www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 5 May 2016, Page No. 16543-16547

Outlier Detection in High Dimension Data Based On Multimodality And Neighbourhood Size Using KNN Method

Deepthi Navile, Dr. Ravikumar G.K

4thsemMTech,
Department of computer science and engineering,
BGSIT, Mandya
deepthinavile18@gmail.com
Professor, Department of Computer science and engineering,
BGSIT,Mandya
ravikumargk@yahoo.com

Abstract—Outlier detection in high-dimensional data presents various challenges resulting from the "curse of dimensionality." A prevailing view is that distance concentration, i.e., the tendency of distances in high-dimensional data to become indiscernible, hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper, we provide evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, we show that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context. Namely, it was recently observed that the distribution of points' reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as hubness. We provide insight into how some points (antihubs) appear very infrequently in k-NN lists of other points, and explain the connection between antihubs, outliers, and existing unsupervised outlier-detection methods. By evaluating the classic k-NN method, the angle-based technique designed for high-dimensional data, the density-based local outlier factor and influenced outlierness methods, and antihub-based methods on various synthetic and real-world data sets, we offer novel insight into the usefulness of reverse neighbor counts in unsupervised outlier detection. Index Terms—Outlier detection, reverse nearest neighbors, high-dimensional data, distance concentration

1 INTRODUCTION

OUTLIER (anomaly) detection refers to the task of identifying patterns that do

not conform to established regular behavior [1]. Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice [2]. The interest in outliers is strong since they may constitute critical and actionable information in various domains, such as intrusion and fraud detection, and medical diagnosis. The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Among these categories, unsupervised methods are more widely applied [1], because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods [3], [4], [5] that mainly rely on a measure of distance or similarity in order to detect outliers. A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless [6], since distance measures concentrate, i.e., pairwise distances become indiscernible as dimensionality increases [7], [8]. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good outlier [9]. This somewhat simplified view was recently challenged [10]. Our motivation is based on the following factors: 1) It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in [10] the actual challenges posed by the "curse of dimensionality" differ from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space [9]. We will present further evidence which challenges this view, motivating the (re)examination of methods. 2)

Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points [11], [12],1 but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data [14] warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method [11]. Our contributions can be summarized as follows: 1) In Section 3 we discuss the challenges that unsupervised outlier detection faces in high-dimensional space. Despite the general impression that all points in a high-dimensional data set seem to become outliers [9], we show that unsupervised methods can detect outliers which are more pronounced in high dimensions, under the assumption that all (or most) data attributes are meaningful, i.e. not noisy. Ourfindings complement the observations from [10] by demonstrating such behavior on data originating from a single distribution without outliers generated by a different mechanism. Also, we explain how high dimensionality causes such pronounced outlierness in comparison with low-dimensional settings. 2) Recently, the phenomenon of hubness was observed [14], which affects reverse nearest-neighbor counts, i.e. k-occurrences (the number of times point x appears among the k nearest neighbors of all other points in the data set). Hubnessis manifested with the increase of the (intrinsic) dimensionality of data, causing the distribution of k-occurrences to become skewed, also having increased variance. As a consequence, some points (hubs) very frequently become members of k-NN lists and, at the same time, some other points (antihubs) become infrequent neighbors. In Section 4 we examine the emergence of antihubs and the way it relates to outlierness of points, also considering lowdimensional settings, extending our view to the full range of neighborhood sizes, and exploring the interaction of hubness and data sparsity. 3) Based on the relation between antihubs and outliers in high- and low-dimensional settings, in Section 5 we explore two ways of using k-occurrence information for expressing the outlierness of points, starting with the method ODIN proposed in [11]. Our main goal is to provide insight into the behavior of k-occurrence counts in different realistic scenarios (high and low dimensionality, multimodality of data), that would assist researchers and practitioners in using reverse neighbor information in a less ad-hoc fashion. 4) Finally, in Section 6 we describe experiments with synthetic and real data sets, the results of which illustrate the impact of factors such as dimensionality, cluster density and antihubs on outlier detection, demonstrating the benefits of the methods, and the conditions in which the benefits are expected.

2 RELATED WORK

According to the categorization in [1], the scope of our investigation is to examine: (1) point anomalies, i.e., individual points that can be considered as outliers without taking into account contextual or collective information, (2) unsupervised methods, and (3) methods that assign an "outlier score" to each point, producing as output a list of outliers ranked by their scores. The described scope of our study is the focus of most outlier-detection research [1]. Among the most widely applied methods within the described scope are approaches based on nearest neighbors, which assume that outliers appear far from their closest neighbors. Such methods rely on a distance or similarity measure to find the neighbors, with Euclidean distance being the most popular option. Variants of neighbor-based methods include defining the outlier score of a point as the distance to its kth nearest neighbor [3] (henceforth referred to as the k-NN method), or as the sum of distances to the k nearest neighbors [4]. Related to these methods are approaches that determine the score of a point according to its relative density, since the distance to the kth nearest neighbor for a given data point can be viewed as an estimate of the inverse density around it [5]. A widely-useddensitybased method is the local outlier factor (LOF) [15], which influenced many variations, e.g., the local correlation integral (LOCI) [16], local distance-based outlier factor (LDOF) [17], and local outlier probabilities (LoOP) [18]. The angle-based outlier detection (ABOD) [19] technique detects outliers in high-dimensional data by considering the variances of a measure over angles between the difference vectors of data objects. ABOD uses the properties of the variances to actually take advantage of high dimensionality and appears to be less sensitive to the increasing dimensionality of a data set than classic distance-based methods. The study in [20] distinguishes three problems brought by the "curse of dimensionality" in the general context of search, indexing, and data mining applications: poor discrimination of distances caused by concentration, presence of irrelevant attributes, and presence of redundant attributes, all of which hinder the usability of traditional distance and similarity measures. The authors conclude that despite such limitations, common distance/similarity measures still form a good foundation for secondary measures, such as shared-neighbor distances, which are less sensitive to the negative effects of the curse. Zimek et al. [10] continue the discussion of problems relevant to unsupervised outlier-detection methods in highdimensional data by identifying seven issues in addition to distance concentration: noisy attributes, definition of reference sets, bias (comparability) of scores, interpretation and contrast of scores, exponential search space, data-snooping bias, and hubness. In this article we will focus on the aspect of hubness, and assume that all attributes carry useful information, i.e., are not overly noisy. Finally, the notion of reverse nearest neighbors, considered important in

DOI: 10.18535/ijecs/v5i5.35

areas outside outlier detection [21], [22], was used to formulate outlier scores in various ways. In [11], the reverse k-nearest neighbor count is defined to be the outlier score of a point in the proposed method ODIN, where a user-provided threshold parameter determines whether a point is designated as an outlier or not. Experiments were performed on low-dimensional data, and offered little insight into the reason why reverse nearest neighbors should constitute meaningful outliers. In [12], a method for detecting outliers based on reverse neighbors was briefly considered, judging that a point is an outlier if it has a zero k-occurrence count. The proposed method also does not explain the mechanism which creates points with low k-occurrences, and can be considered a special case of ODIN with the threshold set to 0. In [23], the relation

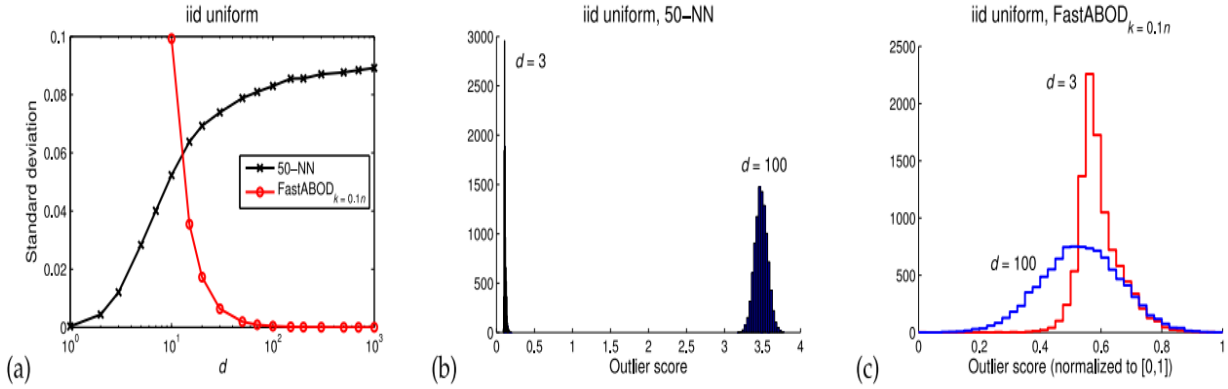

Fig. 1. Outlier scores versus dimensionality d for uniformly distributed data in ½0; 1 d : (a) Standard deviation; (b) Histogram of 50-NN scores; (c) Histogram of normalized ABOD scores.

between reverse nearest neighbors and outliers was explored, but again no investigation was performed on how reverse neighbors are connected with high-dimensional phenomena, focusing instead on application to stream mining and improving the execution time of reverse nearest-neighbor computation.The main focus of [24] was on the efficiency of computing INFLO scores. In contrast to all approaches above, we focus on highdimensional as well as low-dimensional data and use reverse nearest neighbors only through the distribution of k-occurrences, taking into account the inherent relationship between dimensionality, neighborhood size and reverse neighbors that was not observed in previous outlier-detection work. In doing so, we will revisit the outlier scoring method ODIN [11].

3 OUTLIER DETECTION IN HIGH DIMENSIONS: IMPROVING THE PERSPECTIVE In this section we revisit the commonly accepted view that in high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases [9]. In [10] this view was challenged by showing that the exact opposite may take place: as dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods, assuming all dimensions carry useful information. We present an example revealing that this can happen even when no true outliers exist, in the sense of originating from a different distribution than other points. Example 3.1. Let us observe n ¼ 10;000 d-dimensional points, whose components are independently drawn from the uniform distribution in range ½0; 1. We employ the classic k-NN method [3] (k ¼ 50; similar results are obtained with other values of k). We also examine ABOD [19] (for efficiency reasons we use the FastABOD variant with k ¼ 0:1n), and use standard deviation to express the variability in the assigned outlier scores. Fig. 1a illustrates the standard deviations of outlier scores against dimensionality d. Let us observe the k-NN method first. For small values of d, deviation of scores is close to 0, which means that all points tend to have almost identical outlier scores. This is expected, because for low d values, points that are uniformly distributed in ½0; 1 d contain no prominent outliers. This assumption also holds as d increases, i.e., still there should be no prominent outliers in the data. Nevertheless, with increasing dimensionality, for k-NN there is a clear increase of the standard deviation. This increase indicates that some points tend to have significantly smaller or larger outlier scores than others. This can be observed in the histogram of the outlier scores in Fig. 1b, for d ¼ 3 and d ¼ 100. In the former case, the vast majority of points have very similar scores. The latter case, however, clearly shows the existence of points in the right tails of the distributions which are prominent outliers, as well as points on the opposite end with much smaller scores. 1a says little about the expected performance of ABOD, which ultimately depends on the quality of the produced outlier rankings [10]. However, when scores are regularized by logarithmic inversion and linearly normalized to the ½0; 1 range [25], a trend similar to k-NN can be observed, shown in Fig. 1c. As discussed, high dimensionality causes the emergence of some points that tend to be clearly detected as outliers by common unsupervised methods. This happens despite the fact that the

existence of prominent outliers is not expected. Apparently, it is only the increase of dimensionality that caused the generation of the prominently scored outliers. This observation raises several questions: Is such behavior an artefact of the selected data distribution? Is it a property of the distance function used? Can these prominent outliers somehow be characterized? In the example above, we chose the setting involving uniformly distributed random points because of the intuitive expectation that it should not contain any really prominent outliers. Analogous observations can be made with other data distributions, numbers of drawn points, and distance measures. The demonstrated behavior is actually an inherent consequence of increasing dimensionality of data, with the tendency of the detected prominent outliers to come from the set of antihubs—points that appear in very few, if any, nearest neighbor lists of other points in the data.

4 ANTIHUBS AND OUTLIERS

In this section, we observe antihubs as a special category of points in high-dimensional spaces. We explain the reasons behind the emergence of antihubs and examine their relation to outliers detected by unsupervised methods in the context of varying neighborhood size k. Finally, we explore the interplay of hubness and data sparsity. 4.1 Antihubs: Definition and Causes The existence of antihubs is a direct consequence of high dimensionality when neighborhood size k is small compared to the size of the data. To understand this relationship more clearly, let us first briefly review the counterintuitive concentration behavior of distances as dimensionality increases [8]. Distance concentration refers to the tendency of distances in high-dimensional data to become almost indiscernible as dimensionality increases, and is usually expressed through a ratio of a notion of spread (e.g., standard deviation) and magnitude (e.g., the expected value) of the distribution of distances of all points in a data set to some reference point. If this ratio tends to 0 as dimensionality goes to infinity, it is said that distances concentrate. Considering random data with iid coordinates and Euclidean distance, concentration is reflected in the fact that, as dimensionality increases, the standard deviation of the distribution of distances remains constant, while the mean value continues to grow. More visually it can be said that, as dimensionality increases, all points tend to lie approximately on a hypersphere centered at the reference point, whose radius is the mean distance. It is important to note that in high-dimensional space any point can be used as the reference point, producing the concentration effect: the radius of the sphere (the expected distance to the reference point) increases with dimensionality, while the spread of points above and below the surface (e.g., the standard deviation of the distance distribution) becomes negligible compared to the radius. Returning to antihubs, their emergence is an aspect of the "curse of dimensionality" related to distance concentration. This aspect will be generally referred to as hubness [14]. To describe hubness, let us define the notions of k-occurrences, hubs and antihubs. 4.2 The Relation BetweenAntihubs and Outliers Outlier-detection methods can generally be categorized into global and local approaches, i.e., the decision on the outlierness of some data object can be based on the complete (global) database or only on a (local) selection of data objects [27]. Naturally, there can exist a whole continuum of degrees between the two opposing extremes of "global" and "local," where the degree of locality may be tunable using parameters. For example, by raising the value of k when using the classic k-NN outlier detection method, one increases the set of data points used to determine the outlier score of the point of interest, moving from a local to a global notion of outlierness, and ending in the extreme case when k ¼ n 1. Likewise, raising k when determining reverse nearest neighbors, i.e., antihubs, raises the expected size of reverse-neighbor sets (while their size can still vary amongst points).5 Since antihubs have been defined as points with the lowest Nk values, we can explore the relation between Nk scores and outlierness by measuring the correlation between Nk values and outlier scores produced by unsupervised methods. For the data in The measured correlations are plotted in Figs. 3a and 3b, together with the correlation between inverse Nk values and the distance to the data set mean (Fig. 3c) for two values of dimensionality: low (d ¼ 2) and high (d ¼ 100). Furthermore, we consider two portions of points for computing correlations: all points (p ¼ 100%) and p ¼ 5% of points with the highest distance from the data set mean as the strongest outliers. It can be seen that for the highdimensional case correlations for p ¼ 100% are very strong for a wide range of k values, with the exceptions being very low (close to 1) and very high values (close to n ¼ 10;000). For p ¼ 5% agreement between NkandIn summary, the emergence of antihubs is closely connected with outliers both in high-dimensional and lowdimensional data. The examples above illustrate this connection, and suggest that antihubscan be used as an alternative to standard outlier-detection methods. However, from the discussion above one could deduce that antihubs simply provide a crude approximation of established outlier scoring methods for some ranges of values of parameter k. As we will see in the next section, this is not the case, since in more realistic settings involving multimodal data the correlations can behave quite differently. 4.3 Multimodality and Neighborhood Size Real data differs from the synthetic examples from previous sections in many respects, including existence of multiple clusters in the data, and possibility that different regions where data resides have different densities.

5 CONCLUSIONS

In this paper, we provided a unifying view of the role of reverse nearest neighbor counts in problems concerning unsupervised outlier detection, focusing on the effects of high dimensionality on unsupervised outlier-detection methods and the hubness phenomenon, extending the previous examinations of (anti)hubness to large values of k, and exploring the relationship between hubness and data sparsity. Based on the analysis, we formulated the AntiHub method for unsupervised outlier detection, discussed its properties, and proposed a derived method which improves discrimination between scores. Our main hope is that this article clarifies the picture of the interplay between the types of outliers and properties of data, filling a gap in understanding which may have so far hindered the widespread use of reverse-neighbor methods in unsupervised outlier detection. The existence of hubs and antihubs in high-dimensional data is relevant to machine-learning techniques from various families: supervised, semi-supervised, as well as unsupervised .In this paper we focused on unsupervised methods, but in future work it would be interesting to examine supervised and semi-supervised methods as well. Another relevant topic is the development of approximate versions of AntiHub methods that may sacrifice accuracy to improve execution speed. An interesting line of research could focus on relationships between different notions of intrinsic dimensionality, distance concentration, (anti)hubness, and their impact on subspace methods for outlier detection. Finally, secondary measures of

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Survey, vol. 41, no. 3, p. 15, 2009.

[2] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Hoboken, NJ, USA: Wiley, 1987.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2, pp. 427–438, 2000.

[4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in Proc. Conf. Appl. Data Mining Comput. Security, 2002, pp. 78–100.

[5] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237– 253, 2000.

[6] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217–235.

[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. 8th Int. Conf. Database Theory, 2001, pp. 420–434.

[8] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Trans. Knowl.Data. Eng., vol. 19, no. 7, pp. 873–886, Jul. 2007.

[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in Proc. 27th ACM SIGMOD Int. Conf. Manage. Data, 2001, pp. 37–46.

[10] A. Zimek, E. Schubert, and H.-P.Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist.Anal.Data Mining, vol. 5, no. 5, pp. 363–387, 2012.

[11] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430–433.

[12] J. Lin, D. Etter, and D. DeBarr, "Exact and approximate reverse nearest neighbor search for multimedia data," in Proc 8th SIAM Int. Conf. Data Mining, 2008, pp. 656–667.

[13] A. Nanopoulos, Y. Theodoridis, and Y. Manolopoulos, "C2P: Clustering based on closest pairs," in Proc 27th Int. Conf. Very Large Data Bases, 2001, pp. 331–340.

[14] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," J. Mach. Learn. Res., vol. 11, pp. 2487–2531, 2010.

[15] M. M. Breunig, H.-P.Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," SIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.

[16] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in Proc 19th IEEE Int. Conf. Data Eng., 2003, pp. 315–326.

[17] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc 13th Pacific-Asia Conf. Knowl. Discovery Data Mining, 2009, pp. 813–822

. [18] H.-P.Kriegel, P. Kroger, E. Schubert, and A. Zimek, "LoOP: Local € outlier probabilities," in Proc 18th ACM Conf. Inform. Knowl.Manage., 2009, pp. 1649–1652.