Universidade de São Paulo - USP
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação

BEPE Research Project

Evaluation, Model Selection and Unsupervised Outlier Detection in Subspaces

FAPESP Process #2015/06019-0
Evaluation, Model Selection and Unsupervised Outlier Detection in Data Spaces and Subspaces
Avaliação, Seleção de Modelos e Detecção Não Supervisionada de Outliers em Espaços e Subespaços de Dados

Henrique Oliveira Marques
Student

Prof. Dr. José Fernando Rodrigues Júnior
Advisor in Charge

Prof. Dr. Jörg Sander
Host - University of Alberta

São Carlos, March 2017

Abstract

Although there is a growing literature that tackles the unsupervised outlier detection problem, the unsupervised evaluation of outlier detection results has been notably overlooked. In contrast to the unsupervised cluster analysis, where indexes for internal evaluation and validation of clustering solutions have been conceived and shown to be very useful, in the outlier detection domain only recently an index for internal evaluation of top-n (binary) outlier detection results, called IREOS, was firstly proposed by the candidate during his Masters. The index, based on the separability given by a maximum margin classifier (e.g. SVM and KLR), can evaluate and compare different candidate labelings based solely on the data and the assessed solutions themselves. And, consequently, allowing the selection of the most promising solutions corresponding to the more suitable models (algorithms, parameters). In this project, we propose a step towards bridging the gap related to the internal evaluation of subspace outlier detection results. Moreover, we also intend to use some of the principles used by IREOS to give rise a new unsupervised subspace outlier detection algorithm.

1 Introduction

The information technology has evolved extraordinarily. The computer processing capacity has increased over the last 50 years following Moore's law, which states that the processing capacity increases by approximately 100% every 18 months (Moore, 1964, 1965, 1975). At the same time, the storage capacity has increased even faster. As a consequence, the cost of technology has declined, increasing our ability to produce and store data. However, our capacity to collect and store data has far surpassed our capacity to manually analyze and extract knowledge from them (Fayyad et al., 1996). Due to our inability to manually analyze this huge amount of data, researchers from a variety of fields have been engaged in the development of techniques capable of synthesizing, processing and transforming data into useful knowledge in an intelligent and automated way, for what has been called data mining.

The term data mining (Fayyad et al., 1996; Han and Kamber, 2006) widely refers to statistical, mathematical and computational techniques of analysis to efficiently extract useful knowledge from large datasets. There are some central tasks in data mining that together cover most of the existing problems in practical applications (Tan et al., 2006): association analysis, cluster analysis, pattern classification, regression analysis and outlier detection. Among such tasks, the area of outlier (anomaly) detection plays an important role discovering patterns that are exceptional in some sense (Hodge and Austin, 2004; Patcha and Park, 2007; Hadi et al., 2009; Chandola et al., 2009, 2012; Zimek et al., 2012). Detecting such patterns is relevant for two main reasons: (i) in some applications, such patterns represent spurious data (e.g., sensor failures or noise) that should be removed in a preprocessing step for further data analysis; or, more importantly, (ii) in many applications, such patterns represent extraordinary behaviors that deserve some special attention, such as genes associated with certain diseases, frauds in financial systems, employees with unusual productivity profiles, or customers with uncommon purchasing patterns. In this context, it is said that "one person's noise is another person's signal" (Knorr and Ng, 1998; Han and Kamber, 2006).

Outlier detection techniques can be categorized in different ways. For instance, a common distinction is that between the methods that assign binary labels ("outlier" vs. "inlier" for those observations deemed anomalous vs. normal) and methods that assign a score or rank representing a degree to which an observation is considered to be an outlier. Another distinction is that between supervised, semisupervised, and unsupervised outlier detection techniques (Tan et al., 2006). Supervised techniques assume that a set of observed instances labeled as inliers and outliers are available to train a classifier. In the semisupervised scenario, labeled outliers are not available and only previously known inliers can be used in order to obtain a (one class) classification model. When no labeled data are available at all, it is necessary to use unsupervised techniques, which do not assume any prior knowledge about which observations are outliers and which are inliers. One last distinction here[1] is that between the methods that use the full space of features and methods that use only a subset of the features to deem which observations should be detected as outliers, the latter category known as subspace outlier detection (Kriegel et al., 2009b; Zimek et al., 2012).

In this project, we focus on the unsupervised subspace outlier detection scenarios. Without labeled examples, the main complicating factor in this problem is that the notion of "outlierness" is not precisely and generally defined, its formalization rather depends on the application scenario and the detection method to be used. The most common general definitions of outlier remain rather vague, such as these classic examples:

> "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 and Lewis, 1994).

Due to the subjectivity inherent in the unsupervised outlier detection scenario a rich variety of detection methods has been developed, from classic parametric statistical methods (Hawkins, 1980; Barnett and Lewis, 1994) to more recent database-oriented approaches conceived to deal with multivariate, possibly large databases. Considering the latter category, a plethora of detection algorithms has emerged in the past 15 years or so. Examples are DB-Outlier (Knorr and Ng, 1998; Knorr et al., 2000), kNN Outlier (Angiulli and Pizzuti, 2002; Ramaswamy et al., 2000), LOF (Breunig et al., 2000) and its many variants (Jin et al., 2006; Kriegel et al., 2009a, 2011; Papadimitriou et al., 2003; Tang et al., 2002; Zhang et al., 2009) (see, e.g., the work of Schubert et al. (2012b) for a discussion of these and many more variants), and ABOD (Kriegel et al., 2008), just to mention a few. Each of these algorithms, however, uses its own criterion to judge quantitatively the level of adherence of each observation with the concept of outlier, from a particular perspective. This complicates not only the selection of a particular algorithm and/or the choice of an appropriate configuration of parameters for this algorithm in a practical application, but also the assessment of the quality of the solutions obtained, especially in light of the problem of defining a measure of quality that is not tied to the criteria used by the algorithms themselves. These issues are interrelated and refer to the problems of model selection and assessment (evaluation or validation) of results in unsupervised learning. Although these problems have been investigated for decades in the area of unsupervised data clustering (Jain and Dubes, 1988), only recently in outlier detection an index for internal evaluation of top-n (binary) outlier detection results, called IREOS, was firstly proposed by the candidate during his Masters (Marques et al., 2015).

The areas of clustering and unsupervised outlier detection are related to each other and, from a certain perspective, they can even be seen as two sides of the same coin. In fact, when referring to an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism", as in Hawkins' (Hawkins, 1980) definition, it is implicitly assumed that there are one or more mechanisms responsible for generating the normal, "unsuspicious" observations. Clusters are possible candidates to model such mechanisms. In addition, an outlier is not necessarily always an observation which deviates so much from other observations. Indeed, in some application domains, outliers are seen as observations that resemble one another because they are possibly generated by a common mechanism, although they deviate and are generally less frequent than other observations (e.g. some types of fraud or mutations following a common pattern).

Given the aforementioned relationship between clustering and unsupervised outlier detection, it is surprising that while the internal evaluation problem has been extensively studied in data clustering, in unsupervised outlier detection it is still quite open. In the literature, the procedure for evaluation of results in unsupervised

[1] The categorizations mentioned here are not exhaustive, there are other possible categorizations.

outlier detection usually is based on the use of previously labeled datasets, in which the outliers (according to a particular intuition or semantic) are previously known. In this scenario, referred to as external evaluation or validation, the labels are not used by the algorithms, but rather to assess their results only (Zimek et al., 2013, 2012). Specifically, the labeling or ranking produced by the algorithm is compared with the correct and previously known labeling (ground truth). The quantitative result of this comparison is taken as the quality measure of the outlier detection algorithm using, for example, precision-at-n or AUC ROC curve (Schubert et al., 2012a). However, the availability of labeled data required by external evaluation measures is not consistent with the premises of unsupervised learning, and the commonly practiced external evaluation of unsupervised outlier detection algorithms makes sense only when comparing performances of algorithms in controlled experiments. Therefore, although the evaluation procedure described above is useful during the development of new algorithms, it is not feasible in practical applications in the unsupervised context.

In the data clustering domain, the related problems of evaluation and model selection are tackled by using some kind of quantitative index, called validation criterion (Jain and Dubes, 1988). In practice, when labels are not available, internal validation indexes can be used. These indexes are called internal as they do not make use of any external information (such as class labels) in the evaluation of a solution. Instead, internal indexes measure the quality of an obtained clustering solution based only on the solution and the data objects. Most such indexes are also relative in the sense that they can be employed to compare different clustering solutions pointing out which one is better in relative terms. Therefore they can also be used for model selection. Internal, relative indexes have been shown to be effective and useful tools for the unsupervised clustering evaluation and model selection tasks — e.g. see Halkidi et al. (2001); Milligan and Cooper (1985); Vendramin et al. (2010, 2013) and references therein.

In the outlier detection domain only recently an index for internal evaluation of top-n (binary) outlier detection results, called IREOS (Internal, Relative Evaluation of Outlier Solutions), was firstly proposed by the candidate during his Masters (Marques et al., 2015). The index follows the same, common intuition as a multitude of algorithms and criteria: an outlier is an observation that is to some extent farther away and can, therefore, be more easily separated from other observations than an inlier. In order to assess the separability of individual observations is used a maximum margin classifier (e.g. SVM and KLR) (Scholkopf and Smola, 2001; Zhu and Hastie, 2001), as this type of classifier is able to quantify how distant each observation is from the decision boundary while trying to maximize the margin of separability between this boundary and the instances of different classes, the index uses how far the observation is from the decision boundary of the classifier to quantify how easy it is to be separated from the others and takes this magnitude as the basis for quantifying the degree of separability of the observation. This idea is illustrated in Figure 1. Figures 1a, 1b, and 1c highlight different observations labeled as an outlier (red square) in different hypothetical outlier detection solutions. In Figure 1a, the highlighted observation, a genuine global outlier, is far away from a maximum margin classification boundary (dashed line) that discriminates it from the other observations. In Figure 1b, the highlighted observation is arguably a local outlier (w.r.t. the neighboring cluster) and the margin is narrower but still wider than that in Figure 1c. In the case of Figure 1c, the highlighted object is undoubtedly an inlier and not only the margin is very narrow but also the decision boundary needs to be nonlinear (i.e., more complex). In a good solution, consisting mostly of genuine outliers correctly detected by some method, the average degree of separability is expected to be high, whereas in a poor solution containing many false positives this average degree of separability should be lower. IREOS uses this intuition as its fundamental principle, but it also captures additional desirable features for an internal quality measure that does not allow the measure to be trivially optimized and therefore reduced to a mere outlier detection algorithm.

Although the intuition that "an outlier is an observation that is to some extent farther away and can, therefore, be more easily separated from other observations than an inlier" follows the same common intuition as a multitude of algorithms and criteria, it is not tied to any particular criterion used by outlier detection algorithms. Therefore, this concept could be adapted and combined with feature selection techniques, such

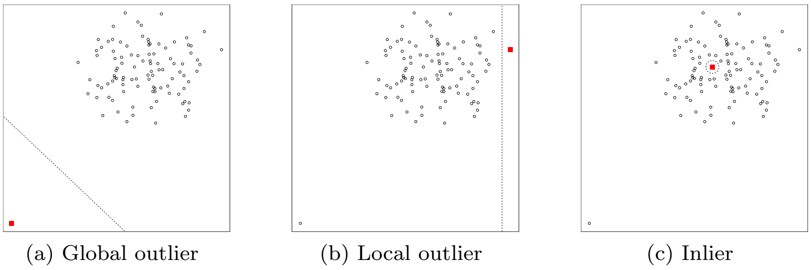
(a) Global outlier (b) Local outlier (c) Inlier

Figure 1: Illustrative data set: three different observation labeled as outliers

as forward selection and backward elimination (Guyon and Elisseeff, 2003), to the development of a new unsupervised subspace outlier detection algorithm. Such adaptation is one of the research topics that this project proposes to investigate.

The other research topic that this project brings is related to the internal evaluation of subspace outlier detection results. Solutions provided by this category of algorithms cannot be evaluated by IREOS. The measure of separability used by IREOS is not commensurable in different subspaces, therefore, the separabilities of two observations detected in different subspaces of the data cannot be directly compared. In this research topic, we intend to investigate means to address this problem, possibly, using procedures similar to those used by subspace (Kriegel et al., 2009b) and ensemble outlier (Lazarevic and Kumar, 2005) detection algorithms.

Both topics of research, that consists the main objectives of the proposal, are further discussed in the section 2. The methodology that will be used in this project, i.e., the datasets and evaluation procedures are described in section 3. In section 4, information concerning the university selected for the internship is provided. This section also provides details regarding the foreign advisor that will be supervising the student during his internship. Finally, in section 5, both the activities and the schedule of the present research proposal are presented.

2 Objectives

The idea of defining outliers with respect to a subspace of the original data space arose from Knorr and Ng (1999). With the intensification of the problem of high dimensionality that datasets from practical applications have suffered, the subspace outlier detection techniques have received increasing attention due to the fact that many irrelevant features may easily mask outliers. Such techniques attempt to reveal the subset of features where the observation becomes more detectable. However, the discovery of such subset of features is not a trivial task. In this context, different techniques of subspace outlier detection have been proposed, such as: Feature Bagging (Lazarevic and Kumar, 2005), PCOut (Filzmoser et al., 2008), OutRank (Müller et al., 2008), SOD (Kriegel et al., 2009b), OUTRES (Müller et al., 2010), HighDOD (Nguyen et al., 2011), HiCS (Keller et al., 2012), among others. These techniques may differ in the exploratory approach of the candidate subspaces and/or in the way that the outlier scores are produced/combined over the different subspaces.

Although there is a growing literature that tackles the unsupervised subspace outlier detection problem, the unsupervised evaluation of subspace outlier detection results has been notably overlooked. In this project, we propose a step towards bridging this gap by the adaptation of IREOS to cover such scenario. In order to evaluate the quality of the solutions produced by these algorithms, the separability of the observations should be measured in the respective subspaces where such observations were evaluated as outliers. In IREOS present form, however, the measure of separability is not commensurable over the different subspaces, therefore, it can not be directly applied to evaluate the observation separability over the different subspaces. The approaches used by the different subspace outlier detection techniques for the commensurability of the scorings produced over the different subspaces are diverse (Lazarevic and Kumar, 2005; Filzmoser et al., 2008; Müller et al., 2008, 2010; Kriegel et al., 2009b; Nguyen et al., 2011; Keller et al., 2012). The study of these different approaches, as well as application or adaptation to the commensurability of the IREOS index over the different subspaces, comprises the initial approach that will be used to extend the IREOS index to evaluate solutions of subspace outlier detection algorithms.

Another research topic that we intend to investigate in this project is related to how some of the principles used by IREOS could be applied, not to evaluate a given outlier detection solution but, to detect them. Although the basic concept of separability applied by the index fits in all classical outlier definitions, it is not used by any particular outlier detection algorithm. The index itself is not trivial to be optimized because IREOS captures desirable features for an internal quality measure that does not allow the measure to be trivially optimized and therefore reduced to an outlier detection algorithm. However, the concept of separability used by the index could be applied/adapted aiming a new subspace outlier detection algorithm. The basic hypothesis to be explored initially is that the outlier scores should be somehow positively related to the degree of separability of each observation. In order to explore the candidate subspaces, we intend to use feature selection techniques, such as forward selection and backward elimination (Guyon and Elisseeff, 2003), as the first possible approach to finding the subspace where each observation has its highest degree of separability. A similar approach was used earlier by Micenkova et al. (2013), however, Micenkova et al. (2013) applied such an approach in a distinct context of post-processing, i.e., after outliers were detected, to find the subspace that better could explain why the observation was labeled as an outlier.

3 Methodology

The present research project will have as its starting point the Masters of the candidate (Marques, 2015), in which the internal evaluation measure (IREOS) was developed, and his PhD, in which the candidate still working with topics related to the measure. Therefore, both the computational implementation of IREOS and collections of real and synthetic datasets for outlier detection are already available. However, we intend to increase the number of datasets used, possibly, using outlier detection evaluation repositories that have been recently proposed in the literature (Campos et al., 2016; Goldstein and Uchida, 2016). Using these datasets, the results will be evaluated using the following methodologies:

- **Index evaluation:** The evaluation of the new index to be developed will be made from the ground truth of the datasets. The ground truth will not be used by the index in the evaluation process of the outlier detection solution, but rather for the comparison between the quality measure provided by the index and the quality measure according to the ground truth. External measures such as precision-at-n and AUC ROC curve (Schubert et al., 2012a) will be correlated with the quality measure provided by the internal evaluation index. It is expected that the better the (internal) index performance, the greater the correlation of this index with external indexes based on datasets with ground truth.

- **Algorithm evaluation:** The results of the new outlier detection algorithm to be developed will also be evaluated using external evaluation measures (e.g. precision-at-n and AUC ROC curve) and compared to the results of the same measures produced by the most well-known outlier detection techniques in the literature. The computational implementations of such techniques are mostly available in open source tools such as ELKI[2] and SOREX[3].

[2] http://elki.dbs.ifi.lmu.de/

[3] http://dme.rwth-aachen.de/de/OpenSubspace/SOREX

4 The University and the Supervisor

The University of Alberta[4] (U of A) has been ranked among the top five universities in Canada and among the top public research universities worldwide. It is currently considered the 107^{th} best university of the world by the Times Higher Education World University Rankings (THE)[5] and 94^{th} best university of the world by the Quacquarelli Symonds World University Rankings (QS)[6], two of the most renowned international university rankings. The University of Alberta is also listed as one of the best universities in North America according to a list released by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[7].

The Prof. Dr. Jörg Sander[8] is a renowned and very active researcher in the field of Data Mining, due to all his relevant scientific contributions, he carries since August 2015 the "ACM Distinguished Scientist" award. He is one of the co-creators of the well-known algorithm DBSCAN (Ester et al., 1996), responsible for the introduction of the density-based paradigm to the data mining community and the first paper to receive the "KDD Test of Time Award"[9], that "recognizes outstanding papers from past KDD Conferences beyond the last decade that have had an important impact on the data mining research community". Prof. Sander is also one of responsible for the introduction of the density-based clustering algorithm OPTICS (Ankerst et al., 1999) and the density-based outlier detection algorithm LOF (Breunig et al., 2000). These algorithms have been described in many books on Data Mining (Tan et al., 2006; Han et al., 2011), Cluster Analysis (Gan et al., 2007) and Outlier Detection (Aggarwal, 2013) and together these 3 papers count over 15,000 citations. He is also one of the co-inventors of state-of-the-art algorithms for projected and subspace clustering, such as P3C (Moise and Sander, 2008; Moise et al., 2009) and STATPC (Moise and Sander, 2008; Moise et al., 2009). He has regularly published papers in journals of recognized excellence, as well as, in selective conferences with high impact in the Data Mining field, such as: IEEE Int. Conf. on Data Mining (IEEE ICDM) and ACM Int. Conf. on Knowledge Discovery and Data Mining (ACM SIGKDD). He has also regularly served as Program Committee member for these and other important related conferences — e.g., the SIAM Int. Conf. on Data Mining (SIAM SDM) and the IEEE Int. Conf. on Data Engineering (IEEE ICDE).

It is important to note that the PhD project of the candidate is inserted in a context of international collaboration that his advisor, Prof. Dr. Ricardo J. G. B. Campello, established during his most recent postdoctoral, that also includes the Prof. Jörg Sander (Prof. Dr. José Fernando Rodrigues Júnior has been the advisor in charge during the leaving of Prof. Ricardo J. G. B. Campello from the University of São Paulo). The candidate is inserted in this international collaboration since his Masters and it already gave rise to two papers with co-authorship of Prof. Jörg Sander (Marques et al., 2015; Swersky et al., 2016). It is expected that the internship under the supervision of Prof. Jörg Sander will be beneficial to the student in different aspects. The internship envisioned in this proposal will be of great value to the professional qualification of the student, as he will have the opportunity to interact with international researchers and other students from a university of recognized excellence.

5 Activities and Schedule

The schedule of activities is described in Table 1. All activities are listed in the following.

1. Study and proposal of an index for evaluation of subspace outlier detection results.

2. Experiments and evaluation of the index for evaluation of subspace outlier detection results.

[4] http://www.ualberta.ca

[5] https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking

[6] http://www.topuniversities.com/university-rankings/world-university-rankings/2016

[7] http://capes.gov.br/cienciasemfronteiras/html/ranking.html

[8] http://www.cs.ualberta.ca/~joerg

[9] http://www.kdd.org/News/view/2014-sigkdd-test-of-time-award

3. Study and proposal of the new unsupervised subspace outlier detection algorithm.

4. Experiments and evaluation of the new unsupervised subspace outlier detection algorithm.

5. Writing of papers and technical reports.

Activity	Month											
	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG
1	●	●	●	●								
2		●	●	●	●							
3						●	●	●	●	●		
4							●	●	●	●	●	
5					●	●					●	●

Table 1: Schedule of activities regarding the referred project.

References

Aggarwal, C. C. (2013). *Outlier Analysis.* Springer.

Angiulli, F. and Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. pages 15–26, Helsinki, Finland.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. pages 49–60. ACM Press.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons, 3rd edition.

Breunig, M. M., Kriegel, H.-P., Ng, R., and Sander, J. (2000). Lof: Identifying density-based local outliers. pages 93–104, Dallas, TX.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30(4):891–927.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58.

Chandola, V., Banerjee, A., and Kumar, V. (2012). Anomaly detection for discrete sequences: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 24(5):823–839.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press.

Fayyad, U. M., Shapiro, G. P., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* MIT Press.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Comput. Stat. Data Anal.*, 52(3):1694–1711.

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability).* SIAM, Society for Industrial and Applied Mathematics, illustrated edition edition.

Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, 11(4):1–31.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Hadi, A. S., Imon, A. H. M. R., and Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jin, W., Tung, A. K. H., Han, J., and Wang, W. (2006). Ranking outliers using symetric neighborhood relationship. pages 577–593, Singapore.

Keller, F., Müller, E., and Böhm, K. (2012). Hics: high contrast subspaces for density-based outlier ranking. Washington, DC.

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, pages 392–403, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Knorr, E. M. and Ng, R. T. (1999). Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 211–222, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Knorr, E. M., Ng, R. T., and Tucanov, V. (2000). Distance-based outliers: Algorithms and applications. *The VLDB Journal*, pages 8(3–4):237–253.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009a). Loop: local outlier probabilities. pages 1649–1652, Hong Kong, China.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009b). Outlier detection in axis-parallel subspaces of high dimensional data. pages 831–838, Bangkok, Thailand.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2011). Interpreting and unifying outlier scores. pages 13–24, Mesa, AZ.

Kriegel, H.-P., Schubert, M., and Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 444–452, New York, NY, USA. ACM.

Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection. pages 157–166, Chicago, IL.

Marques, H. O. (2015). Avaliação e seleção de modelos em detecção não supervisionada de outliers. Master's thesis, Universidade de São Paulo, São Carlos - SP, Brazil.

Marques, H. O., Campello, R. J. G. B., Zimek, A., and Sander, J. (2015). On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, SSDBM '15, pages 7:1–7:12, New York, NY, USA. ACM.

Micenkova, B., Dang, X.-H., Assent, I., and Ng, R. (2013). Explaining outliers by subspace separability. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 518–527.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179.

Moise, G. and Sander, J. (2008). Finding non-redundant, statistically significant regions in high dimensional data: A novel approach to projected and subspace clustering. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 533–541, New York, NY, USA. ACM.

Moise, G., Zimek, A., Kröger, P., Kriegel, H., and Sander, J. (2009). Subspace and projected clustering: Experimental evaluation and analysis. *Knowledge and Information Systems*, 21(3):299–326.

Moore, G. (1964). The future of integrated electronics. *Fairchild Semiconductor internal publication.*

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, 38(8).

Moore, G. (1975). Progress in digital integrated electronics. *Electron Devices Meeting, 1975 International*, 21:11–13.

Müller, E., Assent, I., Steinhausen, U., and Seidl, T. (2008). Outrank: ranking outliers in high dimensional data. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 600–603.

Müller, E., Schiffer, M., and Seidl, T. (2010). Adaptive outlierness for subspace outlier ranking. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1629–1632, New York, NY, USA. ACM.

Nguyen, H. V., Gopalkrishnan, V., and Assent, I. (2011). 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*, DASFAA'11, pages 138–152, Berlin, Heidelberg. Springer-Verlag.

Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. pages 315–326, Bangalore, India.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.*, 51(12):3448–3470.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. pages 427–438, Dallas, TX.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA.

Schubert, E., Wojdanowski, R., Zimek, A., and Kriegel, H.-P. (2012a). On evaluation of outlier rankings and outlier scores. pages 1047–1058, Anaheim, CA.

Schubert, E., Zimek, A., and Kriegel, H.-P. (2012b). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video and network outlier detection. *Data Mining and Knowledge Discovery.*

Swersky, L., Marques, H. O., Sander, J., Campello, R. J. G. B., and Zimek, A. (2016). On the evaluation of outlier detection and one-class classification methods. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining.* Addison Wesley.

Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. pages 535–548, Taipei, Taiwan.

Vendramin, L., Campello, R. J. G. B., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

Vendramin, L., Jaskowiak, P. A., and Campello, R. J. G. B. (2013). On the combination of relative clustering validity criteria. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 4:1–4:12, New York, NY, USA. ACM.

Zhang, K., Hutter, M., and Jin, H. (2009). A new local distance-based outlier detection approach for scattered real-world data. pages 813–822, Bangkok, Thailand.

Zhu, J. and Hastie, T. (2001). Kernel logistic regression and the import vector machine. In *Journal of Computational and Graphical Statistics*, pages 1081–1088. MIT Press.

Zimek, A., Campello, R. J. G. B., and Sander, J. (2013). Ensembles for unsupervised outlier detection: Challenges and research questions. *ACM SIGKDD Explorations*, 15:11–22.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.*, 5(5):363–387.