronteiras de decisão não linear, como no exemplo da Figura 5c, IREOS utiliza uma função *kernel* para transformar o espaço de entrada em um espaço linearmente separável. Originalmente, o índice utiliza o *kernel* radial que demanda a configuração do parâmetro γ. Este parâmetro é positivamente relacionado ao grau de não linearidade da fronteira de decisão, isto é, quanto maior o valor de γ mais complexa será a fronteira de decisão, consequentemente, maior a dificuldade de separação. Para eliminar a necessidade de escolha de um valor particular de γ, o índice parte da premissa que o grau de dificuldade de separação de observações fáceis serão sempre menores que o grau de dificuldade de separação de observações difíceis, embora a diferença relativa entre eles possam se alterar com diferentes valores de γ. Essa ideia é ilustrada na Figura 6, em que é utilizado um classificador de máxima margem com *kernel* para separar as três observações rotuladas como *outliers* na Figura 5. No eixo horizontal o valor de γ é variado de 0 até γ_{max} (valor necessário para que todas observações rotuladas como *outliers* sejam individualmente separadas das demais observações na base de dados), enquanto o eixo vertical apresenta uma medida $p(\mathbf{x}_i, \gamma)$ que quantifica em um intervalo normalizado o quão longe cada observação $\mathbf{x}_i$ está da fronteira de decisão. Observe que, não importa o valor de γ, $p(\mathbf{x}_i, \gamma)$ é sempre maior para $\mathbf{x}_i$ dado pelo *outlier* global (Figura 5a) do que para o *outlier* local (Figura 5b), que por sua vez é maior do que para o *inlier* (Figura 5c), embora as diferenças relativas se alterem.

Figura 6 – Separabilidade



A separabilidade da observação $\mathbf{x}_i$ é medida computando a área sob a curva (*Area Under the Curve* (AUC)) do intervalo de valores de γ, i.e. $\int_{\gamma=0}^{\gamma_{max}} p(\mathbf{x}_i, \gamma)$. Para avaliar a qualidade de uma dada solução $\mathbf{S}$, é tomada a média das curvas de separabilidade de cada observação $\mathbf{x}_i \in \mathbf{S}$, i.e. $\bar{p}(\gamma) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathbf{S}} p(\mathbf{x}_i, \gamma)$, então computa-se a área sob essa curva, $\int_{\gamma=0}^{\gamma_{max}} \bar{p}(\gamma)$. Para definir um índice preliminar a AUC é normalizada em $[0,1]$ dividindo-a pela sua área máxima ($\gamma_{max} \times 1$):

$$\frac{1}{\gamma_{max}} \int_{\gamma=0}^{\gamma_{max}} \bar{p}(\gamma) \tag{3.2}$$

Para calcular $\int_{\gamma=0}^{\gamma_{max}} \bar{p}(\gamma)$, γ precisa ser discretizado em um número finito de valores no intervalo de $[0, \gamma_{max}]$ para treinar o classificador e então computar $\bar{p}(\gamma)$ para cada γ. Caso o intervalo de γ seja uniformemente discretizado em n_γ valores, o índice preliminar pode ser computado da seguinte forma:

$$I(\mathbf{S}) = \frac{1}{\gamma_{max}} \int_{\gamma=0}^{\gamma_{max}} \bar{p}(\gamma) \approx \frac{1}{\gamma_{max}} \frac{(\gamma_{max} - 0)}{n_\gamma} \sum_{l=1}^{n_\gamma} \bar{p}(\gamma_l) = \frac{1}{n_\gamma} \sum_{l=1}^{n_\gamma} \bar{p}(\gamma_l) \tag{3.3}$$

Modelagem de Clumps

Clumps são conjuntos de observações próximas umas às outras, mas muito pequenos para serem considerados *clusters*. Caso o usuário queira considerar *clumps* como parte do modelo de avaliação é inserido um parâmetro opcional que permite ao usuários expressar o que eles julgam muito pequeno para ser interpretado como *cluster*. O parâmetro m_{clSize} define o tamanho máximo do *clump*, que representa o que o usuário acredita ser o número de observações a partir do qual um grupo de observações seja mais um *cluster* do que um aglomerado de potenciais *outliers*.

Outra intuição capturada com a modelagem de *clumps* é com relação à avaliação de observações rotuladas como *outliers* próximas de outras observações (e.g. observação em um *clump*). O índice afeta negativamente a avaliação dessas observações pelo fato de existirem outras observações próximas, entretanto, com a modelagem de *clumps* o impacto negativo é maior se a observação próxima estiver com rótulo diferente (i.e., rotulada como *inlier*). Para isso, o índice utiliza classificadores de margem suave com penalidade individual para cada observação (OSUNA; FREUND; GIROSI, 1997). Esses classificadores permitem que observações fiquem

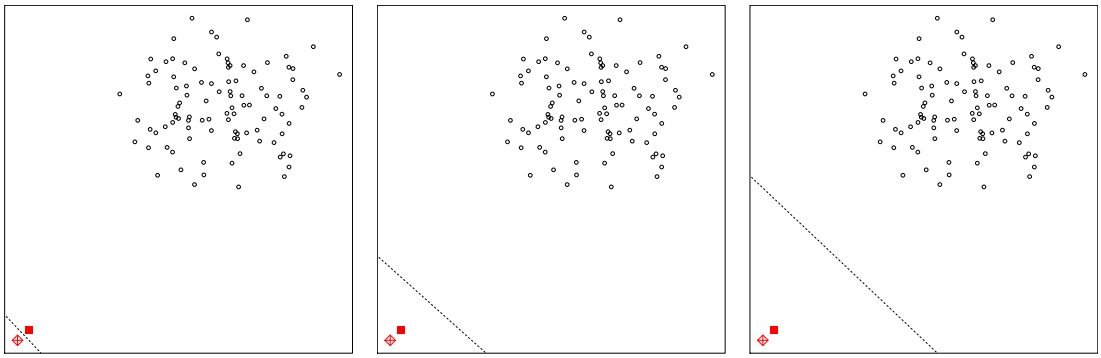
do "lado errado" da margem com relação à sua classe, ao preço de uma penalidade P que é incorporada à função objetivo do classificador:

$$P = \sum_{i=1}^{N} C_i g(\xi_i) \tag{3.4}$$

onde $g(\xi_i)$ mede o quão longe a observação está da margem da sua classe. Quanto mais longe a observação $\mathbf{x}_i$ estiver do lado incorreto da margem da sua classe, maior a penalidade. Os valores de C_i associados às observações $\mathbf{x}_i$ depende das suas rotulações: caso a observação seja rotulada como *inlier*, o custo (penalidade) para violação da margem é integral; entretanto, se a observação for rotulada como *outlier*, apenas uma fração $\beta \in [0,1]$ de C é utilizada, ou seja, na Equação (3.4), $C_i = C$ ou βC dependendo da rotulação de $\mathbf{x}_i$. Repare que, no problema original de classificação, C controla o compromisso entre *overfitting* e *underfitting*; aqui, entretanto, não é de interesse a generalização do classificador, então C precisa apenas ser grande o suficiente para que o classificador consiga classificar as observações $\mathbf{x}_i \in \mathbf{S}$. Quando $\beta = 1$, o índice se reduz ao índice preliminar, em que todas observações são tratadas igualmente independente dos seus rótulos. Para $\beta = 0$, as observações rotuladas como *outliers* não são penalizadas por violação da margem. Repare que, neste outro extremo, quando se avalia a separabilidade de uma observação específica, isto equivale a remover todas as outras observações rotuladas como *outliers* da base de dados. A Figura 7 ilustra esta intuição. Na Figura 7a, quando avaliada a separabilidade da observação original (aqui rotulada em losango) utilizando $\beta = 1$, a violação da margem pela observação adicional com relação a Figura 5, também rotulada como *outlier*, será tão cara quanto a violação de uma observação rotulada como *inlier*, funcionando como o índice preliminar. Na Figura 7b, a avaliação da mesma observação, mas agora utilizando $\beta = 0.5$, faz com que a violação da margem pela observação adicional seja menos penalizada, fazendo com que passe valer a pena violar a margem para a observação adicional fazendo com que a observação avaliada fique mais distante da fronteira de decisão, consequentemente aumentando seu grau de separabilidade. Na Figura 7c, a observação original é avaliada para $\beta = 0$, neste cenário a violação da margem pela observação adicional não será penalizada e a fronteira de decisão será a mesma da Figura 5a, como se a observação adicional não estivesse presente na base de dados. A escolha de β controla então a influência que a rotulação das outras observações terão no momento da avaliação de uma determinada observação.

Para modelagem de possíveis *clumps*, dado um tamanho máximo de *clump* (m_{clSize}), configura-se a fração da penalidade C acima como $\beta = 1/m_{clSize}$. Desta forma, m_{clSize} será o único parâmetro opcional na medida de avaliação. Conforme m_{clSize} aumenta, β diminui e as observações rotuladas como *outliers* no *clump* vão individualmente afetar cada vez menos as outras ao se medir a separabilidade dessas, de forma que um número maior de observações próximas será necessário para conseguir um certo impacto negativo (penalidade). Note que utilizando $\beta = 1/m_{clSize}$ são necessárias m_{clSize} observações rotuladas como *outliers* para conseguir o mesmo impacto de apenas um *inlier*, ou seja, uma penalidade βC m_{clSize} vezes, portanto igual a $m_{clSize} \times \frac{1}{m_{clSize}} \times C = C$. Note também que em problemas top-n de detecção de *outliers*

Figura 7 – Fronteira de decisão do classificador para avaliação da observação em losango



(a) Avaliação da observação em losango para $\beta = 1$

(b) Avaliação da observação em losango para $\beta = 0.5$

(c) Avaliação da observação em losango para $\beta = 0$

seria contraditório utilizar $m_{clSize} > n$. Pelo fato de haver somente n observações rotuladas como *outliers* nas soluções sendo avaliadas, utilizar $m_{clSize} > n$ seria sugerir que aglomerados com mais de n observações devam ser considerados como possíveis *outliers* em um *clump*, mesmo assumindo a presença de apenas n *outliers* na base de dados. Sendo assim, conceitualmente este parâmetro deveria ser utilizado entre $1 \leq m_{clSize} \leq n$. Exceto quando $m_{clSize} = 1$, a separabilidade de cada observação depende dos rótulos das demais observações e, assim, procurar por uma solução que maximize o índice proposto (em vez de utilizá-lo para avaliar uma solução) torna-se uma tarefa complexa.

Em resumo, assim como o índice preliminar, IREOS é calculado utilizando a Equação (3.3). Entretanto, para o cálculo do termo $p(\mathbf{x}_i, \gamma)$, os *outliers* recebem apenas uma fração $\beta = 1/m_{clSize}$ da penalidade atribuída aos *inliers* em caso de violação da margem.

Ajuste para Aleatoriedade

O índice IREOS apresentado pode ser utilizado para comparação em termos relativos de diferentes soluções candidatas, por exemplo, para seleção de modelos. Entretanto, a avaliação em termos absolutos do índice para soluções individuais, por exemplo, para validação estatística, pode ser mal interpretada pelo fato do índice fornecer valores não nulos até mesmo quando as soluções são puramente aleatórias. Para interpretação e avaliação das soluções em termos absolutos é necessário ajustar o índice para aleatoriedade, o que pode ser feito utilizando o clássico *framework* estatístico de ajuste para aleatoriedade (HUBERT; ARABIE, 1985). O índice pode ser ajustado analiticamente, ou, para reduzir o custo computacional, o ajuste pode ser feito alternativamente utilizando simulações de Monte Carlo.

CAPÍTULO

4

RESULTADOS PRELIMINARES

Neste Capítulo são discutidos os resultados preliminares e/ou direções a serem exploradas com relação às hipóteses de pesquisa que serão investigadas neste trabalho. O Capítulo é divido em sete seções, em cada seção uma das hipóteses de pesquisa é discutida com relação às possíveis abordagem que serão utilizadas para sua validação e, em alguns casos, resultados preliminares são apresentados.

4.1 Hipótese I

> *Utilizando o framework proposto por Janssens, Flesch e Postma (2009) para adaptação de métodos não supervisionados de detecção de outliers para o contexto semissupervisionado, outros métodos não supervisionados podem também ser adaptados e um estudo comparativo mais completo e rigoroso do que aquele descrito na referência acima pode ser realizado. Acredita-se que resultados mais consistentes e conclusões mais confiáveis possam ser obtidas deste estudo.*

Os resultados reportados nesta seção foram resultantes de uma colaboração internacional na qual este projeto está também inserido. Os resultados foram submetidos e aceitos na *3rd IEEE International Conference on Data Science and Advanced Analytics* (DSAA'2016) pelo autores Lorne Swersky (University of Alberta), Henrique O. Marques (Universidade de São Paulo), Jörg Sander (University of Alberta), Ricardo J. G. B. Campello (Universidade de São Paulo) e Arthur Zimek (University of Southern Denmark) no artigo intitulado como *"On the Evaluation of Outlier Detection and One-Class Classification Methods"*. O autor desta monografia contribuiu neste artigo programando e adaptando parte dos códigos utilizados, rodando os experimentos, participando ativamente das discussões e também da análise dos resultados.

4.1.1 Avaliação de Métodos de Detecção de Outliers e One-Class Classification

Embora os métodos de detecção não supervisionada e semissupervisionada de detecção de *outliers* tenham surgido no campo da estatística (BARNETT; LEWIS, 1994; MARKOU; SINGH, 2003), pouco tem sido feito na literatura para se comparar ambas as categorias de algoritmos em termos de performance. Devido ao fato dos métodos semissupervisionados possuírem disponível a informação de uma classe, esses geralmente abordam uma tarefa mais fácil. De fato, enquanto os métodos semissupervisionados estimam um modelo da classe *inlier* sem precisarem se preocupar como a possível presença de *outliers* que podem distorcer o modelo, os métodos não supervisionados precisam lidar com a presença de possíveis *outliers* enquanto estimam o modelo. Dada a diferença entre os métodos dessas duas categorias, não é trivial compará-los.

Seguindo a abordagem proposta em Janssens e Postma (2009), métodos não supervisionados podem ser estendidos para usar a informação disponível dos *inliers* para então serem aplicados de forma semissupervisionada. Utilizando esta abordagem, foi realizado um estudo comparativo entre métodos semissupervisionados e métodos não supervisionados adaptados para o cenário semissupervisionado, melhorando um estudo comparativo anterior (JANSSENS; FLESCH; POSTMA, 2009) nos seguintes aspectos:

- Quando reproduzidos os experimentos de Janssens, Flesch e Postma (2009), que reportam médias de 5 repetições usando 5-*fold cross-validation*, foi notada uma grande variabilidade dos resultados. Para aumentar a confiança dos resultados, realizaram-se 30 repetições usando 10-*fold cross-validation*.

- Aumentou-se o número de bases de dados utilizadas de 24 para 433 (400 variantes de uma coleção de imagens, mais 33 outras bases de dados), e o número de métodos comparados de 5 para 11.

- Além da ROC AUC utilizada por Janssens, Flesch e Postma (2009), também foi utilizada *Adjusted Precision-at-n* (AjustedPrec@n), como definido em Campos *et al.* (2016), para medir a performance. Essas medidas se complementam e juntas fornecem um visão mais completa da performance dos métodos (CAMPOS *et al.*, 2016).

- Além do tipo de experimento realizado por Janssens, Flesch e Postma (2009), onde uma classe é rotulada como *inlier* e as outras classes são rotuladas como *outliers*, também foi realizado um segundo tipo de experimento onde uma classe é rotulada como *outlier* e as outras classes são rotuladas como *inliers*. Ambos tipos de experimentos representam possíveis cenários de aplicação real. É mostrado que algumas das conclusões se alteram dependendo do tipo de experimento.

- Também incluiu-se uma adaptação de um método recente de detecção de *outliers* — chamado GLOSH (CAMPELLO *et al.*, 2015b) — para o problema de OCC.

- Por fim, apresentou-se uma discussão original sobre princípios que não deveriam ser violados quando métodos de detecção não supervisionada de *outliers* são adaptados para a tarefa de OCC. Todos os códigos utilizados nos experimentos foram cuidadosamente examinados e ajustados para que esses princípios não fossem violados quando produzidos os resultados reportados.

Adaptando Métodos de Detecção Não Supervisionada de Outliers para OCC

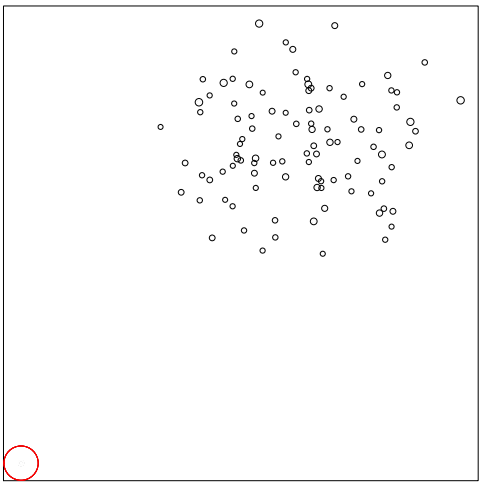
Os métodos de detecção não supervisionada de *outliers* têm em comum o fato de computarem um *score* para cada observação. Para usar um método de detecção não supervisionada de *outliers* em OCC, a estratégia geral é a seguinte: Primeiro executa-se o método não supervisionado nos dados de treinamento e pré-computa-se o *score* para cada *inlier*. Em seguida, calcula-se o *score* para a nova observação a ser classificada, possivelmente usando quantidades pré-computadas (por exemplo, densidades ou distâncias para os vizinhos mais próximos) a partir das observações dos dados de treinamento. Então, para classificar a nova observação, compara-se o seu *score* com os *scorings* pré-computados dos *inliers*.

A Figura 8 exemplifica a estratégia geral citada acima. Em que os círculos em preto representam os *scorings* dos *inliers* do conjunto de treinamento e o círculo em vermelho o *scoring* da nova observação a ser classificada. Utilizando um algoritmo não supervisionado, inicialmente apenas os *scorings* dos *inliers* (conjunto de treinamento) são computados. Em seguida, utilizando o mesmo algoritmo não supervisionado, o *scoring* da observação em vermelho é calculado. A partir do *scoring* da observação em vermelho pode-se classificá-la como *outlier*, devido ao seu *scoring* de *outlier* ser superior a todos os *scorings* dos *inliers* do conjunto de treinamento, caso houvesse ao menos um *inlier* no conjunto de treinamento com *scoring* superior ao da nova observação a ser classificada, esta seria então classificada como *inlier*.

Existem dois importantes aspectos relacionados ao uso das quantidades pré-computadas a partir do dados de treinamento quando o *score* de uma nova observação a ser classificada é computado. Primeiro, quando múltiplas observações são classificadas, não existe a necessidade de recalcular essas quantidades novamente para cada nova observação, uma vez que elas dependem apenas dos dados de treinamento e podem então ser pré-computadas, o que torna os cálculos mais rápidos.

Segundo e mais importante, utilizar quantidades pré-computadas a partir das observações dos dados de treinamento assegura que o modelo não seja de forma alguma afetado pelas novas observações a serem classificadas. Este é um princípio básico de OCC, observações não rotuladas não deveriam afetar o modelo pré-computado a partir dos *inliers*, uma vez que elas podem ser *outliers*. Por exemplo, suponha que um determinado algoritmo funciona através da comparação

Figura 8 – Exemplificação da estratégia geral para adaptação de um método de detecção não supervisionada para OCC.



entre a densidade de uma nova observação a ser classificada e as densidades dos seus vizinhos mais próximos dos dados de treinamento. Neste caso, as densidades dos *inliers* deveriam ser pré-calculadas, não podendo assim ser afetadas pela presença da observação não rotulada que está sendo avaliada; de outra forma, cada observação não rotulada iria afetar o modelo de uma forma diferente, o que significa que as diferentes observações seriam classificadas por diferentes modelos/critérios.

Ao classificar múltiplas observações, recomenda-se também que o procedimento de classificação descrito acima seja realizado de forma independente para cada observação. Desta forma, as diferentes observações não rotuladas não afetarão as avaliações umas das outras. A razão pela qual se classifica apenas uma observação de cada vez, ao invés de múltiplas observações de uma vez só, é porque não se faz suposições sobre a natureza de cada observação em relação ao conjunto de dados combinado como um todo. É possível que observações que deveriam ser classificadas como *outliers*, no sentido de não pertencerem à classe *inlier*, possam estar agrupadas de tal modo que um método não supervisionado não as detectariam como *outliers*.

Configuração dos Experimentos

Foram comparados e avaliados 11 dos algoritmos descritos no Capítulo 2: ABOD, Auto-Encoder, Gaussian, GLOSH, KNNOutlier (kNN_{global}), kNNDD (kNN_{local}), LOCI, LOF, LP, PW e SVDD. Foi utilizado o código do repositório disponível em <http://prlab.tudelft.nl/users/david-tax/> (TAX, 2015) para a maioria dos métodos comparados, exceto por LOF, LOCI, kNN_{local} e GLOSH. No caso do LOF e LOCI, suas implementações foram modificadas para garantir que novas observações a serem classificadas não afetassem o modelo pré-computado a partir dos *inliers*, seguindo as orientações previamente discutidas na seção anterior. Como kNN_{local} não está disponível no repositório, foi utilizada uma implementação própria do algoritmo. GLOSH foi adaptado baseado na implementação do HDBSCAN* disponível em

<http://lapad-web.icmc.usp.br/>.

Foram utilizadas 31 bases de dados reais do *University of California Irvine* (UCI) *Machine Learning Repository* (LICHMAN, 2013) pré-processadas para OCC e disponibilizadas em <http://prlab.tudelft.nl/users/david-tax/>: Abalone, Arrhythmia, Balance-scale, Ball-bearing, Biomed, Breast, Cancer, Colon, Delft1x3, Delft2x2, Delft3x2, Delft5x1, Delft5x3, Diabetes, Ecoli, Glass, Heart, Hepatitis, Housing, Imports, Ionosphere, Iris, Liver, Satellite, Sonar, Spectf, Survival, Vehicle, Vowels, Waveform e Wine.

Também foram utilizadas as bases de dados CellCycle-237 e YeastGalactose, disponibilizadas por Yeung *et al.* (2001), Yeung, Medvedovic e Bumgarner (2003), assim como uma coleção de 400 bases de dados baseadas na *Amsterdam Library of Object Images* (ALOI) (GEUSEBROEK; BURGHOUTS; SMEULDERS, 2005), criada conforme descrito em Horta e Campello (2012). Especificamente, a coleção foi criada a partir da seleção aleatória de 2, 3, 4 ou 5 categorias de imagens da ALOI para serem utilizadas como rótulos de classe e, em seguida, amostradas 25 imagens de cada uma das categorias selecionadas, resultando em conjuntos de dados contendo 2, 3, 4 ou 5 classes e 50, 75, 100 ou 125 imagens (observações). Estas imagens são descritas por seis descritores: momentos de cor (144 atributos), estatísticas da textura extraídos da matriz de co-ocorrência de nível de cinza (88 atributos), histograma da borda (Sobel) (128 atributos), primeiro momento estatístico do histograma dos níveis de cinza (5 atributos), *gray-level run-length matrix features* (44 atributos) e o histograma de níveis de cinza (256 atributos). PCA foi aplicado a cada conjunto de vetores de atributos separadamente e o primeiro componente principal resultante de cada conjunto foi extraído. Os primeiros componentes extraídos são então combinados de tal forma que cada imagem é então descrita por um vetor com seis atributos.

No total, foram utilizadas 433 bases de dados reais multi-classes. Os resultados das bases de dados ALOI e Delft foram sumarizados, por serem variantes obtidas a partir da mesma fonte. Por fim, devido à incapacidade de alguns algoritmos em lidar com observações duplicadas, as duplicatas foram removidas das bases de dados em que estão presentes.

A fim de avaliar o desempenho de um método em uma base de dados de classe única, o seguinte procedimento foi aplicado: Primeiro, foi dividido o conjunto de dados em 2 subconjuntos, um contendo 20% e o outro contendo 80% dos dados. No subconjunto com 80% dos dados foi aplicado um procedimento de 10-*fold cross-validation* para otimizar os parâmetros dos métodos com relação à ROC AUC.

Os parâmetros dos métodos foram optimizados nos seguintes intervalos: $k = 1, 2, \cdots, 50$ para LOF, kNN_{global} e kNN_{local}; $M_{clSize} = M_{pts} = 1, 2, \cdots, 50$ para GLOSH; Número de neurônios = 2, 5, 7, 10, 12, 15, 17, 20, 22, 25 para Auto-Encoder[1], $h = 0.001$ a 50 (discretizado logaritmicamente em 25 diferentes valores) para Gaussian e para o *kernel* Gaussiano usado pelo

[1] Devido à alta demanda computacional do Auto-Encoder, um limite de tempo de 1000s foi imposto à convergência da rede, o que dá um montante máximo de tempo igual a 30 repetições $\times$ 10 *folds* $\times$ 10 parâmetros $\times$ 1000s $\approx$ 35 dias por experimento envolvendo uma única base de dados.

SVDD, $\alpha = 0.1, 0.2, \cdots, 1.0$ para LOCI, LP, PW e SVDD.

Após a otimização dos parâmetros, o subconjunto contendo 20% dos dados (conjunto de teste) é utilizado para medir o desempenho dos métodos (treinado com os valores de parâmetros ótimos do 10-*fold cross-validation*). A fim de obter resultados mais confiáveis, este procedimento foi repetido 30 vezes, e os valores de ROC AUC resultantes são agregados e reportados.

Para efeito de comparação com os resultados reportados por Janssens, Flesch e Postma (2009), também foi calculado a *Weighted* ROC AUC utilizada em seu trabalho, que dá mais peso aos resultados obtidos a partir de experimentos envolvendo classes *inlier* maiores. Vale ressaltar, no entanto, que esta abordagem é questionável, uma vez que é sabido que as curvas ROC são inerentemente ajustadas para o desbalanceamento de classes.

Além dos valores de ROC AUC, também são reportados os valores de AjustedPrec@n (CAMPOS *et al.*, 2016) para os resultados de classificação obtidos nos conjuntos de teste. Enquanto ROC AUC leva todo o conjunto de teste em consideração, AjustedPrec@n avalia apenas as observações top-n, conforme discutido no Capítulo 3.

Um alto valor de ROC AUC indica apenas que, de forma geral, *outliers* são mais propensos a serem ranqueados à frente dos *inliers*; não significa necessariamente que as primeiras posições no *ranking* são dominadas por *outliers*. Portanto, seguindo o estudo extensivo de Campos *et al.* (2016), argumenta-se que não se pode confiar apenas na ROC AUC para julgar a qualidade de um método de detecção; em vez disso, ROC AUC e AdjustedPrec@n se complementam por revelarem diferentes aspectos do *ranking*, sendo ambas relevantes na prática.

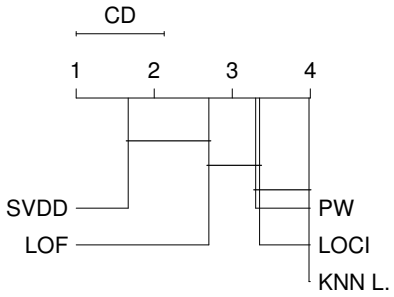
Foram realizados dois tipos de experimentos. No **primeiro tipo** de experimento (Tipo I), foi seguida a única abordagem adotada por Janssens, Flesch e Postma (2009), onde as bases de dados multi-classe são transformadas em bases de dados de classe única re-rotulando uma classe como *inlier*, e as demais classes como *outliers*. Exceto nas bases de dados onde apenas uma única classe foi pré-definida como *inlier* no repositório de dados (<http://prlab.tudelft.nl/users/david-tax/>) — por exemplo, Ecoli — repetiu-se o procedimento para cada classe pré-definida como *inlier* na base de dados, e apenas a média dos resultados foi reportada.

No **segundo tipo** de experimento (Tipo II), inverteu-se as classes *inlier* e *outlier* obtidas no primeiro tipo de experimento para as bases de dados que têm mais de duas possíveis classes *inlier* definidas. Este tipo de experimento é importante, uma vez que modela situações com uma possível classe *inlier* multi-modal. Note-se que os experimentos Tipo II apenas diferem do Tipo I — e, portanto, apenas são reportados — para bases de dados com 3 ou mais classes. Por esta razão, os resultados dos experimentos Tipo II estão disponíveis apenas para um subconjunto das bases de dados consideradas nos experimentos Tipo I.

Resultados

A Figura 9 mostra o *ranking* médio dos 5 métodos comparados em Janssens, Flesch e Postma (2009) em todos os experimentos Tipo I com relação à *Weighted* ROC AUC. Esta figura resume a tentativa de reproduzir os resultados reportados em Janssens, Flesch e Postma (2009), que foram restritos a apenas esta configuração particular. A largura da barra superior (CD) indica a distância crítica do bem conhecido teste estatístico Friedman/Nemenyi (IMAN; DAVENPORT, 1979; NEMENYI, 1963) utilizando nível de significância $\alpha = 0.05$. A figura análoga em Janssens, Flesch e Postma (2009) mostra dois subconjuntos de métodos: (1) os de melhores desempenhos SVDD, LOF, e KNN_{local} (com SVDD e LOF tendo exatamente o mesmo *rank* médio), e (2) um grupo consistindo de PW e LOCI, claramente separado de (1), com um desempenho muito inferior. Os resultados aqui apresentados não concordam com aqueles reportados em Janssens, Flesch e Postma (2009) em vários aspectos: primeiro, LOF e SVDD não estão empatados, segundo, já não existem dois grupos claramente separados, e terceiro, KNN_{local} é o pior em desempenho, ao invés de um dos melhores.

Figura 9 – *Ranking* médio dos métodos comparados em Janssens, Flesch e Postma (2009) em todos os experimentos Tipo I com relação a *Weighted* ROC AUC.

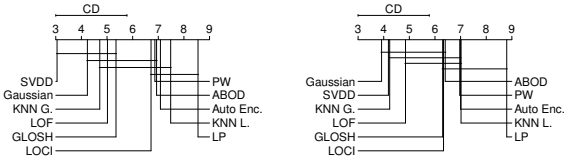


Em seguida, são reportados os resultados dos experimentos realizados seguindo a configuração descrita na seção 4.1.1. Os números detalhados para todos os experimentos são apresentados nas Tabelas 1 - 4. As Tabelas 1 e 2 exibem os valores de ROC AUC para os experimentos Tipo I e Tipo II, respectivamente, para cada método. As Tabelas 3 e 4 mostram os valores de AjustedPrec@n para os experimentos Tipo I e Tipo II, respectivamente, para cada método. Os maiores valores obtidos para cada base de dados são mostrados em negrito. Os resultados gerais dos *rankings* são sumarizados nas Figuras 10 e 11.

A Figura 10 mostra os *rankings* médios dos métodos em todos os experimentos Tipo I com relação a ROC AUC e AdjustedPrec@n. Ao olhar a ROC AUC, pode-se notar que SVDD, Gaussian, e kNN_{global}, nesta ordem, aparecem no topo. A média da ROC AUC para SVDD, Gaussian, e kNN_{global} foram de 0.8, 0.8 e 0.78, respectivamente. Ao olhar para AdjustedPrec@n o quadro geral é semelhante, mas AdjustedPrec@n diz que SVDD não é mais o melhor em performance, sendo agora superado por Gaussian. Isto sugere que os *scorings* produzidos por Gaussian, em média, possuem mais verdadeiros *outliers* com os melhores *scorings* do que o

SVDD, enquanto que para SVDD os *scorings* dos verdadeiros *outliers* tendem a ser maiores do que os dos *inliers*, em geral.

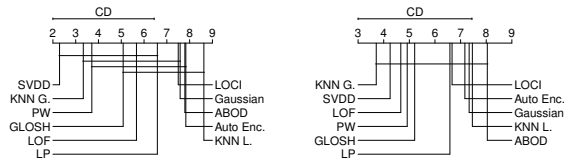
Figura 10 – *Ranking* médio dos métodos em todos os experimentos Tipo I.



(a) ROC AUC (b) AdjustedPrec@n

A Figura 11 mostra os *rankings* médios dos métodos em todos os experimentos Tipo II com relação a ROC AUC e AdjustedPrec@n. Ao comparar ROC AUC com AdjustedPrec@n, pode-se notar novamente algumas inversões nos *ranks*, por exemplo, entre LOF e PW/GLOSH. Em particular, o desempenho relativo do SVDD caiu mais uma vez para AdjustedPrec@n, assim como também caiu nos experimentos Tipo I, mas agora SVDD é superado pelo KNN_{global}, e não por Gaussian. Quando comparados os experimentos Tipo I da Figura 10 e os experimentos Tipo II da Figura 11, uma notável diferença pode ser observada com relação ao Gaussian: enquanto este método estava entre os top 3 métodos nos experimentos Tipo I, o seu desempenho cai drasticamente com relação aos experimentos Tipo II. O desempenho absoluto de Gaussian cai (dos experimentos Tipo I para o Tipo II) de 0.8 para 0.77 em ROC AUC e de 0.48 para 0.38 em AdjustedPrec@n, o que é esperado por Gaussian pressupor que um modelo de classe *inlier* unimodal melhor se encaixe nos dados, como disposto nos experimentos Tipo I. Mas só isso não explica totalmente a queda de Gaussian em termos de desempenho relativo. O que também explica é que outros métodos, particularmente métodos baseados em densidade local, como GLOSH e LOF, possuem um desempenho melhor nos experimentos Tipo II (veja as tabelas para valores detalhados), que correspondem a cenários de aplicação com possíveis classes alvo multi-modais.

Figura 11 – *Ranking* médio dos métodos em todos os experimentos Tipo II.



(a) ROC AUC (b) AdjustedPrec@n

Em geral, com base em ambos os tipos de experimentos, pode-se concluir que: (i) em conformidade com os resultados do estudo anterior em Janssens, Flesch e Postma (2009), conclui-se aqui também que SVDD é um *top performer*, particularmente com relação a ROC AUC; (ii) kNN_{global}, no entanto, pode ser uma opção preferível, uma vez que ele é consistentemente um *top performer*, ainda que muito mais simples que o SVDD; este método não foi incluído no

estudo de Janssens, Flesch e Postma (2009); e (iii) em contraste com Janssens, Flesch e Postma (2009), LOF não possui performance tão boa quanto o SVDD, e kNN$_{local}$ não está entre os *top performers*, mas, ao contrário, consistentemente entre os piores.

Tabela 1 – Primeiro tipo de experimentos — ROC AUC

ROC AUC	GLOSH	KNN L.	LP	ABOD	Auto E.	Gaussian	KNN G.	LOCI	LOF	PW	SVDD
Abalone	0.66	0.66	0.71	0.71	0.72	0.74	0.73	0.76	0.66	0.74	**0.77**
Aloi	0.98	0.97	0.95	**0.99**	0.98	**0.99**	**0.99**	0.98	0.98	0.98	0.98
Arrhythmia	0.63	0.58	0.5	0.51	0.53	0.55	0.52	0.53	0.6	0.5	**0.64**
Balance-Scale	0.88	0.86	0.88	0.86	0.91	**0.93**	0.87	0.87	0.92	0.87	0.91
Ball-Bearing	0.98	0.97	0.5	0.93	**1**	**1**	0.98	0.96	0.99	0.53	0.98
Biomed	**0.84**	0.81	0.51	0.65	0.72	0.8	**0.84**	0.79	0.71	0.69	0.7
Breast	0.96	0.93	0.81	0.93	0.91	**0.98**	0.96	**0.98**	0.96	0.8	**0.98**
Cancer	0.52	0.52	0.5	0.53	0.54	**0.59**	0.54	0.53	0.53	0.51	0.53
CellCycle237	0.81	0.72	0.74	0.81	0.76	0.82	**0.84**	0.72	0.81	0.74	0.83
Colon	0.67	0.63	0.5	0.64	0.58	0.67	0.66	0.59	**0.68**	0.5	**0.68**
Delft	0.95	**0.96**	0.93	0.68	0.83	0.95	0.93	0.89	**0.96**	0.93	**0.96**
Diabetes	0.65	0.63	0.51	0.61	0.62	0.64	0.62	0.63	**0.66**	0.59	0.65
Ecoli	**0.94**	0.93	0.92	0.93	0.89	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
Glass	0.78	0.78	0.81	0.79	0.76	0.78	0.81	0.67	0.81	0.81	**0.82**
Heart	0.6	0.57	0.5	0.59	0.67	**0.73**	0.58	0.58	0.59	0.55	0.61
Hepatitis	0.56	0.53	0.5	0.56	0.73	**0.74**	0.56	0.55	0.54	0.56	0.57
Housing	0.65	0.65	0.58	0.68	0.76	**0.78**	0.69	0.66	0.65	0.7	0.7
Imports	0.7	0.68	0.81	0.66	0.69	0.65	0.71	0.73	0.78	**0.83**	0.74
Ionosphere	**0.74**	0.66	0.66	0.64	0.62	0.64	0.67	0.6	0.64	0.64	**0.74**
Iris	0.97	0.95	**0.98**	0.97	0.96	**0.98**	0.97	0.97	0.97	**0.98**	**0.98**
Liver	0.54	0.55	0.53	0.55	0.54	0.54	0.55	**0.58**	0.55	0.53	0.56
Satellite	0.95	0.92	0.5	0.95	0.9	0.94	**0.96**	0.94	0.93	0.92	**0.96**
Sonar	0.7	0.73	0.76	0.64	0.67	0.66	0.76	0.65	**0.77**	0.76	0.75
Spectf	**0.66**	0.61	0.5	0.56	0.57	0.55	0.52	0.61	0.63	0.64	**0.66**
Survival	0.61	0.58	0.56	0.56	0.57	0.58	0.62	0.62	0.61	0.55	**0.65**
Vehicle	0.75	0.77	0.5	0.76	0.83	**0.9**	0.79	0.76	0.77	0.79	0.82
Vowels	0.99	0.98	**1**	0.99	0.63	0.99	**1**	0.91	0.99	**1**	0.99
Waveform	0.89	0.85	0.87	0.9	0.86	**0.91**	0.89	0.89	0.88	0.88	**0.91**
Wine	0.86	0.85	0.57	0.88	0.85	**0.96**	0.86	0.86	0.86	0.83	0.87
YeastGalactose	**0.99**	**0.99**	0.98	**0.99**	0.97	0.98	**0.99**	0.75	**0.99**	**0.99**	**0.99**
	0.78	0.76	0.69	0.75	0.75	0.8	0.78	0.75	0.78	0.74	0.8

Tabela 2 – Segundo tipo de experimentos — ROC AUC

ROC AUC	GLOSH	KNN L.	LP	ABOD	Auto E.	Gaussian	KNN G.	LOCI	LOF	PW	SVDD
Abalone	0.62	0.62	0.7	0.67	0.65	0.7	0.68	0.67	0.63	0.71	**0.73**
Aloi	**0.96**	0.94	0.95	0.92	0.94	0.92	**0.96**	0.92	**0.96**	**0.96**	**0.96**
Balance-Scale	0.82	0.8	0.79	0.79	0.77	0.78	0.83	0.81	**0.84**	0.82	0.81
CellCycle237	**0.81**	0.69	0.73	0.75	0.76	0.76	0.79	0.71	0.75	0.74	0.78
Glass	0.75	0.71	0.76	0.69	0.69	0.64	0.76	0.7	**0.77**	0.76	**0.77**
Iris	0.95	0.93	**0.97**	0.96	0.94	0.82	0.96	0.95	0.93	**0.97**	**0.97**
Satellite	**0.85**	0.79	0.5	0.74	0.77	0.74	0.84	0.77	0.82	0.82	**0.85**
Vehicle	0.68	0.63	0.5	0.61	0.66	0.71	0.74	0.7	0.67	0.72	**0.75**
Vowels	0.94	0.94	**0.98**	0.75	0.7	0.64	**0.98**	0.09	0.95	**0.98**	**0.98**
Waveform	0.81	0.69	0.75	0.75	0.62	0.76	0.81	0.83	0.75	0.77	**0.86**
Wine	0.74	0.73	0.58	0.76	0.84	**0.85**	0.75	0.71	0.74	0.76	0.77
YeastGalactose	0.95	0.91	**0.97**	0.93	0.93	0.91	**0.97**	0.93	0.95	**0.97**	**0.97**
	0.82	0.78	0.76	0.78	0.77	0.77	0.84	0.73	0.81	0.83	0.85

Tabela 3 – Primeiro tipo de experimentos — AdjustedPrec@n

A.Prec@n	GLOSH	KNN L.	LP	ABOD	Auto E.	Gaussian	KNN G.	LOCI	LOF	PW	SVDD
Abalone	0.21	0.17	0.3	0.32	0.29	0.32	0.33	0.37	0.2	0.31	**0.4**
Aloi	0.92	0.87	0.85	0.93	0.92	**0.95**	0.94	0.91	0.91	0.92	0.92
Arrhythmia	-0.19	0.14	-0.77	0.03	0.08	0.09	0.05	0.05	**0.17**	-0.77	-0.19
Balance-Scale	0.44	0.5	0.53	0.51	0.61	0.59	0.44	0.55	0.61	0.56	**0.64**
Ball-Bearing	0.84	0.78	-0.28	0.7	0.97	**0.98**	0.85	0.76	0.88	-0.2	0.84
Biomed	**0.57**	0.52	-0.54	0.31	0.4	0.54	**0.57**	0.45	0.37	0.06	0.08
Breast	0.84	0.75	0.43	0.75	0.72	0.87	0.84	**0.88**	0.83	0.38	0.87
Cancer	0.05	0.06	-0.32	0.06	0.03	0.08	**0.11**	-0.01	0.08	-0.32	-0.04
CellCycle237	0.41	0.32	0.4	0.45	0.42	0.51	0.5	0.34	0.44	0.45	**0.52**
Colon	0.2	0.16	-0.62	0.18	0.13	0.21	0.21	-0.17	0.2	-0.62	**0.27**
Delft	0.73	0.73	0.67	0.29	0.48	0.71	0.67	0.63	**0.77**	0.67	0.73
Diabetes	0.22	0.19	-0.52	0.18	0.16	0.22	0.18	0.21	**0.24**	0.11	0.23
Ecoli	**0.75**	0.7	0.66	0.72	0.59	0.73	**0.75**	0.72	0.74	0.74	0.74
Glass	0.4	0.38	0.47	0.45	0.37	0.41	0.48	0.24	0.46	**0.49**	**0.49**
Heart	0.16	0.11	-0.86	0.13	0.27	**0.34**	0.1	0.13	0.13	-0.23	0.17
Hepatitis	0.01	0.03	-0.28	0.04	0.24	**0.25**	0.01	0.02	0	-0.26	-0.02
Housing	0.11	0.13	0	0.14	0.2	**0.22**	0.18	0.16	0.13	0.14	0.16
Imports	0.35	0.28	0.45	0.21	0.31	0.22	0.36	0.38	0.41	**0.53**	0.32
Ionosphere	0.12	0.3	0.28	0.29	0.25	0.32	**0.39**	0.14	0.32	0.31	0.17
Iris	0.81	0.78	**0.88**	0.85	0.8	**0.88**	0.82	0.83	0.84	0.86	0.86
Liver	0.01	0.08	-0.62	0.06	0.05	0.06	0.05	**0.13**	0.07	-0.32	0.03
Satellite	0.73	0.59	-0.21	0.73	0.55	0.71	0.75	0.71	0.67	0.71	**0.77**
Sonar	0.26	0.34	0.38	0.18	0.23	0.21	0.39	0.23	**0.41**	0.39	0.3
Spectf	0.04	0.08	-0.26	0.06	0.05	0.08	0.07	**0.12**	0.11	0.04	0.05
Survival	0.15	0.12	0.04	0.1	0.11	0.13	0.19	0.17	0.17	0.06	**0.25**
Vehicle	0.35	0.37	-0.33	0.35	0.46	**0.62**	0.44	0.35	0.38	0.43	0.47
Vowels	0.87	0.83	**0.94**	0.83	0.84	0.82	**0.94**	0.82	0.86	**0.94**	**0.94**
Waveform	0.56	0.47	0.51	0.61	0.5	0.61	0.56	0.61	0.53	0.53	**0.62**
Wine	0.5	0.52	-0.28	0.56	0.49	**0.8**	0.51	0.52	0.51	0.47	0.53
YeastGalactose	**0.96**	0.9	0.88	0.94	0.88	0.9	**0.96**	0.44	0.91	0.92	0.95
	0.41	0.41	0.09	0.4	0.41	0.48	0.45	0.39	0.44	0.28	0.44

Tabela 4 – Segundo tipo de experimentos — AdjustedPrec@n

A.Prec@n	GLOSH	KNN L.	LP	ABOD	Auto E.	Gaussian	KNN G.	LOCI	LOF	PW	SVDD
Abalone	0.12	0.2	0.22	0.21	0.18	**0.3**	0.23	0.26	0.21	**0.3**	0.16
Aloi	**0.81**	0.74	0.67	0.68	0.77	0.74	**0.81**	0.69	0.8	0.78	**0.81**
Balance-Scale	0.43	0.48	0.54	0.43	0.45	0.44	0.43	0.52	**0.59**	0.54	0.56
CellCycle237	**0.31**	0.18	0.04	0.22	0.24	0.23	0.28	0.17	0.3	0.21	0.27
Glass	0.29	0.28	**0.34**	0.2	0.22	0.13	0.32	0.21	**0.34**	**0.34**	0.17
Iris	0.8	0.75	**0.86**	0.83	0.78	0.58	0.82	0.79	0.76	0.85	**0.86**
Satellite	**0.47**	0.44	-0.21	0.21	0.35	0.31	0.46	0.29	0.45	0.32	0.44
Vehicle	0.25	0.17	-0.33	0.14	0.2	0.25	**0.28**	0.23	0.23	0.16	**0.28**
Vowels	0.58	0.58	**0.77**	0.14	0.19	0.08	**0.77**	0.3	0.61	**0.77**	0.76
Waveform	0.45	0.29	0.38	0.35	0.14	0.34	0.45	0.48	0.36	0.4	**0.53**
Wine	0.33	0.38	-0.51	0.37	0.54	**0.56**	0.37	0.37	0.38	0.26	0.31
YeastGalactose	0.81	0.64	0.82	0.74	0.69	0.64	**0.85**	0.73	0.73	0.84	**0.85**
	0.47	0.43	0.3	0.38	0.4	0.38	0.51	0.42	0.48	0.48	0.5

4.2 Hipótese II

Avaliações de rankings/scorings de soluções de detecção de outliers podem ser diretamente realizadas sem a necessidade da binarização da solução em um problema top-n.

Uma das limitações do índice IREOS é a sua restrição em avaliar diretamente apenas soluções binárias. Uma possível abordagem para a avaliação de soluções dadas em forma de *rankings*, seria avaliar a separabilidade de todas as observações do *ranking*, ao invés de avaliar a separabilidade apenas das observações top-n. Para combinar os valores de separabilidade de cada observação pretende-se utilizar uma média ponderada pelo inverso da posição em que a observação foi ranqueada. Ao contrário da abordagem original de utilizar uma média simples para combinação das separabilidades das observações, a média ponderada pelo inverso da posição no *ranking* permite que a posição em que a observação foi ranqueada influencie no valor do índice, de tal forma que soluções com observações com maior grau de separabilidade ranqueadas mais ao topo do *ranking* possuirão uma avaliação melhor que soluções que possuírem observações com maior grau de separabilidade ranqueadas mais abaixo no *ranking*. O problema de tal abordagem é o custo computacional envolvido para o cálculo de separabilidade de todas as observações da base de dados. Para evitar este alto custo computacional pretende-se utilizar uma heurística, que parte da suposição que a maior parte das observações da base de dados são *inliers* e a separabilidade de tais observações não é de interesse, portanto, não precisam ser calculadas. O problema passa a ser definir quais observações necessitam ser avaliadas.

A abordagem investigada neste momento inicialmente calcula o valor esperado de separabilidade de uma observação utilizando simulações de Monte Carlo conforme discutido em Marques *et al.* (2015). Em seguida é calculada uma estimativa de densidade local de cada observação utilizando um *kernel* Gaussiano. Para que não seja necessário o cálculo de separabilidade de todas as observações, o valor esperado é propagado com o seguinte procedimento: utilizando

a estimativa de densidade local computada, é escolhida a observação de menor densidade para calcular o seu valor de separabilidade, caso o valor de separabilidade seja maior que o valor esperado, é calculado o valor de separabilidade da próxima observação menos densa, até que seja encontrada uma observação com separabilidade menor igual ao valor esperado. Quando essa observação é encontrada, ela e todas as demais observações que ainda não tenham o valor de separabilidade computado, que necessariamente são ainda mais densas que ela, recebem como valor de separabilidade o valor esperado. Esse procedimento é utilizado sob a seguinte justificativa: se uma observação possui separabilidade abaixo do valor esperado, uma observação ainda mais densa dificilmente possui uma separabilidade acima do valor esperado. Isso evita calcular a separabilidade de observações que não deveriam ter importância para o cálculo do índice e, como todas essas assumirão o mesmo valor de separabilidade, não importará a ordem entre elas no *ranking*, não haverá penalização no índice dada qualquer inversão de posições entre elas.

Tendo o valor de separabilidade de todas as observações, adquirido a partir do classificador de máxima margem ou por meio da propagação do valor esperado, o cálculo do índice é feito utilizando a mesma Equação (3.3), entretanto agora para o cálculo $\bar{p}(\gamma)$ é utilizada a média ponderada pelo inverso da posição da observação no *ranking* ao invés da média simples:

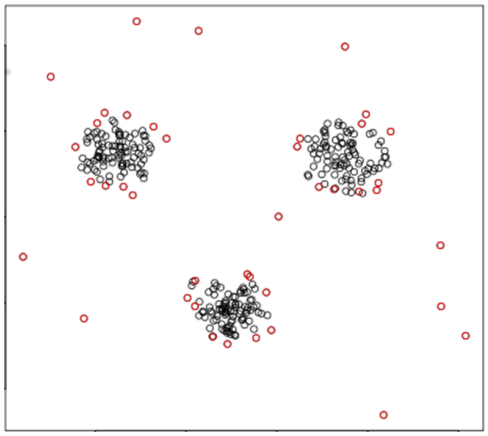
$$\bar{p}(\gamma) = \frac{\sum_{\mathbf{x}_i \in \mathbf{X}} inv\_rank(\mathbf{x}_i) p(\mathbf{x}_i, \gamma)}{\sum_{\mathbf{x}_i \in \mathbf{X}} inv\_rank(\mathbf{x}_i)} \tag{4.1}$$

onde $inv\_rank(\mathbf{x}_i)$ é o inverso da posição da observação $\mathbf{x}_i$ no *ranking* a ser avaliado, i.e., $inv\_rank(\mathbf{x}_i) = N - rank(\mathbf{x}_i)$, sendo N o número de observações na base de dados.

A estratégia apresentada acima é ilustrada na Figura 12. Inicialmente é calculado o valor esperado de separabilidade e, em seguida, a estimativa de densidade local de cada observação. Seguindo a estratégia apresentada acima, as observações são ordenadas a partir dos seus valores de estimativa de densidade local e suas separabilidades são avaliadas e comparadas com o valor esperado de separabilidade. No momento em que o valor de separabilidade for inferior ao valor esperado de separabilidade, todas as demais observações não avaliadas, que necessariamente possuem maior densidade, recebem como valor de separabilidade o valor esperado, ao invés de serem avaliadas pelo IREOS. Na Figura 12 as observações em vermelho serão as únicas a serem avaliadas pelo IREOS utilizando esta estratégia, as demais, ao invés de serem avaliadas por IREOS, receberam como avaliação de separabilidade o valor esperado. Por fim, a partir dos diferentes *rankings* de soluções a serem avaliadas, a Equação (4.1) é utilizada para computar o valor do índice para cada solução a ser avaliada.

A modelagem de *clumps* é um ponto ainda a ser investigado, originalmente todas as observações rotuladas com *outliers* recebem uma penalização menor por violarem margem em relação às observações rotuladas como *inliers*, conforme discutido na seção 3.2.1. Aqui, como não existe a distinção entre *inlier* e *outlier*, mas apenas o grau com que cada uma se categoriza como tal, pretende-se penalizar as observações que violam a margem de forma proporcional ao

Figura 12 – Exemplificação da estratégia de avaliação de *rankings* usando IREOS, em vermelho as observações que serão avaliadas.



grau com que elas se caracterizam como tal segundo o método de detecção.

4.3 Hipótese III

Avaliações de soluções de detecção de outliers em diferentes subespaços dos dados podem ser comparadas tornando a separabilidade das observações em seus diferentes subespaços comensuráveis utilizando procedimentos similares aos utilizados por algoritmos de detecção em subespaços (KRIEGEL et al., 2009b) e ensembles (LAZAREVIC; KUMAR, 2005) para garantir comensurabilidade de seus scorings.

As soluções geradas por algoritmos de detecção de *outliers* em subespaços é outra categoria de soluções que não pode ser avaliada pelo IREOS. Para avaliar a qualidade das soluções produzidas por esses algoritmos, a separabilidade das observações deve ser medida nos respectivos subespaços onde tais observações foram avaliadas como *outliers*. Em sua forma atual, entretanto, IREOS não pode ser aplicado diretamente pois sua medida de separabilidade de observações não é comensurável em diferentes subespaços. Para tal cenário, a medida deverá ser modificada. Certas questões, como a comensurabilidade do índice IREOS em diferentes subespaços, deverão ser investigadas e solucionadas, pois a separabilidade entre as observações que estão em diferentes subespaços não podem ser diretamente comparadas. As abordagens utilizadas pelas diferentes técnicas de detecção de *outliers* em subespaços para a comensurabilidade dos *scorings* produzidos nos diferentes subespaços são diversas (ZIMEK; SCHUBERT; KRIEGEL, 2012; NGUYEN; GOPALKRISHNAN; ASSENT, 2011; MüLLER; SCHIFFER; SEIDL, 2010; KRIEGEL *et al.*, 2009b; MüLLER *et al.*, 2008). O estudo dessas diferentes abordagens, já iniciada de forma muito breve e apresentada na seção 2.1.6, assim como a aplicação ou adaptação dessas para a comensurabilidade do índice IREOS em diferentes subespaços, compreende a abordagem inicial que será utilizada para a extensão do índice IREOS para avaliar soluções de algoritmos de detecção de *outliers* em subespaços.

4.4 Hipótese IV

Procedimentos similares aos utilizados em agrupamento de dados para determinar automaticamente o números de grupos através da aplicação de índices internos de validação (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010) podem ser adaptados permitindo a utilização do IREOS para determinar automaticamente o número de outliers (n) presente nos dados.

Índices internos de validação têm sido utilizados em agrupamento de dados para estimar o número de grupos existentes nos dados. Tais índices podem ser monotônicos em função do número de grupos, caso em que se pode tentar identificar um chamado "joelho" ou "cotovelo", que é uma queda/aumento abrupto do índice e que pode indicar o número natural de grupos existentes no dados (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010). Conjectura-se que procedimentos similares possam ser desenvolvidos utilizando IREOS como base para estimar o número de *outliers* presente nos dados.

Utilizando um procedimento análogo ao utilizado em agrupamento de dados para a determinação do número de grupos, IREOS foi aplicado preliminarmente para a determinação automática do número de *outliers* em uma base de dados. Em agrupamento de dados, para procurar pelos chamados "joelho"/"cotovelo", os índices internos são aplicados sistematicamente para avaliar múltiplas soluções com diferentes números de grupos (c), aumentando iterativamente tal número de 2 até um determinado número máximo (c_{max}) (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010). Como esses critérios são monotônicos com o aumento do número de grupos, procura-se encontrar um contraste evidente na qualidade de soluções com quantidades adjacentes de grupos, que pode sugerir o número adequado de grupos. A busca pelo número ideal de partições em uma determinada base de dados geralmente se dá comparando o aumento de qualidade, de acordo com o critério interno utilizado, da partição com $c-1$ grupos para a partição com c grupos em relação ao aumento de qualidade da partição com c grupos para a partição com $c+1$ grupos (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010).

Similarmente, o mesmo procedimento foi aplicado utilizando IREOS para tentar determinar o número de *outliers* em uma base de dados. Como IREOS avalia apenas soluções binárias, quando uma solução é dada em forma de *scorings*, primeiro deve-se rotular as n observações com maiores *scorings* como *outliers* e as demais como *inliers* (binarização da solução). Dada que uma determinada solução tenha ranqueado as observações de forma correta pelo algoritmo de detecção, espera-se que as observações que se caracterizem mais como *outliers* sejam ranqueadas no topo do *ranking* e conforme se desce no *ranking*, as observações passem a se caracterizar cada vez menos como *outlier*, então espera-se que exista um decréscimo na qualidade da solução avaliada pelo índice IREOS quando se aumenta o número de observações rotuladas como *outliers*. Variando o número de observações rotuladas como *outliers* de 1 até um determinado número n pode-se procurar por um decréscimo proeminente na qualidade da solução, que pode sugerir o

número ideal de *outliers*.

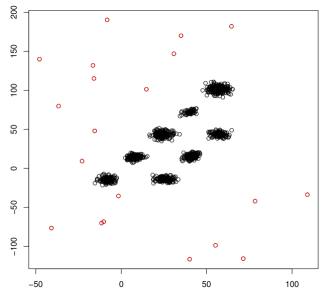
A busca pelo número ideal de *outliers* na base de dados pode ser feita comparando o decréscimo e acréscimo de qualidade da solução com n *outliers* quando comparado, respectivamente, com a solução com $n-1$ e $n+1$ *outliers*. A Figura 13b mostra a avaliação do índice IREOS (sem o ajuste estatístico por chance/aleatoriedade que é opcional ao índice) para uma mesma solução dada em forma de *scorings* com diferentes números n de observações rotuladas como *outliers*, variando n de 1 até 50. A base de dados utilizada é uma base sintética gerada a partir de misturas de Gaussianas e que possui 20 *outliers*, conforme ilustrada na Figura 13a. Pode-se claramente perceber que as 20 primeiras soluções possuem qualidade muito próxima devido às 20 primeiras observações se caracterizarem igualmente como *outlier*, a partir da 21^a observação inserida na solução, pelo fato dela não ser tão *outlier* quanto as demais, um decréscimo na qualidade é percebido.

Para facilitar a detecção desses "joelhos" pode-se transformar o seguinte gráfico em picos utilizando a Equação (4.2), onde o problema passa a ser encontrar o pico de maior altura, que corresponde idealmente ao número correto de *outliers*.

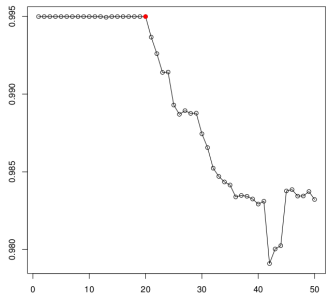
$$\underset{n}{\text{argmax}} \quad IREOS(n)((IREOS(n-1)-IREOS(n))-(IREOS(n)-IREOS(n+1)) \tag{4.2}$$

Devido às diferenças abruptas que podem acontecer em parte mais baixas do *rankings*, essas diferenças são ponderadas pelo valor de IREOS da solução, evitando que soluções ruins sejam selecionadas. O resultado da aplicação da Equação (4.2) nas avaliações feita pelo índice IREOS apresentadas Figura 13b pode ser visualizado na Figura 13c.

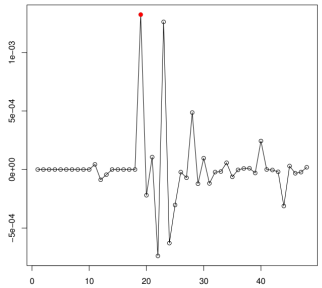
Figura 13 – Utilização de IREOS para determinar o número exato de *outlier*, em vermelho o número correto.



(a) Base de dados 2D utilizada na determinação automática do número de *outliers*



(b) Índice IREOS para uma mesma solução dada em *scorings* variando n de 1 até 50



(c) Índice IREOS da Figura 13(b) transformado em picos utilizando Equação (4.2)

Apesar dessa abordagem funcionar muito bem para bases de dados sintéticas que possuem *outliers* bem separados e tenham sido julgadas corretamente por um algoritmo de detecção, experimentos inicias mostraram que essa abordagem não é muito promissora/robusta em vários

cenários de aplicação menos controlados, i.e., cenários de aplicação real. Devido a isso, abordagens mais elaboradas, para avaliação de *rankings* de detecção de *outliers*, estão atualmente em andamento.

4.5 Hipótese V

> *Soluções com diferentes números de observações rotuladas como outliers podem ser comparadas por meio de testes estatísticos que assumam variância e tamanho da amostra diferentes (WELCH, 1947; MANN; WHITNEY, 1947), dado que a distribuição das avaliações das possíveis soluções segue uma distribuição Normal se certas suposições forem respeitadas.*

IREOS utiliza, como principal componente do índice de avaliação, uma soma das médias do grau de separabilidade das observações rotuladas como *outliers* para cada γ. Se respeitadas as suposições do Teorema do Limite Central, é garantido que a distribuição do índice seguirá, pelo menos aproximadamente, uma distribuição Normal. A comparação entre as médias de soluções com diferentes número de observações rotuladas como *outliers* não pode ser realizada diretamente porque esta média é monotonicamente decrescente com o número de *outliers*, se ordenadas corretamente de acordo com suas separabilidades. Conjectura-se, entretanto, como hipótese, que tal comparação poderia eventualmente ser feita através de testes estatísticos que assumem variância e tamanho da amostra (neste contexto, número de *outliers*) diferentes, tais como o teste paramétrico de Welch's t (WELCH, 1947) e o teste não paramétrico de Mann-Whitney (MANN; WHITNEY, 1947). Essa hipótese será investigada como uma primeira possível abordagem para estender IREOS para avaliação comparativa de soluções com diferentes números de *outliers*. A validação desta hipótese está relacionada à hipótese IV, uma vez que a comparação entre soluções com diferente número de observações rotuladas como *outliers* poderia auxiliar na determinação do número de *outliers* presente na base de dados, ainda que por procedimentos diferentes dos que se pretende investigar na hipótese IV.

4.6 Hipótese VI

> *O conceito de separabilidade utilizado por IREOS para avaliação das soluções pode ser adaptado com vistas ao desenvolvimento de um novo paradigma de detecção não supervisionada de outliers, particularmente no cenário de subespaços, utilizando técnicas de seleção de atributos, tais como forward selection e backward elimination (GUYON; ELISSEEFF, 2003), similarmente à abordagem utilizada por Micenkova et al. (2013) para pós-processamento da solução de um algoritmo de detecção de outliers.*

Outra possível questão que também se deseja investigar neste projeto são formas pelas quais alguns dos princípios utilizado pelo IREOS poderiam ser aplicados não para avaliar uma dada solução de detecção de *outliers*, mas sim para detectá-los. Embora buscar pela solução que otimize o índice IREOS tenha se mostrado um problema não trivial, sendo em princípio computacionalmente inviável otimizá-lo (o que é uma propriedade dos métodos genuínos de avaliação interna), o conceito básico de separabilidade aplicado pelo índice não é utilizado por nenhum algoritmo de detecção de *outliers* em particular, ainda que se encaixe em todas definições clássicas de *outlier*, e então, poderia ser adaptado objetivando novos métodos de detecção de *outliers*, não apenas no espaço completo mas também possivelmente em subespaços dos atributos. Pretende-se investigar a possibilidade de se adaptar esse conceito com vistas ao desenvolvimento de um novo paradigma de detecção não supervisionada de *outliers*. A hipótese básica a ser explorada inicialmente é que os *scorings* de *outliers* devem ser de alguma forma positivamente relacionados com o grau de separabilidade de cada observação. No caso particular de detecção em subespaços, pretende-se utilizar técnicas de seleção de atributos, tais como *forward selection* e *backward elimination* (GUYON; ELISSEEFF, 2003), como uma primeira possível abordagem para encontrar o subespaço em que cada observação possui o seu maior grau de separabilidade. Uma abordagem similar foi utilizada anteriormente por Micenkova *et al.* (2013); entretanto, Micenkova *et al.* (2013) aplicaram tal abordagem num contexto distinto, de pós-processamento, ou seja, após a detecção dos *outliers*, para encontrar o subespaço que poderia explicar o porquê de cada observação ser rotulada como *outlier*, e não para detectá-las.

4.7 Hipótese VII

> *IREOS pode ser utilizado para avaliar, segundo a ótica da separabilidade na qual o índice se baseia, rótulos externos de bases de dados de detecção de outliers, que em sua maioria são oriundas de outras áreas (CAMPOS et al., 2016; GOLDSTEIN; UCHIDA, 2016), para verificar se tais rotulações externas são coerentes com a disposição espacial do dados no espaço de atributos em que esses estão descritos.*

Um problema existente é que, ao contrário de outras áreas, em detecção de *outliers* não existem coleções de bases de dados *benchmark* originalmente projetadas para esta tarefa. Os poucos *benchmarks* disponíveis para avaliação em detecção não supervisionada de *outliers* (CAMPOS *et al.*, 2016; GOLDSTEIN; UCHIDA, 2016) utilizam tipicamente versões modificadas de bases de dados que são originalmente projetadas para outros problemas, por exemplo, para problemas de classificação e agrupamento de dados. Um procedimento comum aplicado a essas bases é a subamostragem de uma das classes para assegurar que as respectivas observações são pouco frequentes (LAZAREVIC; KUMAR, 2005; ZHANG; HUTTER; JIN, 2009; SCHUBERT *et al.*, 2012; MICENKOVA *et al.*, 2013). Esta classe subamostrada é então considerada a classe *outlier* simplesmente porque agora ela é rara, porém a semântica da classe original

não é considerada neste tipo de rotulação arbitrária. A realização deste procedimento de forma simplista e sem que se use um critério apropriado pode fazer com que não haja qualquer compatibilidade entre a semântica dos dados e as rotulações das observações/classes como *outliers* e *inliers*. Neste contexto, pretende-se investigar a hipótese que IREOS poderia avaliar, segundo a ótica da separabilidade, os rótulos produzidos por meio de tais procedimentos para verificar se tais rotulações são coerentes com a disposição espacial do dados naquele espaço de atributos, servindo assim como uma possível ferramenta para seleção de bases de dados *benchmark* na área de detecção não supervisionada de *outliers*. Para isto, a primeira abordagem adotada será a avaliação individual de cada observação rotulada como *outlier* no *ground truth*. Devido ao ajuste do índice para aleatoriedade, é esperado que observações poucas separáveis, isto é, que mais se caracterizam como *inliers* naquele espaço de atributos, recebam valores próximos de 0. Ainda, como uma abordagem complementar, as observações rotuladas como *outliers* no *ground truth* poderão ser avaliadas conjuntamente e a utilização do parâmetro opcional do índice para análise exploratória dos dados (m_{clSize}), que modela a possível presença de aglomerados de *outliers*, poderá ser utilizado para verificar a possível existência de aglomerados de *outliers*. Tais aglomerados podem ocorrer devido às observações rotuladas como *outliers* muitas vezes formarem originalmente uma classe/grupo de uma base de dados de classificação/agrupamento que foi simplesmente subamostrada para dar origem a uma base de dados de detecção de *outliers*. O procedimento de subamostragem, entretanto, não garante que as observações remanescentes da classe/grupo subamostrada se caracterizem como *outliers*. Ao invés, eles podem ainda se caracterizar como uma classe/grupo mais esparsa, o que pode ser avaliado por IREOS.

CAPÍTULO

5

PLANO DE TRABALHO

Neste Capítulo é apresentado o cronograma e o status de cada uma das atividades propostas inicialmente no projeto de pesquisa.

5.1 Cronograma Original do Projeto

O cronograma inicialmente proposto no projeto de pesquisa é apresentado na Tabela 5.

A) Disciplinas: O aluno deverá cumprir o número mínimo de créditos em disciplinas, necessários como parte dos requisitos para a obtenção do título de Doutor em Ciências de Computação e Matemática Computacional.

B) Revisão Bibliográfica: Uma revisão abrangente da literatura será realizada e atualizada ao longo do desenvolvimento da pesquisa a fim de complementar a revisão preliminar. Conceitos básicos e avançados de aprendizado não supervisionado, detecção de *outliers*, agrupamento de dados, probabilidade e estatística, dentre outros, que se inserem no âmbito do presente projeto de pesquisa, serão estudados.

C) Qualificação: Até o final do 18° mês o candidato deverá elaborar e apresentar, como parte dos requisitos para a obtenção do título de doutor, uma monografia com os desenvolvimentos parciais do seu trabalho, bem como defendê-la e aprová-la perante a uma banca de qualificação.

D) Proficiência: Também como parte do requisitos para obtenção do título, o candidato deverá realizar e ser aprovado em um teste de proficiência em língua inglesa até o final do 24° mês.

E) Estudo Conceitual dos Problemas e Proposta de Soluções: Os problemas de interesse relacionados à avaliação, seleção de modelos e algoritmos de detecção não supervisio-

nada de *outliers* serão investigados e possíveis soluções serão desenvolvidas seguindo a metodologia e almejando os objetivos descritos no projeto inicial de pesquisa.

F) Estudo Experimental: O desempenho das possíveis técnicas e procedimentos desenvolvidos serão avaliados em experimentos computacionais extensivos envolvendo bases de dados artificiais e reais.

G) Elaboração da Tese e Divulgação dos Resultados: O trabalho desenvolvido será documentado apropriadamente na forma de artigos científicos e da tese de doutorado, com defesa prevista para Fevereiro de 2018.

Etapas	Meses					
	1-6	7-12	13-18	19-24	25-30	31-36
A	•	•				
B	•	•	•	•	•	
C			•			
D				•		
E	•	•	•	•	•	
F		•	•	•	•	
G			•	•	•	•

Tabela 5 – Cronograma das atividades previstas

5.2 Atividades Desenvolvidas

Abaixo são listas as atividades realizadas referentes a cada item do cronograma inicial.

A) Disciplinas: Todas as disciplinas necessárias para a obtenção do número mínimo de créditos foram concluídas com conceito A.

B) Revisão Bibliográfica: Uma contínua revisão bibliográfica vem sendo realizada desde o início do projeto e pretende-se continuá-la até a defesa da tese.

C) Qualificação: A escrita e defesa da monografia de qualificação está sendo realizada dentro do cronograma previsto.

D) Proficiência: Antecipando o cronograma inicialmente estabelecido, o aluno já foi aprovado no exame de proficiência.

E) Estudo Conceitual dos Problemas e Proposta de Soluções: O estudo conceitual e uma proposta para utilização do IREOS para avaliação de soluções dadas em forma de *scorings* continuam em estudo.

F) Estudo Experimental: Alguns estudos preliminares foram feitos no sentido da utilização do índice IREOS para avaliação de soluções dadas em forma de *scorings* e determinar automaticamente o número de *outliers* presentes em uma base de dados. Além disso, um estudo comparativo entre métodos de classificação de classe única e métodos não supervisionados de detecção de *outliers* adaptados para o cenário supervisionado foi realizado, envolvendo diversos métodos de ambas as classes de algoritmos e uma grande coleção de base de dados.

G) Elaboração da Tese e Divulgação dos Resultados: Os resultados já começaram a ser divulgados por meio da publicação do artigo referente à hipótese I.

De forma geral, o cronograma tem sido realizado dentro do prazo inicialmente previsto, tendo algumas das atividade antecipadas, e.g. proficiência e divulgação dos resultados. Os passos seguintes do projeto deverão seguir o cronograma inicialmente proposto.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: **Proceedings of the 20th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 487–499. ISBN 1-55860-153-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645920.672836>. Citado na página 27.

ANGIULLI, F.; FASSETTI, F. Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 3, n. 1, p. 4:1–4:57, mar. 2009. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/1497577.1497581>. Citado na página 13.

ANGIULLI, F.; PIZZUTI, C. Fast outlier detection in high dimensional spaces. In: . Helsinki, Finland: [s.n.], 2002. p. 15–26. Citado 3 vezes nas páginas 11, 18 e 20.

____. Outlier mining in large high-dimensional data sets. **IEEE Transactions on Knowledge and Data Engineering**, p. 203–215, 2005. Citado 2 vezes nas páginas 11 e 13.

BARNETT, V.; LEWIS, T. **Outliers in Statistical Data**. 3rd. ed. [S.l.]: John Wiley & Sons, 1994. Citado 4 vezes nas páginas 10, 11, 18 e 42.

BAY, S. D.; SCHWABACHER, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: **Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2003. (KDD '03), p. 29–38. ISBN 1-58113-737-0. Disponível em: <http://doi.acm.org/10.1145/956750.956758>. Citado na página 13.

BECKMANN, N.; KRIEGEL, H.-P.; SCHNEIDER, R.; SEEGER, B. The r*-tree: An efficient and robust access method for points and rectangles. In: **Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 1990. (SIGMOD '90), p. 322–331. ISBN 0-89791-365-5. Disponível em: <http://doi.acm.org/10.1145/93597.98741>. Citado na página 15.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006. Hardcover. ISBN 0387310738. Citado na página 35.

BREUNIG, M. M.; KRIEGEL, H.-P.; NG, R.; SANDER, J. Optics-of: Identifying local outliers. In: . Prague, Czech Republic: [s.n.], 1999. p. 262–270. Citado na página 11.

____. Lof: Identifying density-based local outliers. In: . Dallas, TX: [s.n.], 2000. p. 93–104. Citado 4 vezes nas páginas 10, 11, 18 e 21.

CAM, L. L. Maximum likelihood: An introduction. **International Statistical Review / Revue Internationale de Statistique**, [Wiley, International Statistical Institute (ISI)], v. 58, n. 2, p. 153–171, 1990. ISSN 03067734, 17515823. Disponível em: <http://www.jstor.org/stable/1403464>. Citado na página 28.

CAMPELLO, R. J. G. B.; MOULAVI, D.; ZIMEK, A.; SANDER, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 10, n. 1, p. 5:1–5:51, jul. 2015. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/2733381>. Citado 2 vezes nas páginas 18 e 23.

____. ____. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 10, n. 1, p. 5:1–5:51, jul. 2015. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/2733381>. Citado na página 43.

CAMPOS, G. O.; ZIMEK, A.; SANDER, J.; CAMPELLO, R. J. G. B.; MICENKOVÁ, B.; SCHUBERT, E.; ASSENT, I.; HOULE, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. **Data Mining and Knowledge Discovery**, v. 30, n. 4, p. 891–927, 2016. ISSN 1573-756X. Disponível em: <http://dx.doi.org/10.1007/s10618-015-0444-8>. Citado 6 vezes nas páginas 15, 31, 34, 42, 46 e 57.

CHAN, P. K.; STOLFO, S. J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In: **In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining**. [S.l.]: AAAI Press, 1998. p. 164–168. Citado na página 31.

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://doi.acm.org/10.1145/1541880.1541882>. Citado na página 9.

____. Anomaly detection for discrete sequences: A survey. **IEEE Trans. on Knowl. and Data Eng.**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 24, n. 5, p. 823–839, maio 2012. ISSN 1041-4347. Disponível em: <http://dx.doi.org/10.1109/TKDE.2010.235>. Citado na página 9.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Editorial: Special issue on learning from imbalanced data sets. **ACM SIGKDD Explorations Newsletter**, p. 1–6, 2004. Citado na página 31.

DEAN, J.; GHEMAWAT, S. Mapreduce: Simplified data processing on large clusters. **Commun. ACM**, ACM, New York, NY, USA, v. 51, n. 1, p. 107–113, jan. 2008. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/1327452.1327492>. Citado na página 15.

DOMINGOS, P. Metacost: A general method for making classifiers cost-sensitive. In: **Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 1999. (KDD '99), p. 155–164. ISBN 1-58113-143-7. Disponível em: <http://doi.acm.org/10.1145/312129.312220>. Citado na página 31.

DRUMMOND, C.; HOLTE, R. C. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In: . [S.l.: s.n.], 2003. p. 1–8. Citado na página 31.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2nd. ed. [S.l.]: Wiley, 2001. Citado na página 30.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: . [S.l.]: AAAI Press, 1996. p. 226–231. Citado na página 23.

FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery and Data Mining**. [S.l.]: MIT Press, 1996. Citado na página 9.

GAN, G.; MA, C.; WU, J. **Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability)**. illustrated edition. [S.l.]: SIAM, Society for Industrial and Applied Mathematics, 2007. Paperback. ISBN 0898716233. Citado na página 12.

GEUSEBROEK, J.-M.; BURGHOUTS, G. J.; SMEULDERS, A. W. The amsterdam library of object images. **International Journal of Computer Vision**, v. 61, n. 1, p. 103–112, 2005. ISSN 1573-1405. Disponível em: <http://dx.doi.org/10.1023/B:VISI.0000042993.50813.60>. Citado na página 45.

GHOTING, A.; PARTHASARATHY, S.; OTEY, M. E. Fast mining of distance-based outliers in high-dimensional datasets. **Data Min. Knowl. Discov.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 16, n. 3, p. 349–364, jun. 2008. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-008-0093-2>. Citado na página 13.

GOLDSTEIN, M.; UCHIDA, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. **PLoS ONE**, Public Library of Science, v. 11, n. 4, p. 1–31, 04 2016. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0152173>. Citado 2 vezes nas páginas 15 e 57.

GRUBBS, F. E. Procedures for detecting outlying observations in samples. **Technometrics**, v. 11, n. 1, p. 1–21, 1969. Citado 2 vezes nas páginas 10 e 18.

GUTTMAN, A. R-trees: A dynamic index structure for spatial searching. In: **Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 1984. (SIGMOD '84), p. 47–57. ISBN 0-89791-128-8. Disponível em: <http://doi.acm.org/10.1145/602259.602266>. Citado na página 15.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 1157–1182, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944968>. Citado 3 vezes nas páginas 15, 56 e 57.

HADI, A. S. Identifying multiple outliers in multivariate data. **Journal of the Royal Statistical Society. Series B (Methodological)**, Wiley for the Royal Statistical Society, v. 54, n. 3, p. pp. 761–771, 1992. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2345856>. Citado na página 19.

HADI, A. S.; IMON, A. H. M. R.; WERNER, M. Detection of outliers. **Wiley Interdisciplinary Reviews: Computational Statistics**, John Wiley & Sons, Inc., v. 1, n. 1, p. 57–70, 2009. ISSN 1939-0068. Disponível em: <http://dx.doi.org/10.1002/wics.6>. Citado na página 9.

HAIR, J. F.; BLACK, B.; BABIN, B.; ANDERSON, R. E.; TATHAM, R. L. **Multivariate Data Analysis**. 6th. ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2005. ISBN 0-02-349020-9. Citado na página 19.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2nd. ed. [S.l.]: Morgan Kaufmann, 2006. Citado 2 vezes nas páginas 9 e 10.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. [S.l.]: Morgan Kaufmann, 2011. Citado 2 vezes nas páginas 30 e 31.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. [S.l.]: Springer-Verlag New York, Inc., 2013. Citado 2 vezes nas páginas 13 e 35.

HAWKINS, D. **Identification of Outliers**. [S.l.]: Chapman and Hall, 1980. Citado 3 vezes nas páginas 10, 11 e 18.

HE, Z.; XU, X.; DENG, S. Discovering cluster-based local outliers. **Pattern Recognition Letters**, v. 24, n. 9–10, p. 1641 – 1650, 2003. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865503000035>. Citado na página 18.

HINNEBURG, A.; HINNEBURG, E.; KEIM, D. A. An efficient approach to clustering in large multimedia databases with noise. In: . [S.l.]: AAAI Press, 1998. p. 58–65. Citado na página 23.

HODGE, V.; AUSTIN, J. A survey of outlier detection methodologies. **Artif. Intell. Rev.**, Kluwer Academic Publishers, Norwell, MA, USA, v. 22, n. 2, p. 85–126, out. 2004. ISSN 0269-2821. Disponível em: <http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9>. Citado na página 9.

HORTA, D.; CAMPELLO, R. J. G. B. Automatic aspect discrimination in data clustering. **Pattern Recognition**, v. 45, n. 12, p. 4370 – 4388, 2012. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320312002415>. Citado na página 45.

HOULE, M. E.; KRIEGEL, H.-P.; KRöGER, P.; SCHUBERT, E.; ZIMEK, A. Can shared-neighbor distances defeat the curse of dimensionality? In: **Proceedings of the 22Nd International Conference on Scientific and Statistical Database Management**. Berlin, Heidelberg: Springer-Verlag, 2010. (SSDBM'10), p. 482–500. ISBN 3-642-13817-9, 978-3-642-13817-1. Disponível em: <http://dl.acm.org.ez67.periodicos.capes.gov.br/citation.cfm?id=1876037.1876078>. Citado na página 25.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v. 2, n. 1, p. 193–218, December 1985. Citado 2 vezes nas páginas 34 e 39.

IMAN, R.; DAVENPORT, J. Approximations of the critical region of the friedman statistic. In: ____. [S.l.: s.n.], 1979. Citado na página 47.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X. Citado 2 vezes nas páginas 12 e 13.

JANSSENS, J. H. M.; FLESCH, I.; POSTMA, E. O. Outlier detection with one-class classifiers from ml and kdd. In: **Machine Learning and Applications, 2009. ICMLA '09. International Conference on**. [S.l.: s.n.], 2009. p. 147–153. Citado 7 vezes nas páginas 14, 41, 42, 46, 47, 48 e 49.

JANSSENS, J. H. M.; POSTMA, E. O. One-class classification with lof and loci: An empirical comparison. In: **Proceedings of the 18th Annual Belgian-Dutch on Machine Learning**. [S.l.: s.n.], 2009. p. 56–64. Citado na página 42.

JIN, W.; TUNG, A. K. H.; HAN, J. Mining top-n local outliers in large databases. In: **Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2001. (KDD '01), p. 293–298. ISBN 1-58113-391-X. Disponível em: <http://doi.acm.org/10.1145/502512.502554>. Citado na página 13.

JIN, W.; TUNG, A. K. H.; HAN, J.; WANG, W. Ranking outliers using symetric neighborhood relationship. In: . Singapore: [s.n.], 2006. p. 577–593. Citado 3 vezes nas páginas 11, 18 e 21.

KHAN, S. S.; MADDEN, M. G. A survey of recent trends in one class classification. In: **Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science**. Berlin, Heidelberg: Springer-Verlag, 2010. (AICS'09), p. 188–197. ISBN 3-642-17079-X, 978-3-642-17079-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1939047.1939070>. Citado na página 27.

KNORR, E. M.; NG, R. T. Algorithms for mining distance-based outliers in large datasets. In: **Proceedings of the 24rd International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. (VLDB '98), p. 392–403. ISBN 1-55860-566-5. Disponível em: <http://dl.acm.org/citation.cfm?id=645924.671334>. Citado 2 vezes nas páginas 10 e 11.

____. Finding intensional knowledge of distance-based outliers. In: **Proceedings of the 25th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (VLDB '99), p. 211–222. ISBN 1-55860-615-7. Disponível em: <http://dl.acm.org/citation.cfm?id=645925.671529>. Citado na página 25.

KNORR, E. M.; NG, R. T.; TUCANOV, V. Distance-based outliers: Algorithms and applications. **The VLDB Journal**, p. 8(3–4):237–253, 2000. Citado na página 11.

KRIEGEL, H.-P.; KRöGER, P.; SCHUBERT, E.; ZIMEK, A. Loop: local outlier probabilities. In: . Hong Kong, China: [s.n.], 2009. p. 1649–1652. Citado 2 vezes nas páginas 11 e 21.

____. Outlier detection in axis-parallel subspaces of high dimensional data. In: . Bangkok, Thailand: [s.n.], 2009. p. 831–838. Citado 5 vezes nas páginas 10, 14, 18, 25 e 53.

KRIEGEL, H.-P.; SCHUBERT, M.; ZIMEK, A. Angle-based outlier detection in high-dimensional data. In: . Las Vegas, NV: [s.n.], 2008. p. 444–452. Citado 3 vezes nas páginas 11, 18 e 24.

LAZAREVIC, A.; KUMAR, V. Feature bagging for outlier detection. In: . Chicago, IL: [s.n.], 2005. p. 157–166. Citado 3 vezes nas páginas 14, 53 e 57.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>. Citado na página 45.

MAHALANOBIS, P. C. On the generalized distance in statistics. In: **Proceedings of the National Institute of Sciences of India**. [S.l.: s.n.], 1936. p. 49–55. Citado na página 19.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 18, n. 1, p. 50–60, 03 1947. Disponível em: <http://dx.doi.org/10.1214/aoms/1177730491>. Citado 2 vezes nas páginas 14 e 56.

MARKOU, M.; SINGH, S. Novelty detection: A review part 1: Statistical approaches. **Signal Process.**, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 83, n. 12, p. 2481–2497, dez. 2003. ISSN 0165-1684. Disponível em: <http://dx.doi.org/10.1016/j.sigpro.2003.07.018>. Citado na página 42.

MARQUES, H. O. **Avaliação e Seleção de Modelos em Detecção Não Supervisionada de Outliers**. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos - SP, Brazil, 2015. Citado 2 vezes nas páginas 12 e 35.

MARQUES, H. O.; CAMPELLO, R. J. G. B.; ZIMEK, A.; SANDER, J. On the internal evaluation of unsupervised outlier detection. In: **Proceedings of the 27th International Conference on Scientific and Statistical Database Management**. New York, NY, USA: ACM, 2015. (SSDBM '15), p. 7:1–7:12. ISBN 978-1-4503-3709-0. Disponível em: <http://doi.acm.org/10.1145/2791347.2791352>. Citado 4 vezes nas páginas 12, 13, 35 e 51.

MICENKOVA, B.; DANG, X.-H.; ASSENT, I.; NG, R. Explaining outliers by subspace separability. In: **Data Mining (ICDM), 2013 IEEE 13th International Conference on**. [S.l.: s.n.], 2013. p. 518–527. ISSN 1550-4786. Citado 3 vezes nas páginas 15, 56 e 57.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, v. 50, p. 159–179, 1985. Citado na página 13.

MüLLER, E.; ASSENT, I.; STEINHAUSEN, U.; SEIDL, T. Outrank: ranking outliers in high dimensional data. In: **Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on**. [S.l.: s.n.], 2008. p. 600–603. Citado na página 53.

MüLLER, E.; SCHIFFER, M.; SEIDL, T. Adaptive outlierness for subspace outlier ranking. In: **Proceedings of the 19th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2010. (CIKM '10), p. 1629–1632. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871690>. Citado na página 53.

NEMENYI, P. **Distribution-free multiple comparisons**. Tese (Doutorado) — Princeton, 1963. Citado na página 47.

NGUYEN, H. V.; GOPALKRISHNAN, V.; ASSENT, I. An unbiased distance-based outlier detection approach for high-dimensional data. In: **Proceedings of the 16th International Conference on Database Systems for Advanced Applications - Volume Part I**. Berlin, Heidelberg: Springer-Verlag, 2011. (DASFAA'11), p. 138–152. ISBN 978-3-642-20148-6. Disponível em: <http://dl.acm.org/citation.cfm?id=1997305.1997322>. Citado na página 53.

ORAIR, G. H.; TEIXEIRA, C.; WANG, Y.; JR., W. M.; PARTHASARATHY, S. Distance-based outlier detection: Consolidation and renewed bearing. **Proceedings of the VLDB Endowment**, p. 3(2):1469–1480, 2010. Citado na página 11.

OSUNA, E.; FREUND, R.; GIROSI, F. **Support Vector Machines: Training and Applications**. Cambridge, MA, USA, 1997. Citado na página 37.

PAPADIMITRIOU, S.; KITAGAWA, H.; GIBBONS, P.; FALOUTSOS, C. Loci: Fast outlier detection using the local correlation integral. In: . Bangalore, India: [s.n.], 2003. p. 315–326. Citado 3 vezes nas páginas 11, 21 e 22.

PARZEN, E. On estimation of a probability density function and mode. **The annals of mathematical statistics**, JSTOR, v. 33, n. 3, p. 1065–1076, 1962. Citado na página 28.

PATCHA, A.; PARK, J.-M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. **Comput. Netw.**, Elsevier North-Holland, Inc., New York, NY, USA, v. 51, n. 12, p. 3448–3470, ago. 2007. ISSN 1389-1286. Disponível em: <http://dx.doi.org/10.1016/j.comnet.2007.02.001>. Citado na página 9.

PEKALSKA, E.; TAX, D. M. J.; DUIN, R. P. W. One-class LP classifier for dissimilarity representations. In: BECKER, S.; THRUN, S.; OBERMAYER, K. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: MIT Press: Cambridge, MA, 2003. v. 15. Citado na página 29.

PHAM, N.; PAGH, R. A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In: **Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2012. (KDD '12), p. 877–885. ISBN 978-1-4503-1462-6. Disponível em: <http://doi.acm.org/10.1145/2339530.2339669>. Citado na página 18.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. In: . Dallas, TX: [s.n.], 2000. p. 427–438. Citado 4 vezes nas páginas 10, 11, 18 e 20.

RIDDER, D. de; TAX, D. M.; DUIN, R. P. W. An experimental comparison of one-class classification methods. In: ROMENY, B. T. H.; EPEMA, D.; TONINO, J.; WOLTERS, A. (Ed.). **Proc. 4th Annual Conference of the Advanced School for Computing and Imaging (ASCI'98)**. Delft, The Netherlands: ASCI, 1998. p. 213–218. Citado na página 29.

ROUSSEEUW, P. J.; ZOMEREN, B. C. v. Unmasking multivariate outliers and leverage points. **Journal of the American Statistical Association**, American Statistical Association, v. 85, n. 411, p. pp. 633–639, 1990. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2289995>. Citado na página 19.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. Cambridge, MA, USA: MIT Press, 2001. ISBN 0262194759. Citado na página 29.

SCHUBERT, E.; WOJDANOWSKI, R.; ZIMEK, A.; KRIEGEL, H.-P. On evaluation of outlier rankings and outlier scores. In: . Anaheim, CA: [s.n.], 2012. p. 1047–1058. Citado 2 vezes nas páginas 12 e 57.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison Wesley, 2006. Citado 4 vezes nas páginas 9, 11, 30 e 31.

TANG, J.; CHEN, Z.; FU, A. W.-C.; CHEUNG, D. W. Enhancing effectiveness of outlier detections for low density patterns. In: . Taipei, Taiwan: [s.n.], 2002. p. 535–548. Citado 2 vezes nas páginas 11 e 21.

TAX, D. M. J. **One-class classification**. [S.l.]: TU Delft, Delft University of Technology, 2001. Citado 2 vezes nas páginas 28 e 30.

TAX, D. M. J. **One-class classification: Concept-learning in the absence of counter-examples**. Tese (Doutorado) — University of Delft, 2001. Citado na página 27.

____. **DDtools, the Data Description Toolbox for Matlab**. 2015. Version 2.1.2. Citado na página 44.

TAX, D. M. J.; DUIN, R. P. W. Support vector data description. **Machine learning**, Springer, v. 54, n. 1, p. 45–66, 2004. Citado na página 29.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining**, v. 3, n. 4, p. 209–235, 2010. Citado 3 vezes nas páginas 13, 14 e 54.

VLADIMIR, V. N.; VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer Heidelberg, 1995. Citado na página 29.

WELCH, B. L. The generalization of 'student's' problem when several different population varlances are involved. **Biometrika**, v. 34, n. 1-2, p. 28–35, 1947. Disponível em: <http://biomet.oxfordjournals.org/content/34/1-2/28.short>. Citado 2 vezes nas páginas 14 e 56.

YEUNG, K.; MEDVEDOVIC, M.; BUMGARNER, R. Clustering gene-expression data with repeated measurements. **Genome Biology**, v. 4, 2003. Citado na página 45.

YEUNG, K. Y.; FRALEY, C.; MURUA, A.; RAFTERY, A. E.; RUZZO, W. L. Model-based clustering and data transformations for gene expression data. **Bioinformatics**, v. 17, n. 10, p. 977–987, 2001. Disponível em: <http://bioinformatics.oxfordjournals.org/content/17/10/977.abstract>. Citado na página 45.

ZADROZNY, B.; ELKAN, C. Learning and making decisions when costs and probabilities are both unknown. In: **Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2001. (KDD '01), p. 204–213. ISBN 1-58113-391-X. Disponível em: <http://doi.acm.org/10.1145/502512.502540>. Citado na página 31.

ZADROZNY, B.; LANGFORD, J.; ABE, N. Cost-sensitive learning by cost-proportionate example weighting. In: **Data Mining, 2003. ICDM 2003. Third IEEE International Conference on**. [S.l.: s.n.], 2003. p. 435–442. Citado na página 31.

ZHANG, J.; LOU, M.; LING, T. W.; WANG, H. Hos-miner: A system for detecting outlyting subspaces of high-dimensional data. In: **Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30**. VLDB Endowment, 2004. (VLDB '04), p. 1265–1268. ISBN 0-12-088469-0. Disponível em: <http://dl.acm.org/citation.cfm?id=1316689.1316810>. Citado 2 vezes nas páginas 18 e 26.

ZHANG, K.; HUTTER, M.; JIN, H. A new local distance-based outlier detection approach for scattered real-world data. In: . Bangkok, Thailand: [s.n.], 2009. p. 813–822. Citado 4 vezes nas páginas 11, 18, 21 e 57.

ZIMEK, A.; CAMPELLO, R. J. G. B.; SANDER, J. Ensembles for unsupervised outlier detection: Challenges and research questions. **ACM SIGKDD Explorations**, v. 15, p. 11–22, 2013. Citado na página 14.

ZIMEK, A.; SCHUBERT, E.; KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. **Stat. Anal. Data Min.**, John Wiley & Sons, Inc., New York, NY, USA, v. 5, n. 5, p. 363–387, out. 2012. ISSN 1932-1864. Disponível em: <http://dx.doi.org/10.1002/sam.11161>. Citado 3 vezes nas páginas 9, 10 e 53.