ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th - 18th July -2014

Detection and Deletion of Outliers from Large Datasets

Nithya.Jayaprakash[1], Ms. Caroline Mary[2]

M. tech Student, Dept of Computer Science, Mohandas College of Engineering and Technology, India[1]

Assistant Professor, Dept of Computer Science, Mohandas College of Engineering and Technology, India[2]

ABSTRACT: The paper proposes a method for detecting and deleting distance based outliers in very large data sets. This is based on the outlier detection solving set algorithm. This method introduces parallel computation so as to save more time and having excellent performance. First, weights are assigned to each of the data in the data sets. Based on the weights outliers from all the data sets are obtained by using the distance based method and finally they are all deleted. By deleting the outliers, it increases the space for storing more data.

KEYWORDS: Outliers, Distance based outliers, Unsupervised approaches.

I. INTRODUCTION

An Outlier is an observation that is distinct from the rest of the data or an outlier is an exception in a large multi dimensional data set. As an example, consider a list of hockey players. The exception or outlier to that list is the person who plays hockey well or the person who doesn't play hockey well. In a data base there exist several data sets, in which there exist several outliers. Our main aim is to detect those outliers. There exist several ways for finding an outlier. They are statistical based method, density based method, distance based method etc. Here we use the distance based method [1]. Based on the distances to its neighbors [3] outliers are detected. A data value that seems to be out of place with respect to the rest of the data is said to be an outlier. In our concept we are assigning weights to each and every data. Based on the weights a threshold value is set. If any data having weights greater than the threshold value it is considered as an outlier. Thus, outliers from all the data sets are obtained and finally they are all deleted in order to add more data to the datasets.

Outlier detection is a data mining task. Data mining is the process of extracting valid, previously unknown and actionable information from a large data base. The main goal of outlier detection is to isolate the observations which are dissimilar from the rest of the data. This task has practical applications in several fields such as fraud detection, intrusion detection, data cleaning, medical diagnosis etc [1]. Data mining includes supervised and unsupervised approaches. Here we are using the unsupervised approach. With unsupervised learning it is possible to learn larger and more complex models than with supervised learning. In unsupervised learning, the learning can proceed hierarchically from the observations into ever more abstract levels of representation. Each additional hierarchy needs to learn only one step and therefore the learning time increases linearly in the number of levels in the model hierarchy.

A single iteration of the main cycle of the sequential Solving Set algorithm can be efficiently translated according to a parallel/distributed implementation [1]. The outlier detection solving set is a subset S of the data set D that includes a sufficient number of objects from D to allow considering only the distances among the pairs in S and D to obtain the top-n outliers [1]. The distance based method distinguishes an object as outlier based on the distances. If there exist any abnormality in the distance it is considered as an outlier. So that we use the outlier detection solving set algorithm. After detecting outliers from all the datasets the outliers are deleted. This increases the space for storing more data in the data sets. Deletion of outlier data is a controversial practice owned by many scientists and science instructors, while

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th - 18th July -2014

mathematical criteria provide an objective and quantitative method for data rejection, they do not make the practice more scientifically or methodologically sound, especially in small sets or where a normal distribution cannot be assumed. Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known. An outlier resulting may be excluded.

II. RELATED WORK

Identifying And Eliminating Mislabeled Training Instances by Carla E. Brodley, Mark A Friedl [4] presents a new approach to identifying and eliminating mislabeled training instances. The main goal of this method is to improve the quality of the training data. Here they use filters. All the data's are passed through the filters. If any data contains error it will not be passed by the filter. In this way the classification accuracy can be increased.

Hung and Cheung [5] presented a parallel version, called PENL, of the basic NL algorithm [6], [1]. A distance-based outlier is a point for which less than k points lie within the distance N in the input data set. The main problem of this method is that it does not provide an approximate value for the distance N.

Detecting Distance Based Outliers in Streams of Data by Fabrizio Angiulli, Fabio Fassetti [6] proposed a method for detecting distance-based outliers in data streams. Here outliers are detected from large data streams. A data stream is a large volume of data coming as an unbounded sequence. The stream portions that falls between last landmark and the current time are considered. That is, the data's in between certain time intervals are considered. The data's are considered based on its characteristics. The main disadvantage of this method is that the data's with different characteristics are detected even if they belong to the same task.

III. THE PROPOSED OUTLIER DETECTION METHOD

In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory or it may be that some observations are far from the centre of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. Outliers are detected by using the outlier detection solving set algorithm [1]. In every data set there exist some exceptional data's. First, a graph is created according to the data set. Then weights are assigned to each of the data's in the graph. Algorithm proceeds by assigning weights to the data's in each row. According to the weight a threshold value is set. If any data that exceeds the threshold value, then that row is considered as an outlier and it is removed from the data set. In this way outliers from all the data sets are obtained.

After deletion of the particular row more data's can be added to the data set. Automatically weights are assigned. And the same procedure will be repeated. The main advantage of this method is that there will be no delay for searching the data. Thus there will not be any memory wastage. In the existing system, the outliers are only mined or detected. They are not deleted. So if we search in the same data set the same outliers will be obtained. These introduce memory wastage because more outliers are stored rather than data. So it becomes a memory consumption process.

The proposed method can be divided into three main criteria's as shown in fig.1.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th - 18th July -2014

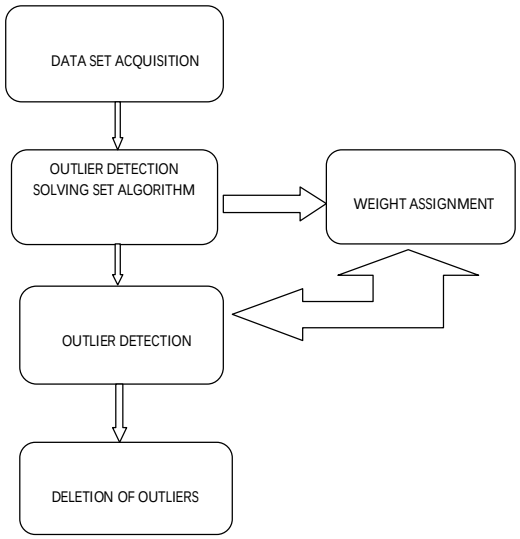

Figure.1.1. System Overview.

A. Design of Data sets

A dataset (or data set) is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the dataset. The dataset may comprise data for one or more members, corresponding to the number of rows. The term dataset may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Here we are using the medical data sets. The medical data set consist of 800 rows. These medical data sets include the details of some diseases and also the areas where these diseases can be occurred. First, we load all the data's to our data base and then imported to our environment.

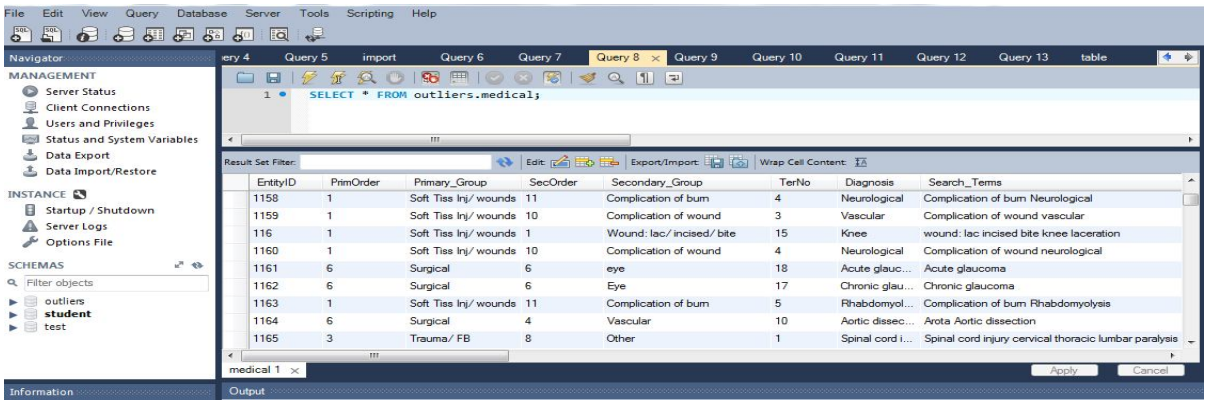

Figure. 2.2. Medical Data set

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th - 18th July -2014

B. Design of Algorithm

This part includes the weight assignment. Here single thread is used. Weights are assigned to each of the fields in the row. Then the maximum weight is considered. That maximum weight is subtracted with each of the weight in the particular row. Then that maximum weight is divided by two. Thus obtain the threshold value. If any if the subtracted value exceeds the threshold value, then that row is considered as an outlier and this process is repeated for each row in the data set. In order to decrease the processing time, we can also use multiple threads for detecting outliers.

C. Deletion Of Outliers

By using the same procedure outliers from all the data sets can be obtained and at last they are all deleted. That increases space to store more data. So if we search in the same data sets same outliers will not be obtained.

IV. EXPECTED OUTPUT

I am expecting more accuracy with no error rate than the existing method. This method can also be efficiently translated into parallel/distributed systems. It provides high performance with in less time.

IV.CONCLUSION

Identifying distance based outliers is an important data mining activity. It can be used in several applications such as fraud detection and also in military operations. This method provides vast time savings and also having high performances. Although it is a learned method, it is having high accuracy. Data's are detected by using the simple mechanisms. By deletion it increases the capacity of the data base. And there will not be any loss of original data. That is, detection and deletion can be done without any side effects.

V. ACKNOWLEDGEMENTS.

The authors would like to thank the teaching faculties of Mohandas College of Engineering.

REFERENCES

[1] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, "Distributed Strategies For Mining Outliers in Large Data Sets," *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 7, July 2013.
[2] J. Han and M. Kamber, Data Mining, Concepts and Technique Morgan Kaufmann, 2001.
[3] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," *IEEE Trans. Knowledge and Data Eng.*, vol. 18, no. 2, pp. 145-160, Feb. 2006.
[4] Carla E.Brodley , Mark A Friedl, "Identifying And Eliminating Mislabeld Training Instances.
[5] E. Hung and D.W. Cheung, "Parallel Mining of Outliers in Large Database," Distributed and Parallel Databases, vol. 12, no. 1, pp. 5- 26, 2002.
[6] E. Knorr and R. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," Proc. 24rd Int'l Conf. Very Large Data Bases (VLDB), pp. 392-403, 1998.
[7] Fabrizio Angiulli , Fabio Fassetti ," Detecting Distance Based Outliers In Streams Of Data".
[8] A. Koufakou and M. Georgiopoulos, "A Fast Outlier Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes," Data Mining Knowledge Discovery, vol. 20, pp. 259-289, 2009.
[9] F. Angiulli and F. Fassetti, "Dolphin: An Efficient Algorithm for Mining Distance-Based Outliers in very Large Datasets," *Trans. Knowledge Discovery from Data*, vol. 3, no. 1, article 4, 2009.
[10] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High- Dimensional Data Sets," *IEEE Trans. Knowledge and Data Eng.*, vol. 2, no. 17, pp. 203-215, Feb. 2005.
[11] A. Asuncion and D. Newman, UCI Machine Learning Repository 2007.
[12] S.D. Bay and M. Schwabacher, "Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule," *Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, 2003.
[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Survey*, vol. 41, no. 3, pp. 15:1-15:58,

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th - 18th July -2014

2009.

[14] H. Dutta, C. Giannella, K.D. Borne, and H. Kargupta, "Distributed Top-K Outlier Detection from Astronomy Catalogs Using the DEMAC System," *Proc. SIAM Int'l Conf. Data Mining (SDM)*, 2007.

[15] A. Ghoting, S. Parthasarathy, and M.E. Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets," *Data Mining Knowledge Discovery*, vol. 16, no. 3, pp. 349-364, 2008.

[16] S.E. Guttormsson, R.J. Marks, M.A. El-Sharkawi, and I. Kerszenbaum, "Elliptical Novelty Grouping for on-line Short- Turn Detection of Excited Running Rotors," *Trans. Energy Conversion*, vol. 14, no. 1, pp. 16-22, 1999.

[17] Y. Tao, X. Xiao, and S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 394-403, 2006.

[18] Y. Tao, X. Xiao, and S. Zhou, "Mining Distance-Based Outliers from Large Databases in Any Metric Space," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD*), pp. 394-403, 2006.[19] www.google.com