International Journal of Computer Applications Technology and Research
Volume 4– Issue 11, 797 - 801, 2015, ISSN: 2319–8656

Hubness in Unsupervised Outlier Detection Techniques for High Dimensional Data –A Survey

R.Lakshmi Devi
Mother Teresa Women's University
Kodaikanal, India

R.Amalraj
Dept. of Computer Science
Sri Vasavi Arts College
Erode, India

Abstract: -- Outlier detection in high dimensional data becomes an emerging technique in today's research in the area of data mining. It attempts to find objects that are considerably unrelated, unique and inconsistent with respect to the majority of data in an input database. It also poses various challenges resulting from the increase of dimensionality. Due to the "increase of dimensionality," distance becomes meaningless. Hubness is an aspect for the increase of dimensionality pertaining to nearest neighbors which has come to an attention. This survey article, discusses some important aspects of the hubness in detail and presents a comprehensive review on the state-of-the-art specialized algorithms for unsupervised outlier detection for high dimensional data and role of hubness.

Keywords: Hubness, High dimensional data, Outliers, Outlier detection, Unsupervised.

1. INTRODUCTION

An outlier is an observation which appears to be inconsistent with the remainder of that set of data. In data mining, detection of outliers is more interesting and an important research area. Many applications which apply outlier detection are Fraud detection for credit card, Loan application processing, Intrusion detection cyber-security, Network performance, insurance or health and so on. Most of these applications are high dimensional which means that data can contain hundreds of dimensions. Moreover in high dimensional space, the data is sparse. The sparsity of high dimensional data denotes that every point is an almost equally good outlier [1]. Many algorithms have been proposed in recent years for handling such a problem. This paper focuses a broad overview of extensive research on outlier detection techniques for high dimensional data and a role of hubness.

The various sections to be discussed are categorized in the following order: an introduction to the outlier detection is explained in Section 2, various outlier detection techniques for high dimensional data in Section 3, and Section 4 provides a detailed introduction to the phenomenon of hubness and also examines how hubness is used in various techniques. Finally the chapter is concluded in Section 5.

2. OUTLIER DETECTION

Outlier (anomaly) detection refers to the process of finding patterns that do not conform to standard behavior. These non-conforming patterns are often referred to as anomalies, outliers, exceptions in different application domains.

There are three categories of outliers:

1. Point Outliers: If specific data instance is considered as inconsistent with respect to the rest of data, then the instance is stated as a point outlier.

2. Context Outlier: If a data instance is inconsistent in a specific context, then it is stated as a contextual outlier.

3. *Collective Outlier:* If a collection of related data instances is inconsistent with respect to the entire data set, it is stated as a collective outlier [2].

Outliers can also be categorized into vector, sequence, trajectory and graph outliers etc., depending on the type of data from where outliers can be detected.

Vector outlier: Vector outliers are detected from vector like representation of Data such as the relational database.

Sequence outlier: In many applications, data are existing as a sequence. A good example is the computer system Call log.

Trajectory outlier: Recent developments in satellites and tracking facilities utilize a huge amount of trajectory data of moving objects. For eg. Animal movement data.

Graph outlier: Graph outliers are graph entities that are strange from their Peers. Examples are nodes, edges and sub graphs [3].

The labels of a data instance represent if that instance is *normal* or *Anomalous.* Depending on the labels connected the data instance, anomaly detection techniques can be in one of the following three modes:

Supervised anomaly detection. Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as anomaly class.

Semi-Supervised anomaly detection. Techniques that operate in a semi supervised mode, assume that the training data has labeled instances for only the normal class. Since they do not require labels for the anomaly class, they are more widely applicable than supervised techniques.

Unsupervised anomaly detection. Techniques that operate in unsupervised mode do not require training data, and thus are most widely applicable [2].

Among these categories, unsupervised methods are more widely applied [4], because the other categories require accurate and representative labels that are often expensive to obtain. [4] Shows 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.

3. OUTLIER DETECTION FOR HIGH DIMENSIONAL DATA

There are many applications in high dimensional domains in which data contain dozens or even hundreds of dimensions. In high dimensional space, the data are sparse and the idea of proximity fail to achieve their effectiveness. This is due to the increase of dimensionality that concentrates the high dimensional data tend to be equi-distant to each other as dimensionality increases.

To face the challenge associated with high data dimensionality, two different categories of work is to be conducted. The first category of methods project the high dimensional data to lower dimensional data. Dimensionality detection techniques such as Principle Component Analysis (PCA), Independent component Analysis (ICA), singular value Decomposition (SVD) etc. can be applied to the high dimensional data before outlier detection is performed.. Second category is to redesign the mechanism to accurately capture the proximity relationship between data points in the high dimensional space [3].

The earlier research methods for unsupervised outlier detection in high dimensional data are discussed below.

Since the distance based method for outlier detection is not effective due to various noise in high dimensional data [5] utilizes an approach named ABOD (Angle Based Outlier Detection) which uses the variance of angles between the points as an outlier degree with the assumption that angles are more stable than distances Also compares ABOD with the distance-based method LOF for a real world data set to show that ABOD to perform well on high-dimensional data.

The paper [6] proposes method for finding distance-based *outliers* based upon the k nearest neighbor points. To calculate the number of data points falling, two algorithms such as nested loop join and index join algorithms are used. Also partition-based algorithm is used. This algorithm divides the data set into different subsets and then cuts entire partitions rapidly as it is determined that they cannot contain outliers.

Distance based method to deal with the problem of finding outliers for k dimensional data sets where k >=5 is focused in paper [7]. Applying three algorithms such as index based, nested loop based, and cell based, authors come to the conclusion that cell based is for k<=4 and nested loop is the choice for k>=5 and also finds that there is no limit on the size of the dimensions.

The approach used in [8] is based on the relationship between k- Nearest Neighbor and Reverse Nearest Neighbor which involves two phases. First phase is dealing with the problem of finding a query point using kNN.The second step introduces Boolean Range Query which checks the existence of a point in a given region can be used for RNN queries problems.

Density based local outlier is identified in [9] with the help of LOF for many KDD (Knowledge discovery in databases) applications, such as detecting criminal activities in E-commerce, finding the rare instances or the outliers. Author introduces LOF (Local outlier factor) which is local meaning that it considers only the restricted neighborhood. Also proves that LOF is useful for finding the meaningful outliers in real world dataset.

A new method (LOCI—Local Correlation Integral method) for finding outliers in large, multidimensional data sets is recognized in [10] which Introduces the multi-granularity deviation factor (MDEF), to detect both remote outliers as well as outlying clusters and proposes a method which is associated with MDEF to check the existence of an outlier point by comparing MDEF value with local average. It is automatic and can be computed quickly.

LDOF (Local Distance-based Outlier Factor) approach [11] to handle KDD applications, measures the outlier-ness of an object. LDOF uses the relative location of an object to its neighbors to determine the degree to which the object deviates from its neighborhood. According to its violation degree the outlier is found. The neighborhood size is chosen only by analyzing the properties of LDOF. Top n technique is also established here to simplify the parameter setting. Also proves that it remains stable even for large range of neighborhood sizes compared to top-n KNN and top-n LOF.

The approach used in this paper [12] is to standardize the computation of an outlier score for each database object and introduces a LoOP (Local Outlier Probability) outlier detection model which is the combination of the idea of local, density-based outlier scoring with a probabilistic, statistically-oriented approach. This method is formulated to provide an outlier score which is in range of [0, 1].This is ease with which outlier score can be computed and interpreted for the comparison of datasets.

Improved K-means technique for outlier detection in high dimensional dataset is explored in [13]. This paper solves the finding of outlier detection by applying the existing Clique method for high dimensional dataset to generate subspace and then the improved k-means algorithm is applied on generated subspaces for identifying outliers.

A hybrid approach for outlier detection in high dimensional data by combining both density based and distance based approach is identified in [14] so that it can take the benefits of both density and distance based clustering methods. DBSCAN (Density based Spatial Clustering Application with noise) a density based technique and k-means are combined in this hybrid approach.

An efficient outlier detection methods has been proposed in [15] which is based on fuzzy c means clustering using Artificial Bee colony algorithm. Fuzzy clustering is used to choose the cluster heads, ABC to select the members of the clusters. When the ABC-FCM algorithm is first performed, it is found that the small clusters are the outlier clusters. Other outliers are then determined based on computing differences between objective functions values when points are temporarily removed from the data set. If a noticeable change occurred on the objective function value, the points are considered.

Reverse nearest neighbors count is recognized in unsupervised distance-based outlier detection [4]. The concept of hubness is introduced here and explores the interplay of hubness and data sparsity. Outier detection methods are implemented based on the properties of antihubs. The relationship between dimensionality, neighborhood size, and reverse neighbors are taken into account for the effectiveness of the mod

When the above outlier detection methods are analyzed, recent research proves that the concept of hubness is extensively tend to be needed for handling the problem of the increase of dimensionality in outlier detection for high dimensional data. The following section discusses characteristics of hubness and hubness based techniques for outlier detection in high dimensional data.

4. HUBNESS BASED OUTLIER DETECTION

4.1 Hubness

High dimensionality of data space resulting from the increase of dimensionality occur in many domains and face challenges for traditional data mining techniques, both in terms of effectiveness and efficiency. Main aspect of the increase of dimensionality is distance concentration, which denotes the tendency of distances between all pairs of points in high dimensional data to become almost equal [16]. Hubness is an aspect of the increase of dimensionality related to nearest neighbors.

Let $D \subset R^d$ be a finite set of n points. For point $x \in D$ and a given distance or similarity measure, the number of k-occurrences, denoted $N^k(x)$, is the number of times x occurs among the k nearest neighbors of all other points in D. For $q \in (0, 1)$, *hubs* are the $\lceil nq \rceil$ points $x \in D$ with the highest values of $N^k(x)$ [4].

It has been shown that hubness, as a phenomenon, appears in high-dimensional data as an inherent property of high dimensionality [4].

4.2 Role of Hubness in machine learning

Recently several papers considered the consequence of hubness in high dimensional data on different data mining and machine learning tasks [16] [18-20]. If we survey the origin of hubness phenomenon, it mainly proves that it is an essential property of high dimensional vector space, and it explore its effect on applications based on measuring distances in vector spaces, notably classification, clustering and information retrieval[18].

[16] Explores the aspect of the increase of dimensionality that is demonstrated through the usage of the concept known as hubness and thoroughly examines the emergence of it and shows that it is an intrinsic property of high dimensional data. Subsequently [16] utilizes the effect of hubness on lot of machine learning tasks belonging to supervised, semi supervised and unsupervised learning families and also discusses the interaction of hubness with dimensionality reduction.

Clustering is the process where similar elements are grouped together. This process becomes hard when the sparse of data is high especially in high dimensional data and does not distinguish the distance between the data points properly. In [20] the authors successfully explore the concept of hubness in clustering high dimensional data by using point hubness scores to guide the search, but choose a centroid-based cluster at the end. With this concept they propose an algorithm called Global hubness-proportional K-means (GHPKM) also compares the proposed algorithm to kernel K-means and one standard density-based method, GDBScan. Finally proves that the proposed algorithm is more robust than the others.

The existence of hubness phenomenon and its applications are comprehensively experimented also in other application fields like music retrieval[21] and classification[16][19][22-24][28-30], image feature representation[25], data reduction[16][26], collaborative filtering[27] and Text retrieval[17].

Paper [22] presents various fuzzy measures for k nearest neighbor classification especially designed for high dimensional data with the usage of the concept known as hubness which express fuzziness of elements appearing in k neighborhoods of other points.

4.3 Role of Hubness in outlier detection

In Recent research, various papers discussed the influence of so called hubness in high-dimensional data on different data mining outlier detection tasks. Papers which concentrates on hubness is given below.

The concept of hubness is observed in [14], which affects reverse nearest-neighbor counts, i.e. k-occurrences. Hubness is demonstrated with the increase of the dimensionality of data, causing the distribution of k-occurrences to become skewed. So hubs very frequently become members of k-NN lists and, at the same time, antihubs become infrequent neighbors.

[31] explores a new important feature of the curse concerning to the distribution of k-occurrences (the number of times a point appears among the k nearest neighbors of other points in a data set) and shows that, as dimensionality increases, this distribution becomes considerably skewed and hub points emerge (points with very high k-occurrences) And also observes the importance of hubness and finds that it is an essential property of high dimensional data.

Identifying unsupervised outliers especially in high dimensional data becomes a tedious procedures. The paper [32] proposes a new approach for unsupervised outlier detection in high dimensional data. Antihub phenomenon is introduced to tackle high dimensional data and an algorithm Recursive Antihub^2 extending Antihub^2 is constructed to reevaluate the outlier score of a point by considering N_k scores of the neighbors of x in addition to $N_k(x)$ itself. Also proves that this new algorithm improves the computational complexity with reduced number of iterations. In paper [4] also, hubness take place major role in finding unsupervised outliers in high dimensional data.

5. CONCLUSION

In this paper, the survey is discussed with different ways in which problem of unsupervised outlier detection for high

dimensional data has been formulated in literature and have attempted to provide an overview of huge literature on various techniques. Implementation of high dimensional data in most of the applications become an issue nowadays due to increase of dimensionality. Hubness is the recently known concept for handling the problems related with the increase of dimensionality and it is understood that it is an intrinsic property of the data where dimension is high. So the role of hubness has been examined in this paper. Also reviewed some of the recent advancements in unsupervised outlier detection for dealing with more complex high dimensional data with the usage of hubness. Outlier detection for high dimensional data is a fast growing emerging technique of today's research and more new methods regarding this technique will emerge in the future.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of this paper.

7. REFERENCES

[1] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data" ACM SIGMOD RECORD February 2002.

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

[3] Zhang, Ji. "Advancements of outlier detection: A survey." ICST Transactions on Scalable Information Systems 13.1 (2013): 1-26.

[4] Miloˇs Radovanovi´c, Alexandros Nanopoulos, and Mirjana Ivanovi´ "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection" IEEE Transactions On Knowledge And Data Engineering, October 2014.

[5] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc 14th ACM SIGKDD Int Conf on Knowledge and Data Mining (KDD), 2008, pp. 444–452.

[6] Ramaswamy, Sridhar, Rajeev Rastogi, and Kyuseok Shim. "Efficient algorithms for mining outliers from large data sets." ACM SIGMOD Record. Vol. 29. No. 2. ACM, 2000.

[7] Knorr, Edwin M., Raymond T. Ng, and Vladimir Tucakov. "Distance-based outliers: algorithms and applications." The VLDB Journal—the International Journal on Very Large Data Bases 8.3-4 (2000): 237-253.

[8] Singh, Amit, Hakan Ferhatosmanoglu, and Ali Şaman Tosun. "High dimensional reverse nearest neighbor queries." Proceedings of the twelfth international conference on Information and knowledge management. ACM, 2003.

[9] Breunig, Markus M., et al. "LOF: identifying density-based local outliers." ACM SIGMOD record. Vol. 29. No. 2. ACM, 2000.

[10] Papadimitriou, Spiros, et al. "Loci: Fast outlier detection using the local correlation integral." Data Engineering, 2003. Proceedings. 19th International Conference on. IEEE, 2003.

[11] Zhang, Ke, Marcus Hutter, and Huidong Jin. "A new local distance-based outlier detection approach for scattered real-world data." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2009. 813-822.

[12] Kriegel, Hans-Peter, et al. "LoOP: local outlier probabilities." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[13] Bhatt, Mr Pushpendra, and Tidake Bharat. "A Review Paper on Improved K-Means Technique for Outlier Detection in High Dimensional Dataset." International Journal of Engineering Sciences & Research Technology, 4(1) January, 2015.

[14] Randive, Neha, et al. "Hybrid Approach for Outlier Detection in High Dimensional Data."Int. Journal of Engineering Research and Applications, Vol. 4, Issue 4(Version 9), April 2014, pp.31-35.

[15] Chalotra, Poonam, and Maitreyee Dutta. "An Outlier Detection Method Based On Artificial Bee Colony Fuzzy Clustering." International Journal of Advanced Research in Computer Science 3.1 (2012).

[16] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, Hubs in space: Popular nearest neighbors in high-dimensional data, J Mach Learn Res 11 (2010), 2487–2531.

[17] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, "On the existence of obstinate results in vector space models," in Proc. 33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2010, pp. 186–193.

[18] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, Nearest neighbors in high-dimensional data: the emergence and influence of hubs, In Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, Canada, 2009. 865–872.

[19] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, Timeseries classification in many intrinsic dimensions, In Proceedings of the 10th SIAM International Conference on Data Mining (SDM), Columbus, OH, 2010, 677–688.

[20] N. Tomaˇsev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "The role of hubness in clustering high-dimensional data,"IEEE T Knowl Data En, vol. 26, no. 3, pp. 739–751, 2014.

[21] J. J. Aucouturier, "Ten experiments on the modelling of polyphonic timbre," Ph.D. dissertation, University of Paris 6, Paris, France, 2006.

[22] N. Tomaˇsev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "Hubness-based fuzzy measures for high-dimensional k nearest neighbor classification," in Proc. 7th Int. Conf. on Machine Learning and Data Mining (MLDM), 2011, pp. 16–30.

[23] Tomašev, Nenad, et al., "A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian kNN," in Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM), 2011, pp. 2173–2176.

[24] N.Tomaˇsev and D. Mladeni´c, "Nearest neighbor voting in High dimensional data: Learning from past occurrences," Computer Science and Information Systems, vol. 9, no. 2, pp. 691–712, 2012.

[25] N.Tomaˇsev, R. Brehar, D. Mladeni´c, and S. Nedevschi, "The influence of hubness on nearest-neighbor methods in object recognition," in Proc. 7th IEEE Int. Conf. on Intelligent Computer Communication and Processing (ICCP), 2011, pp. 367–374.

[26] K.Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and effective instance selection for time-series classification," in Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Part II, 2011, pp. 149–160.

[27] A. Nanopoulos, M. Radovanovi´c, and M. Ivanovi´c, "How does high dimensionality affect collaborative filtering?" in Proc. 3rd ACM Conf. on Recommender Systems (RecSys), 2009, pp. 293–296.

[28] Tomaˇsev N, Mladeni´c D (2013a) Hub co-occurrence modeling for robust high-dimensional knn classification. In: Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol 8189, Springer Berlin Heidelberg, pp. 643–659

[29] Tomašev, Nenad, and Dunja Mladenić. "Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification." Knowledge and information systems 39.1 (2014): 89-122.

[30] Tomašev, Nenad, et al. "Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series." Feature Selection for Data and Pattern Recognition. Springer Berlin Heidelberg, 2015. 231-262.

[31]Radovanović, Miloš, Alexandros Nanopoulos, and Mirjana Ivanović. "Nearest neighbors in high-dimensional data: The emergence and influence of hubs." Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009

[32]Sylvia, J.MichaelAntony, and T.C. Rajakumar "Recursive Antihub^2 Outlier Detection in High Dimensional Data.", Global Journal of Advanced Research, Vol-2, Issue-8 PP. 1269-1274