International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03Issue 12
August 2016

Reverse Nearest Neighbours in Unsupervisedd Distance-Based Outlier Detection

Mr. Pramod.N
PG Scholar, Department of computer science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India
Prof. Honnaraju.B
Assistant professor, Department of computer Science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India

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 havea 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 anti hub-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.

INTRODUCTION:

OUTLIER (anomaly) detection refers to the task of identifying patterns that do not conform to established regular behavior. Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice. The interest in outliers is strong since they may

constitute critical and actionable information in various domains, such as intrusion and fraud detection, andmedical 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, because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods that mainly rely on a measure of distance or similarity in order todetect outliers.

A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless, since distance measures concentrate, i.e., pairwise distances become indiscernible as dimensionality increases,. 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. This somewhat simplified view was recently challenged.

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 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. 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 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 warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method.

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 , 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. Our findings complement the observations from 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,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). Hubness is 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 and also we examine the emergence of antihubs and the way it relates to outlierness of points, also considering low dimensional 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 the system. 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.

EXISTING SYSTEM:

In the Existing system, the system has implemented based on the (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.

PROPOSED SYSTEM:

In the proposed system, the system provides 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.

INPUT DESIGN:

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in

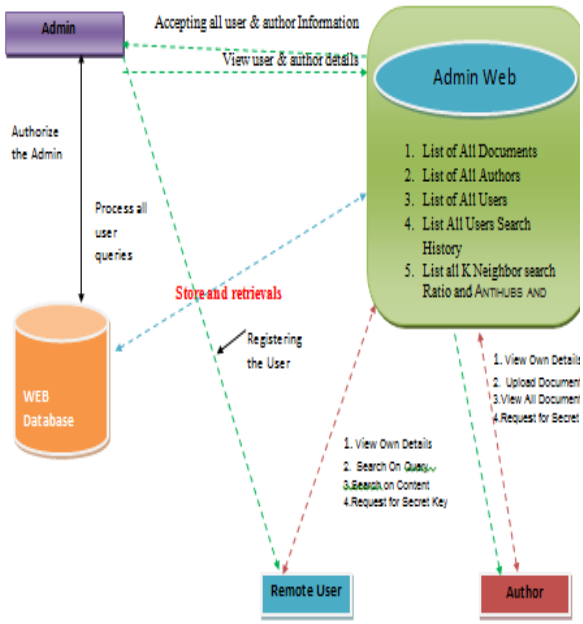
the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN:

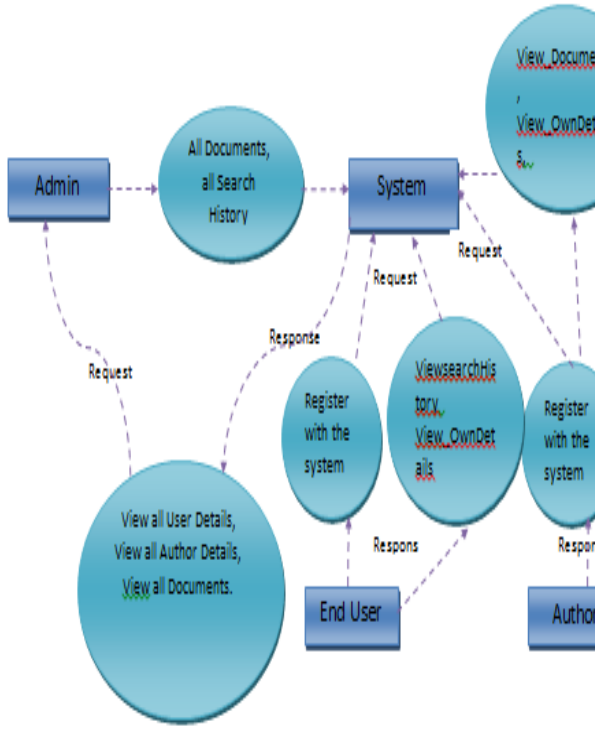
The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.

The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

Architecture Diagram



DATAFLOW DIAGRAM:



CONCLUSION

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 distance/similarity, such as shared-neighbor distances warrant further exploration in the outlier-detection context.

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. Radovanovi_c, A. Nanopoulos, and M. Ivanovi_c, "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.