International Journal of Research in Information Technology

(IJRIT)

www.ijrit.com ISSN 2001-5569

Unsupervised Learning for Outlier Detection from High-Dimensional Datasets

G. Sravan Kumar [1], Kotla Anusha [2]

[1]Associate Professor, Computer Science and Engineering, Sreyas Institute of Engineering & Technology , JNTU Hyderabad, India.

[1] *sravanreddy.golamari@gmail.com*

[2] Student, Computer Science and Engineering, Sreyas Institute of Engineering & Technology, JNTU Hyderabad, India.

anushakotla35@gmail.com[2]

Abstract

Data mining is widely used for mining trends or patterns from different sources of data. When huge amount of data is analyzed, it is possible that the data source has hidden strange patterns. One kind of such patterns is known as outliers. Outlier is a tuple or record that is detached, strange, or away from other records in one way or other. Outlier detection in data mining is used to extract hidden trends that can be interpreted to have business intelligence. In this paper we proposed a framework for detecting outliers. The notion of reverse nearest neighbour is used in our work to detect outliers. Euclidean distance measure is used to find similarity between objects in the search space. We collected datasets from UCI machine learning repository. We built a prototype application to demonstrate the proof of concept. Our results revealed that the outlier detection can help understand useful trends in the data sources for making well-informed decisions.

Index Terms *– Data mining, outlier detection, distance based measure, high-dimensional* **data**

I.INTRODUCTION

Outlier detection has been around for many years in data mining domain. Outlier is an abnormal tuple in a database which can provide valuable information. In many applications outliers provide required business intelligence that can be used to make well informed decisions. For instance, outliers in banking sector can provide details of fraud as other transactions differ from them. Outliers can thus provide valuable information for making strategies to solve issues if any. Our work is related to knowledge discovery from databases (KDD). The overview of KDD is presented in Figure 1.

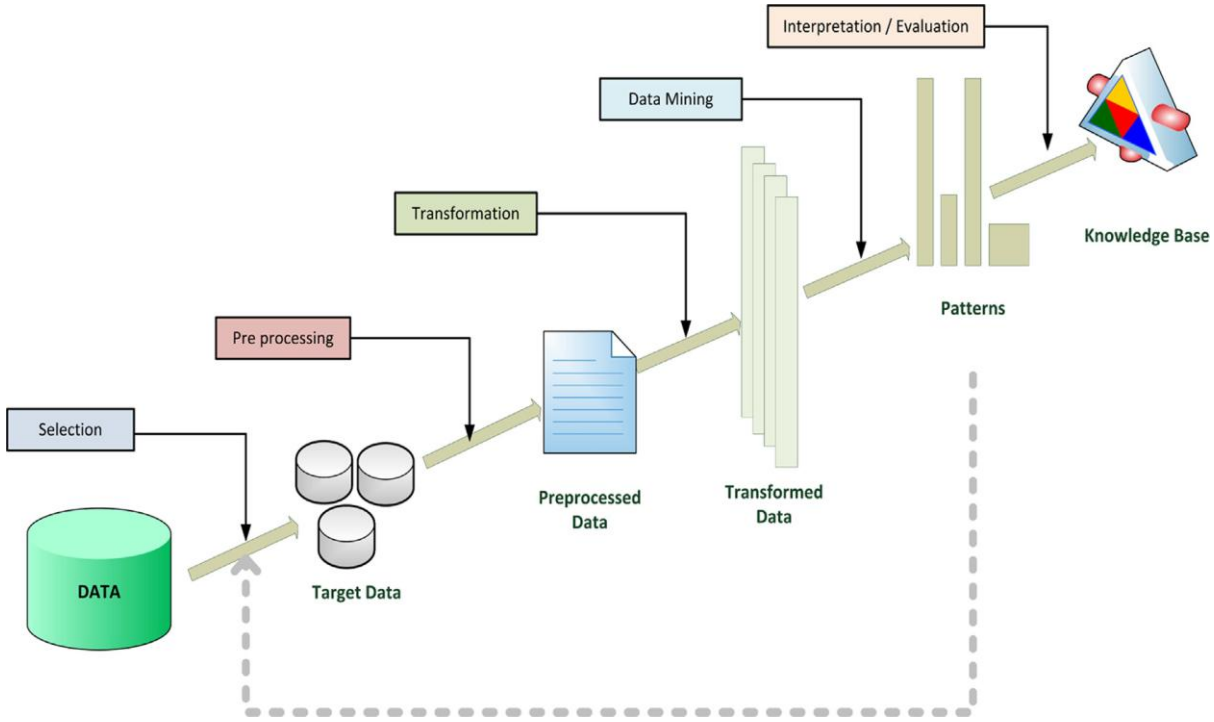

Figure 1 – General Knowledge Discovery in Databases (KDD) overview

As shown in Figure 1, it is evident that knowledge discovery needs a series of steps such as selection of data, pre-processing, transformation, data mining and interpretation. The outcome of the process is business intelligence that can be used to make well informed decisions. In this paper our focus is on the outlier detection. Outliers can provide required business intelligence in order to make strategic decisions in the real world. We proposed a methodology that takes high-dimensional data as input and produces outliers. We built a prototype application to demonstrate the proof of concept. The empirical results reveal that the proposed methodology is useful in finding outliers from high-dimensional data and overcomes the issue of curse of dimensionality.
The remainder of the paper is structured as follows. Section II provides review of literature. Section III presents the proposed system in detail. Section IV presents experimental results while section V concludes the paper.

II.RELATED WORKS

Outlier detection can have different aspects such as point anomalies supervised methods and methods that utilize outlier scope for each point that can be used to have ranking [1]. Most of the outlier detection methods that were found in the literature are using nearest neighbours. They assume that outliers have characteristics that are very far from others or neighbours. Those methods employ distance measure or similarity measure in order to find nearest neighbours. One of the well known similarity measure is known as Euclidean Distance. There are variants of neighbour-based methods that make use of outlier score of given point that is considered as distance to its nearest neighbour [3]. Therefore it is known as k-NN method. The concept of density and inverse density are used in [5] for performing outlier detection. Local Outlier Factor (LOF) is one of the density based methods widely used as explored in [15]. Later on LOF variants came into existence. For instance LOCI [16] makes use of local correlation. Local instance based outlier detection (LDOF) [17] is yet another method which is based on density. Local outlier probabilities (LoOP) are another method explored in [18]. Angle based outlier detection (ABOD) is the focus of the research in [19]. This technique is used to apply for high dimensional data.

The problem of "curse of dimensionality" is the main focus in [20]. They identified three problems with the main problem. They are poor discrimination of distances, irrelevant attributes, and redundant attributes. These problems cause issues with traditional approaches. Unsupervised outlier detection methods are explored in [10] and they identified seven issues such as exponential search space, hubness, data-snooping bias, interpretation, bias of scores, definition of reference sets, and noisy attributes. The notion of reserve nearest neighbours is explored in [21] and [22] for outlier detection process. That feature is also considered for formulating outlier

scores. In this paper our focus is on the reverse nearest neighbour concept for outlier detection. Towards this we proposed a framework that supports outlier detection which is distance based.

III.PROPOSED SYSTEM

In this paper we proposed a methodology for outlier detection from high-dimensional data. The methodology is presented in Figure 2. A list of pre-processing attributes is considered besides taking data from high-dimensional database. We proposed methods such as anti-hub calculations and measure performance in order to identify some data points that appear differently. A distance measure such as Euclidean distance is used to know similarity between data points. The concept of anti-hub is used to make the detection accurate. The "curse of dimensionality" is overcome by using the proposed methodology showing that dimensionality has different impact.

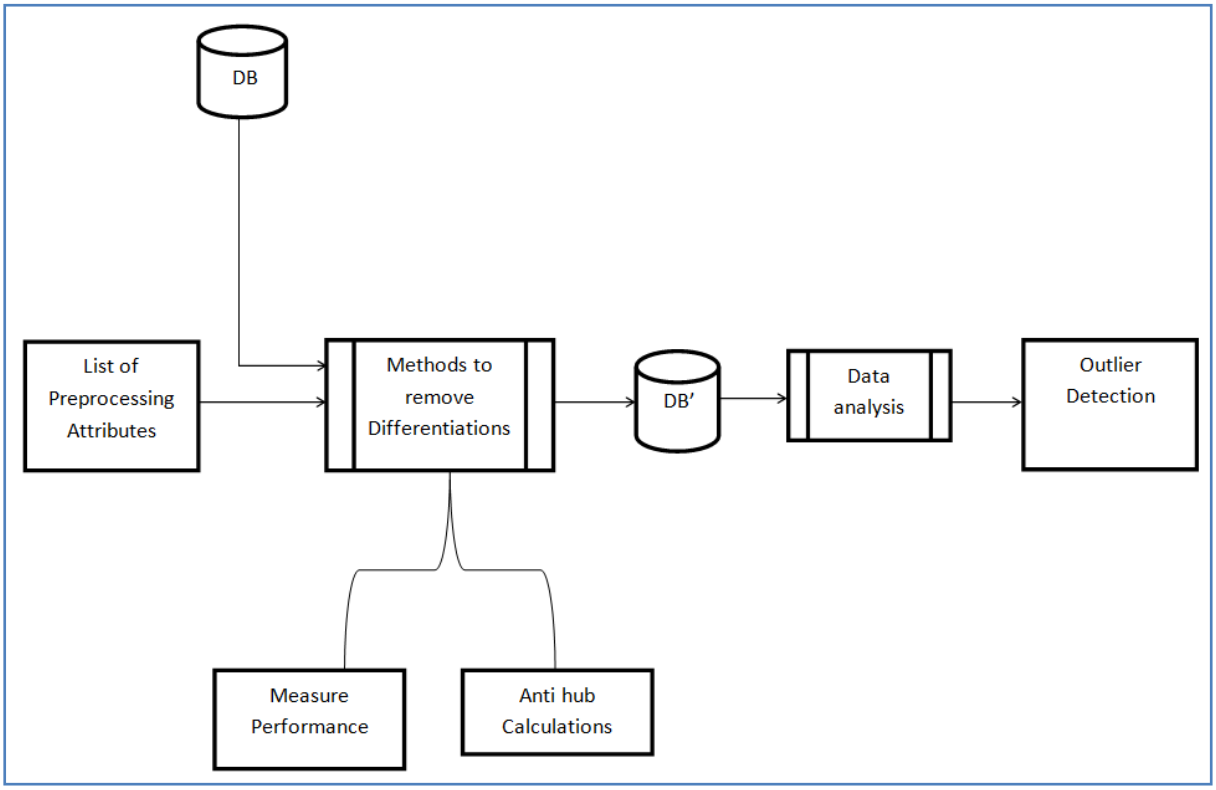

Figure 2 – Proposed methodology for outlier detection

Once the database is transformed by identifying different data points that are likely to become outliers, the data is keep in another database which is subjected to data analysis. The data analysis is the process of finalizing and validating outliers that are obtained from high-dimensional data. In the process we used reverse nearest neighbour concept where unsupervised outlier detection is made more accurate. The outliers obtained from the analysis are presented to end users.

Proposed Algorithm

```
Algorithm: Outlier Detection Algorithm
Inputs: Dataset D, distance measure m, number of neighbours n
Output: Outliers
    Initialize distance measure m
    Initialize neighbours vector N
    Initialize outliers vector O
    Initialize anti hubs vector AH
    For each d in D
      Extract neighbours into N
      For each n in N
        Find differently appearing data points
        Add them to AH
      End For
    End For
    For each ah in AH
        Validate ah
        Add ah to O
    End For
    Display O
```

Algorithm 1 – Outlier detection algorithm

As shown in the algorithm, the procedure or methodology described earlier is presented in the form of step by step procedure which can produce outliers. The outlier detection process takes high-dimensional data as input and generates outliers by following the given procedure.

IV.IMPLEMENTATION

We implemented the proposed system using a prototype application built in Java/JEE platform. The implementation supports multiple users to have data mining operations in a GUI based environment. The users can perform respective activities. The use case of the proposed system is as shown in Figure 3.

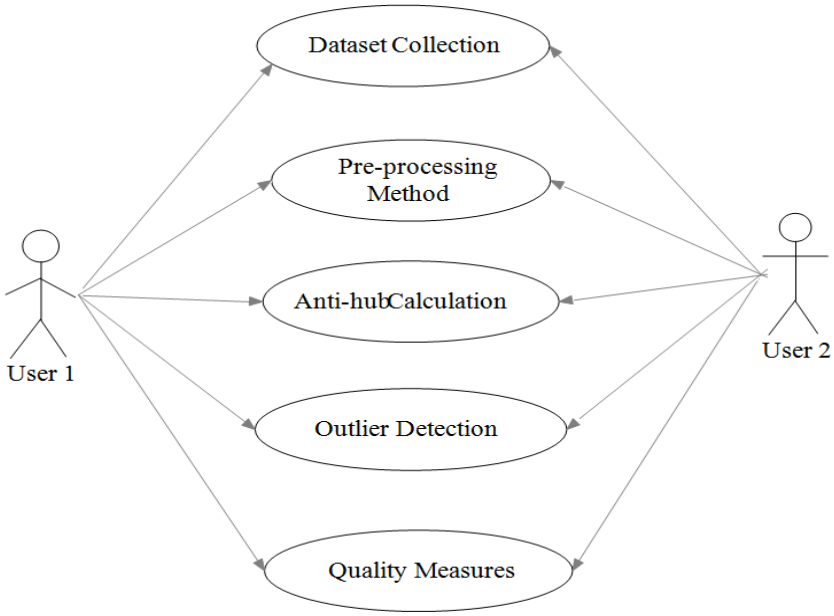

Figure 3 – Use case diagram for proposed system

As shown in Figure 3, the users of the proposed system can perform various operations. The operations include dataset collection which is essential for doing experiments. The datasets collected from UCI machine learning repository are used by users in order to make experiments. Then pre-processing is applied in order to have data to have good quality before subjected to actual outlier detection methodology proposed by us. Then the proposed algorithm is employed in order to find anti-hub results. The results of anti-hub are the data points that show difference from others. The anti-hubs are then validated using analysis as part of outlier detection. In the process distance based measures and reverse nearest neighbour concept were used in order to have high quality outliers that can help in obtaining business intelligence. The sequence of events and the interaction among different objects involved in our implementation is presented in Figure 4.

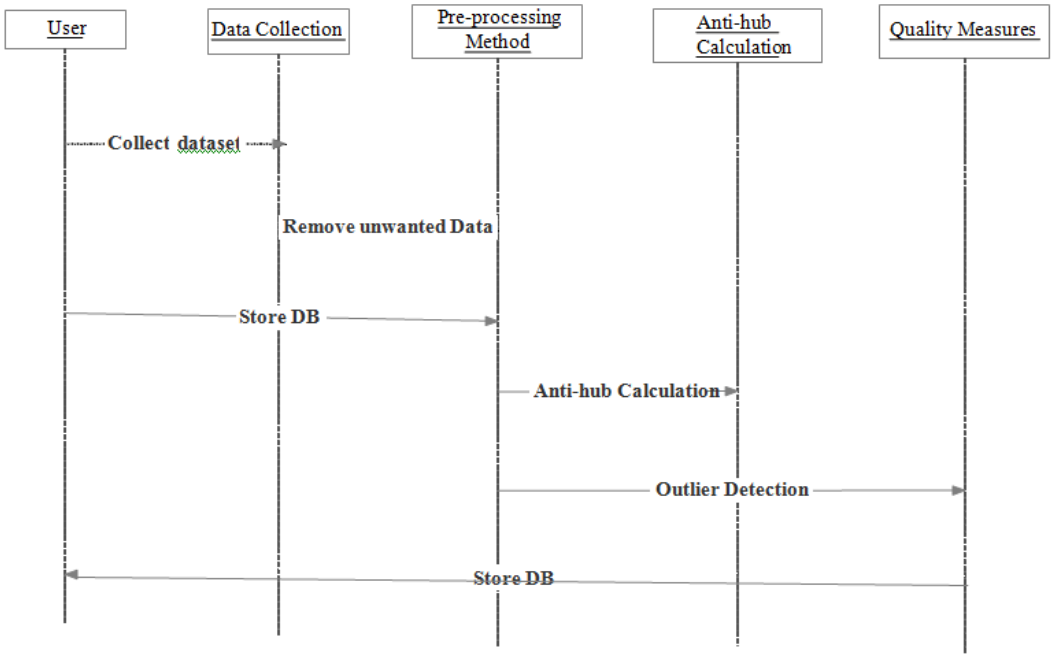

Figure 4 – Sequence diagram showing interactions among objects

As shown in Figure 4, it is evident that there are five objects among which interaction is made before detection of outliers. They are user, data collection object, pre-processing object, anti-hub computation object, and quality measures object. All objects are contributing towards obtaining quality outliers from high-dimensional data.

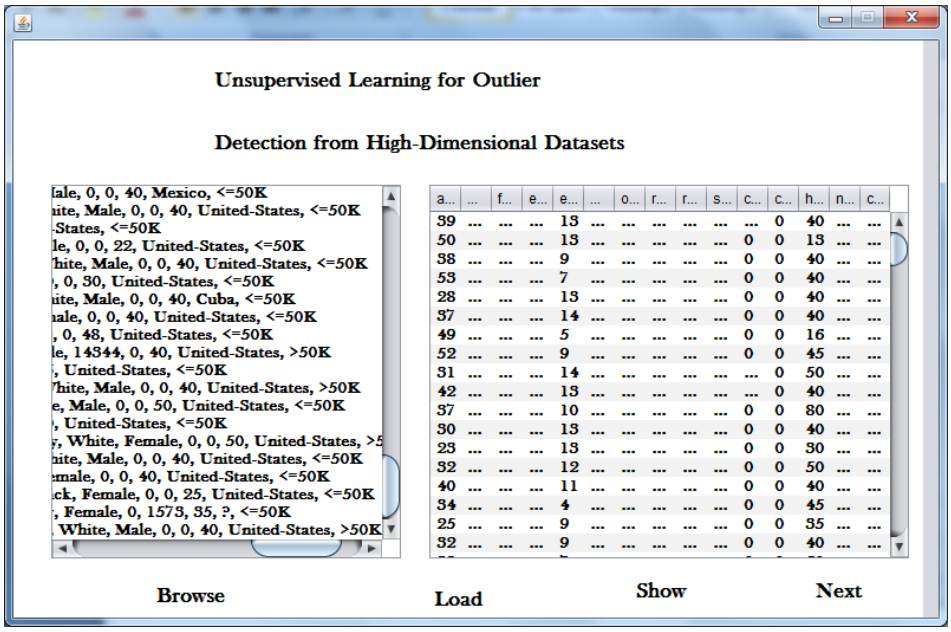

Figure 5 – UI with dataset loaded

As shown in Figure 5, the UI is presented with different controls that can be used to perform operations. For instance browse button is used to load dataset needed.

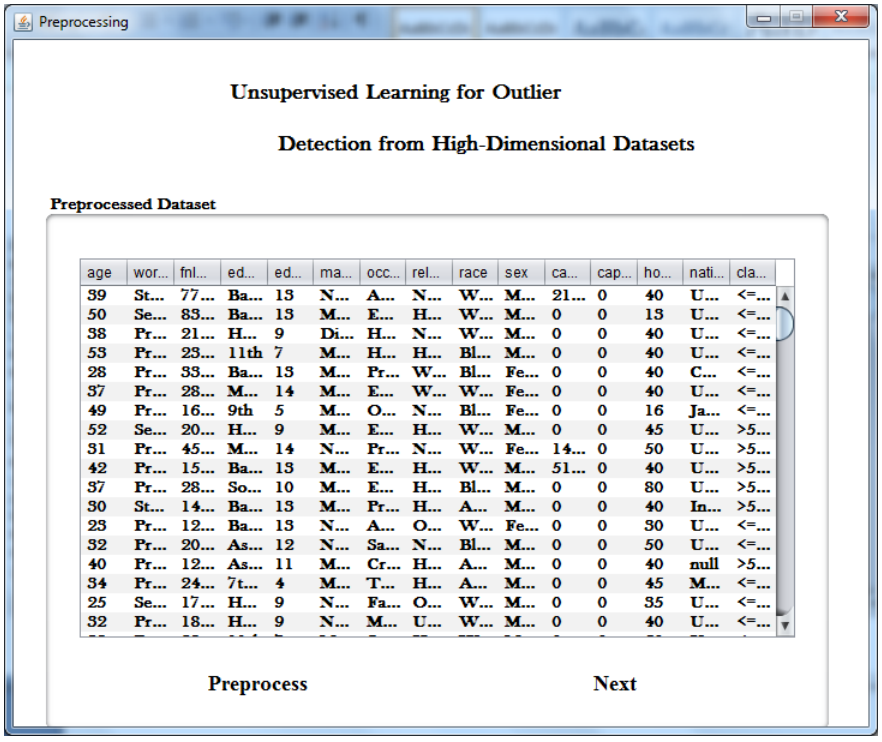

Figure 6 – Intermediate results

As shown in Figure 6, the intermediate results are presented. The dataset is subjected to pre-processing and the data is presented before outlier detection.

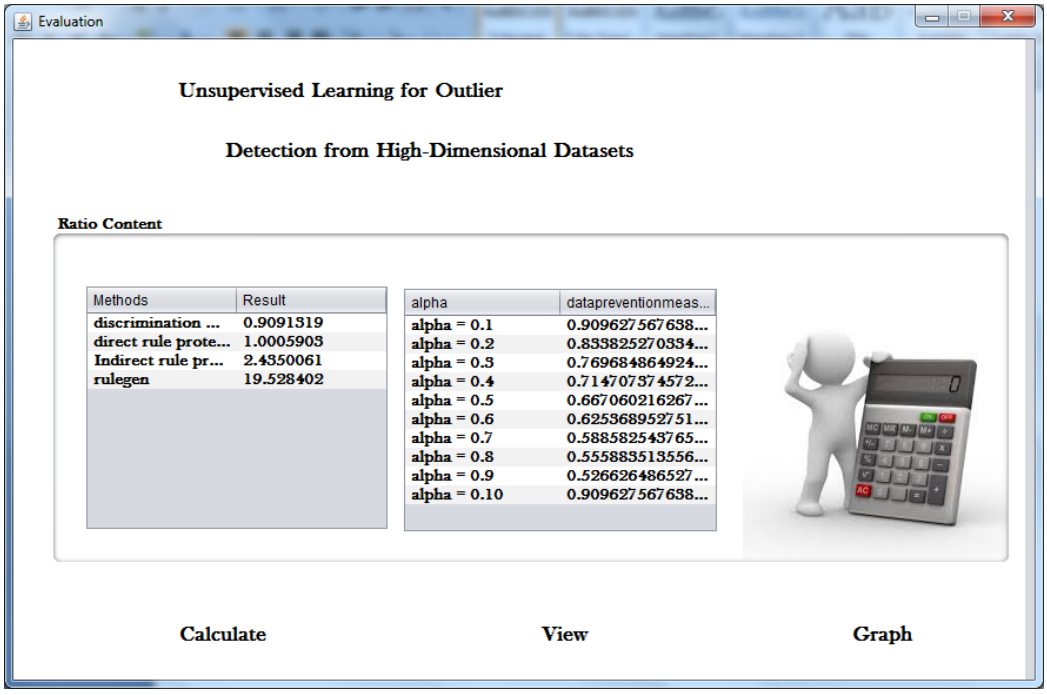

Figure 7 – Outlier detection

As shown in Figure 7, the outlier detection process is underway and the results provide different methods with result besides alpha and other measures. This will lead to outlier detection as discussed in our methodology.

V. EXPERIMENTAL RESULTS

We made experiments with the proposed methodology which is implemented in our prototype application. The application is used to demonstrate the proof of concept. Information loss dynamics in the process of outlier detection is shown in Figure 8.

Figure 8 – Shows kNN search ratio and unmatched data ratio

As shown in Figure 8, it is evident that the results show the kNN search ratio versus unmatched data ratio in terms of information loss. Information loss is less when threshold is increasing for both. Information loss is more for kNN search ratio.

VI.CONCLUSIONS AND FUTURE WORK

In this paper we studied the importance of outlier detection from data sources. We focused on nearest neighbour and reverse nearest neighbour approaches for detecting outliers. This approach is distance based and Euclidean distance measure is used for finding similarity between two objects in the process of finding outliers. Outlier in a database reveals abnormal situation which is very useful in making strategies in real time. For instance it is possible to have credit card fraud detection in banking sector. The utility of outliers in the enterprises is more when they are interpreted and used constructively. We incorporated the concept of anti-hub for the purpose of finding outliers from high-dimensional data. We built a prototype application to demonstrate the proof of concept. The empirical results revealed that the framework is useful. In future we intend to improve the framework and evaluate it for outlier detection in different domains.

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.

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

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

[20] M. E. Houle, H.-P. Kriegel, P. Kr€oger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in Proc 22nd Int. Conf. Sci. Statist.DatabaseManage., 2010, pp. 482–500.

[21] A. Singh, H. Ferhatosmano_glu, and A. ¸Saman Tosun, "High dimensional reverse nearest neighbor queries," in Proc 12th ACM Conf. Inform. Knowl. Manage., 2003, pp. 91–98.

[22] Y. Tao, M. L. Yiu, and N. Mamoulis, "Reverse nearest neighbour search in metric spaces," IEEE Trans. Knowl. Data Eng., vol. 18, no. 9, pp. 1239–1252, Sep. 2006.

Bibilography

G. Sravan Kumar Currently working as Associate professor in Sreyas Institute of Engineering and Technology, Hyderabad. He pursued his B.E and M.Tech (P.hd) in Computer Science and Engineering from JNTUH

Kotla Anusha pursuing M.Tech in Sreyas Institute of Engineering and Technology, Hyderabad. She pursued her B.Tech in Information Technology from JNTUH.