Ranking with Distance based Outlier Detection Techniques: A Survey

Jitendra R. Chandvaniya
M.Tech Student Computer
Engineering Dept
School of Engineering
R K University
Rajkot, India

Rajanikanth Aluvalu
Computer Engineering Dept
School of Engineering
R K University
Rajkot, India

ABSTRACT

Outlier Detection is very much popular in Data Mining field and it is an active research area due to its various applications like fraud detection, network sensor, email spam, stock market analysis, and intrusion detection and also in data cleaning. Here we will study some outlier detection technique which are mainly based on distance-based outlier detection with ranking approach and give some idea about the new technique which we will implement in future.

General Terms

Distance-Based Outlier Detection, Neighborhood, Clustering, Ranking.

Keywords

Distance-Based Outlier Detection, Nearest Neighbor, Ranking and Pruning

1. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD) [1]. Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models [2]. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. Outlier Detection is very much popular in Data Mining field and it is an active research area due to its various applications like fraud detection, network sensor, email spam, stock market analysis, and intrusion detection and also in data cleaning [3]. The importance of outlier detection is due to the fact that outliers in data trans- late to significant information in a wide variety of application domains. For example, an unusual traffic design in a computer network might mean that a hacked computer is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used to detect inconsistent patterns in patient medical records which could be symptoms of a new disease. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse [4]. Outliers can also translate to critical entities such as in military surveillance, where the Presence of an unfamiliar region in a satellite image of enemy area could indicate enemy troop movement or anomalous readings from a space craft would signify a fault in some component of the craft. So, it is very important to detect outlier in large data set very efficiently, there are various methods available for outlier detection like a very simple and easy method is distance-based outlier detection in this method outlier can be detected based on its distance to predefined points in a given data set, find out the nearest neighbor and based on it detect points as outliers [5] but it is very challenging task to develop such method which efficiently detect outliers within less time and which can be apply to large dimensional data set effectively.

2. RELATED WORK

There are various work done in outlier detection using various techniques in this section we will see one by one techniques, Very first algorithm based on distance proposed by Edwin Knorr and Raymond T. Ng *Algorithms for Mining Distance-Based Outliers in Large Datasets* [5] In this author determine two algorithms, first one is a nested loop algorithm that runs in *O(dN*N)*time, on other hand is cell-based algorithm that is linear with respect to *N* where *N* is the number of points of the data set, but exponential in *d* where *d* is the dimensions of the data set. This method efficiently works if *d<=4*, while nested loop algorithm is effective work for small data set to be mined. Sridhar Ramaswamy proposed *Efficient Algorithms for Mining Outliers from Large Data Sets* [6], In this author rank each point on the basis of its distance to *kth* nearest neighbor Outliers detection in this method done using partition-based algorithm, first it partitions the points using clustering algorithm, then prunes those partitions that cannot contain outliers. Wen Jin define *Ranking Outliers Using Symmetric Neighborhood Relationship* [7], here author use measure on local outliers based on a symmetric neighborhood relationship. The proposed measure considers both neighbors and reverse neighbors of an object when estimating its density distribution and then based on it detect top-n outliers. Carlos H. C. Teixeira proposed *An Efficient Algorithm for Outlier Detection in High Dimensional Real Databases* [8] main aim of this algorithm is a fast strategy to estimate the *unusualness* of a record within the database and use a rank-ordered approach to evaluate records. Algorithm partitions the database and ranks the objects that are candidates to be an outlier, it reduce the number of comparisons among objects. Author evaluates different ranking heuristics in a wide-ranging set of real and synthetic databases. . Nguyen Hoang Vu and Vivekanand Gopalkrishnan define Efficient *Pruning Schemes for Distance-Based Outlier Detection* [9] in the first phase, partition the data into clusters, and make an early estimate on the lower bound of outlier scores. Based on this lower bound, the second phase then processes relevant clusters using the traditional block nested-loop algorithm. Here two efficient pruning rules are utilized to quickly discard more non-outliers and reduce the search space, another Approach defined by rajendra is *Distance Based Fast Outlier*

Detection Method [10], in this local distance-based outlier factor used to measure the degree to which an object departs from its neighborhood. Then author use pruning strategy to prune out some of the point using clustering algorithm from the given data set which are probably not the member of outliers.

Srinivasan Parthasarathy define *Distance Based Outlier Detection: Consolidation and Renewed Bearing* [11] In this author use combination of optimization strategies which can give more efficiency, here author use different pruning strategies and ranking strategies and combining them for outlier detection purpose, in this author conclude that combination of ROCO and ANNI is able to achieve one of the best execution times. But for large number of objects like Uniform30D database which has 30 dimensions this algorithm is not applicable.

Rajendra Pamula proposed *An Outlier Detection Method based on Clustering*[12], author use K-means clustering to partition the dataset and LDOF technique to decide final set of outlier. Bhaduri define an *Algorithms for speeding up distance-based outlier detection* [13], in this paper author has introduced sequential and distributed algorithm. Sequential algorithm (iOrca) and Distributed algorithms Door and iDOor, combination with index scheme with distributed processing algorithm works speedily and it is also applicable for the large datasets. Srinivasan Parthasarathy proposed *Locality Sensitive Outlier Detection: A Ranking Driven Approach* [14], here author develop a light-weight ranking scheme that is driven by locality sensitive hashing, which reorders the database points according to their likelihood of being an outlier. Here ranking scheme improves the effectiveness of the distance-based outlier detection process by up to 5-fold. Ms. S. D. Pachgade found Outlier *Detection over Data Set Using Cluster-Based and Distance-Based Approach* [15], in this author use combination of cluster-based and distance based outlier detection This approach deals with only numerical data and it cannot deal with more complex datasets. Yanyan Huang define *A Hybrid Distance-Based Outlier Detection Approach* [16], in this author uses average distance as neighborhood distance, and record the number of data object points within the neighborhood, so that average number of neighbors can be calculated. . Vijay Kumar define algorithm for detection of outlier using cluster-based approach [17] in this approach first they do Partition Around Medoids (PAM) clustering algorithm. After that small clusters are determined and consider as a outlier cluster. Then using absolute distances between the medoid of the current cluster and each one of the points in the same cluster, through this calculation detecting the rest of the outliers (if any).

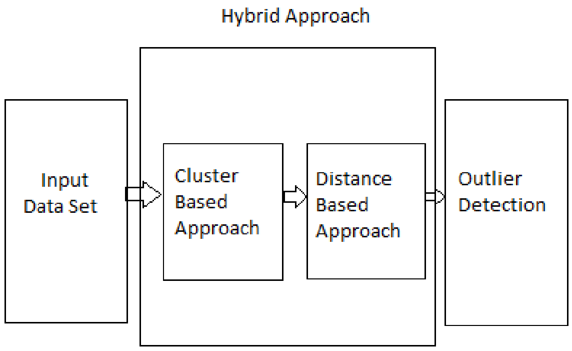

Fig 1: Hybrid Approach of Outlier Detection

H. Huang, K. Mehrotra define *Rank-Based Outlier Detection, here* in this paper author propose new approach for outlier detection, based on this new ranking measure the purpose of this measure is whether a point is *important* for its nearest neighbor. Here author use notation low cumulative rank which says that the point is central. Centrally located point in a cluster has relatively low cumulative sum of ranks because it is among the nearest neighbor of its own nearest neighbors. So, rank measures an object's outlines. Sakthi Nathiarasan used *Outlier Detection based on Utility and Clustering (ODUC) algorithm* [18] here author said that that there are Not all the outliers are essential for business improvement and in other aspect so interestingness of the user while applying any data mining technique . In this system of outlier detection is based on mainly utility and k-means clustering, first it prune the data objects whose utility value is lesser than user's minimum threshold value and the second step is to employ repeated k-means clustering. This algorithm works in two phases in first phase it prune out data objects whose utility is lesser then user's minimum threshold value and in second phase it perform repeated k-means clustering to find and prune the data objects which lies nearer to the centroid of the cluster during each iteration.

3. RANKING APPROACHES

3.1 ROCN (Ranking Object Candidates for Neighbors)

This strategy ranks the order in which neighbors of points are processed so this will reduce the current value of D^k *(p)* faster. It reorders the neighboring clusters so that the search for neighbors proceeds from closest to distant, so we can say that ROCN helps in improving the performance of neighbor while evaluating a given point. Ranking the search for neighbors in a partition level using estimated distances among them. These estimates could be determined by calculating the distances between the centers (centroids) of the partitions or even between MBR structures

As above we have seen various papers in which they have used ROCN strategy, we find that ROCN is significant and able to improve the execution time but when we apply it to large dimensional dataset like Uniform30D this strategy not gave best execution time [6, 7, 8, 11, 14].

3.2 ROCO (Ranking Object Candidate for Outlier)

This strategy used to decide which objects are more likely to be outlier. Aim of this strategy is to focus on the value of D^k *min*. The strategy estimate the kth-NN distance for each Object p. These estimates are then used as a ranking where in objects with greater kth NN distances will be considered first as candidate outliers. Ranking objects that are candidates for outliers this strategy is more likely to density-based heuristic, in which the intuitions that have low-density regions (partitions) tend to contain higher-score objects. We define density as |P|/R(P) , where |P| is the number of objects in partition P, and R(P) is the MBR diagonal length of P. As above we have seen various papers in which they have used ROCO strategy, we find that this strategy is very much use full and give much better result compare to ROCN but still it not much effective with large dimensional dataset [11, 14, 19].

3.3 OTHER METHODS OF RANK BASED OUTLIER DETECTION

3.3.1 LOF (Local outlier Factor) approach

In this approach author proposed that each data point of the given data set should be assigned a degree of outlines and they refer it as the "Local Outlier Factor" (LOF) [20, 21] of the data point and it is calculated as given below.

$$L_k(p) = [\sum o\epsilon Nk(p) \frac{lk(0)}{lk(p)} lk(o)/|Nk(p)|]$$

$L_k(p)$ is calculated for selected values of k in a pre-specified range, max $L_k(p)$ is retained , and a p with large LOF is declared to be outlier [20]

3.3.2 COF (Connectivity-based outlier factor) approach

This is modified technique of LOF this says that when a cluster and a neighboring outlier have similar neighborhood densities. COF can be measure as Given below formulae.

$$COF_k(p)=[A_{Nk(p)}(P)/[\sum_{o\epsilon Nk(p)} A_{Nk(p)}(o)/|N_k(p)|]^{-1}$$

Here larger value of $COF_k(p)$ denote higher Possibility that p is an outlier [20.21].

3.3.3 INFLOW (INFLuential measure of outlier by symmetric relationship approach)

This method is based on the concept of symmetric neighborhood relationship in this considers neighbors and reverse neighbors of a data point when estimating its density distribution. Using this outlier cab be detected as given below method.

$$INFLO_k(p) = 1/den(p) * \sum_{o\varepsilon ISk(p)} den(o)/|(IS_k(p))|$$

Where den (p) = $1/d_k(p)$ [20,21].

Here in below Table 1 experimental results shown, this results shows the comparison between LOF, COF and INFLO algorithms, Nrc denotes the number of outliers within top m instances, Pr represent precision, Re represent recall, and RP represent Rank-Power. Synthetic Dataset is used which contains 74 instances, including six planted outliers and has four clusters of different densities consisting of 36, 8, 8, 16, instances. Also used four different values of k.

Table 1. Comparison between LOF, COF and INFLO [21]

m	**LOF**				**COF**				**INFLO**			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	**5**	**1.00**	**0.83**	**1.00**	**5**	**1.00**	**0.83**	**1.00**	**5**	**1.00**	**0.83**	0.882
10	**6**	**0.60**	**1.00**	0.95	**6**	**0.60**	**1.00**	0.95	**6**	0.60	**1.00**	0.875
15	**6**	**0.40**	**1.00**	0.95	**6**	**0.40**	**1.00**	0.95	**6**	**0.40**	**1.00**	0.875
30	**6**	**0.20**	**1.00**	0.95	**6**	**0.20**	**1.00**	0.95	**6**	**0.20**	**1.00**	0.875
m	**LOF**				**COF**				**INFLO**			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	**5**	**1.00**	**0.83**	**1.000**	**5**	**1.00**	**0.83**	**1.000**	**5**	**1.00**	**0.83**	**0.882**
10	**6**	**0.60**	**1.00**	0.913	**6**	**0.60**	**1.00**	0.913	**6**	**0.60**	**1.00**	**0.995**
15	**6**	**0.40**	**1.00**	0.913	**6**	**0.40**	**1.00**	0.913	**6**	**0.40**	**1.00**	**0.995**
30	**6**	**0.20**	**1.00**	0.913	**6**	**0.20**	**1.00**	0.913	**6**	**0.20**	**1.00**	**0.995**
m	**LOF**				**COF**				**INFLO**			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	3	0.60	0.50	**1.000**	**4**	**0.80**	**0.67**	**1.000**	3	0.60	0.50	**1.000**
10	4	0.40	0.67	0.667	**5**	**0.50**	**0.83**	0.789	4	0.40	0.67	0.833
15	4	0.27	0.67	0.667	**5**	**0.33**	**0.83**	**0.789**	4	0.27	0.67	0.833
30	4	0.13	0.67	0.667	5	0.17	0.83	**0.789**	5	0.17	0.83	0.8360
m	**LOF**				**COF**				**INFLO**			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	3	0.60	0.50	**1.000**	**4**	**0.80**	**0.67**	**1.000**	2	0.40	0.33	**1.000**
10	4	0.40	0.67	0.625	**5**	**0.50**	**0.83**	0.789	4	0.40	0.67	0.526
15	**5**	**0.33**	**0.83**	0.484	**5**	**0.33**	**0.83**	**0.789**	4	0.27	0.67	0.526
30	5	0.17	0.83	0.484	5	0.17	0.83	**0.789**	5	0.13	0.67	0.526

4. CONCLUSION

As we have seen above different approaches are there for outlier detection but still not them all satisfied fully for outlier detection which can work for large dimensions and take less execution time. LOF gives one of the best performance when there is number of instance of dataset D is less than 30, while COF gives best performance when instance size is less than 10, while the value of k (number of outliers) is equals to 10 INFLO gives best performance than other algorithm but as increase the value of k it becomes less effective for outlier detection.

ROCN and ROCO we have seen both of them they are suitable for small dimension dataset and give best result with small instance and less dimensional. So my future work is to develop an Adaptive Parallel Algorithm for Outlier detection using Ranking Strategies. In this proposed algorithm we will apply parallel technique on ranking and pruning to detect outliers. We hope that our Proposed Algorithm will overcome the existing issues of outlier detection using ranking strategies.

5. ACKNOWLEDGMENTS

Our thanks to the school of engineering, R K University.

6. REFERENCES

[1] LUKAS A. KURGAN and PETR MUSILEK, *Department of Electrical and Computer Engineering, University of Alberta,"* A survey of Knowledge Discovery and Data Mining process models*",* The Knowledge Engineering Review, Vol. 21:1, 1–24. 2006, Cambridge University

[2] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok Computer Science Department, Columbia University," A Data Mining Framework for Building Intrusion Detection Models",IEEE-1999[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] VARUN CHANDOLA University of Minnesota, "Outlier Detection: A Survey"

[4] Karanjit Singh and Dr. Shuchita Upadhyaya, Department of Computer Science and Applications, Kurukshetra University Kurukshetra, Haryana, India "Outlier Detection: Applications And Techniques", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[5] E. Knorr and R. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In Proceedings of VLDB'98, pages 392–403, 1998.

[6] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. In Proceedings of SIGMOD'00, pages 427–438, 2000

[7] Wen Jin1, Anthony K. H. Tung2, Jiawei Han3, and Wei Wang4," Ranking Outliers Using Symmetric Neighborhood Relationship"

[8] Carlos H. C. Teixeira, Gustavo H. Orair, Wagner Meira Jr, Srinivasan xParthasarathy "An Efficient Algorithm for Outlier Detection in High Dimensional Real Databases"-2008

[9] Nguyen Hoang VuandVivekanand Gopalkrishnan," Efficient Pruning Schemes for Distance-Based Outlier Detection", W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 160–175, 2009.

[10] Rajendra Pamula,Jatin, "Distance Based Fast Outlier Detection Method", India Conference (INDICON), 2010 Annual IEEE.

[11] Gustavo H. Orair Carlos H. C. Teixeira Wagner Meira Jr., Ye Wang Srinivasan Parthasarathy "Distance Based Outlier Detection: Consolidation and Renewed Bearing" *Proceedings of the VLDB Endowment*, Vol. 3, No. 2 Copyright 2010 VLDB Endowment 21508097/10/09... $ 10.00

[12] Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi," An Outlier DetectionMethod based on Clustering", 2011 Second International Conference on Emerging Applications of Information Technology © 2011 IEEE.

[13] Kanishka Bhaduri, **"**Algorithms for speeding up distance-based outlier Detection", *SIGKDD* '11 San Diego, CA, USA

[14]Ye Wang ,Srinivasan Parthasarathy, Shirish Tatikonda,"Locality Sensitive Outlier Detection: A Ranking Driven Approach" *Computer Science and Engineering Department, The Ohio State University, OH, USA,2011*

[15] Ms. S. D. Pachgade, Ms. S. S. Dhande, " Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach", Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[16] Yanyan Huang, Zhongnan Zhang*, Minghong Liao, Yize Tan, Shaobin Zhou," A Hybrid Distance-Based Outlier Detection Approach,2012 International Conference On System and Informatics(ICSAI 2012),987-1-4673-0199-2/12/$31.00 © 2012 IEEE.

[17] Vijay Kumar, Sunil Kumar, Ajay Kumar Singh," Outlier Detection: A Clustering-Based Approach", International Journal of Science and Modern Engineering (IJISME), ISSN: 2319-6386, Volume-1, Issue-7, June 2013

[18] Sakthi Nathiarasan A ,*M. E- Student, "* Algorithm for Outlier Detection Based on Utility and Clustering (ODUC)*", Department of Computer Science and Engineering Adhiyamaan College of Engg , Hosur, India,* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013

[19] M. Wu and C. Jermaine. A bayesian method for guessing the extreme values in a data set? In VLDB '07: Proceedings of the 33rd international conference on Very large data bases, pages 471–482. VLDB Endowment, 2007

[20] H. Huang, K. Mehrotra, C. K. Mohan," Rank-Based Outlier Detection", *Electrical Engineering and Computer Science Technical Reports.* Paper 47,2011

[21] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, \Lof: Identifying density-based local outliers," In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 93{104, 2000.