

Weighted Density based Error Optimization for Classical Dataset

Sumeet V. Shingi

ME Student

Department of Computer Science & Engineering

PES College of Engineering, Aurangabad, Maharashtra, India

Abstract— Error optimization has many significant applications being a branch of data mining. In statistics, there are different methods which are brought –up into the pathway of detection of errors independently. When dealing with very large classical datasets, there are various obstacles which are not present on small scale. With respect to numerical data, exiting methods cannot be applied directly since difference in between datasets with respect to their classes is defined. This paper contributes its efforts for evaluating error optimization algorithm for classical dataset based on given density. Proper formulation of uncertainty of objects and its density in every attributes provides weighted density. With this exercise, understanding of its different applications will be more sophisticated for beginners.

Key words: Error optimization, Deviation based method, Methods of outliers

I. INTRODUCTION

An error may be a piece of data or observation that deviates drastically from the given norm of dataset. An error may be caused by unusual behavioral habitat of an element but it can be indicated as measurement error or that the given dataset has a heavy tailed distribution. Here is a simple scene of example in outlier detection, a measurement process consistently produces readouts between 1 to 10 but in some rare cases we get measurements of greater than 20. These are measurement beyond the norm that is what we called as an error since it lies the normal distribution.

The process of obtaining a valid and comprehensive information from large datasets is known as work done for data mining. Further, this can be implemented for organizational decision making. Various different types of problems (eg. data redundancy, the value of attributes is not specific, data is not complete and error) exist in mining particular information in huge datasets. D. Hawkins defines outlier as: “An outlier is an observation that deviates so much from other observations to arouse suspicion that it was generated by a different mechanism.”

II. RELATED SURVEY

A. Types of outliers

Following are three different types of outliers:

- 1) Point
- 2) Contextual
- 3) Collective

1) Point

Suppose there is an element which have differential behavioral nature than that of the rest of elements present in a particular group. Such an element is treated as a point error with respect to that of other elements. Example, if we consider credit card fraud detection with data set corresponding to irrespective of everyone’s credit card transactions assuming data definition by only one feature: amount spent. A transaction for which the amount spent is

very high compared to the normal range of expenditure for that person will be a point outlier.

2) Contextual

Error obtained within a particular context can be determined as contextual error.

Here, there are two attributes observed, as below:

- 1) Neighboring attribute
- 2) Behavioral attribute

1) Neighboring attribute

The role of neighboring attribute is to determine its neighborhood for that particular instance. Example, in spatial data sets, the longitude and latitude of allocation are the contextual attributes.

2) Behavioral attribute

The role of behavioral attribute is to determine non-neighboring features of particular instance, which is exactly opposite to previous attribute. Example, the amount of rainfall occurred at particular location is behavioral attribute with respect to that of entire world.

Let us consider an example in relation of contextual outlier which can be found in the credit card fraud detection with contextual as time of purchase. Suppose an person usually has a weekly shopping bill of Rs.1500 except during the week in which Diwali comes, when it reaches Rs.10000. A new purchase of Rs.10000 in a week in October will be considered as a contextual outlier, since it does not states the normal behavior of the individual in the context of time.

3) Collective

Collective outliers can be observed in the sequence data, graph data and spatial data. It should be noted that point outliers can occur in any data set whereas the collective outliers can occur only in data sets in which data instances are related. If a collection of related data instances consist of errors within entire data set, than it is termed as a collective outlier. The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous.

B. Methods of outliers

In order to work on larger datasets, different study work is bought into the action to conduct outliers detection. Researchers and many more organizations are per forming different experiments to optimize the errors. The ancient work done considers numerical data and a prior knowledge of distribution.

1) Statistical Distribution based Method

The functionality of this method depend upon the data distribution. In this method, errors are found with respect to the model which is been implemented, whether it is Normal or Poisson, for best fit.

The two demerits of this method is: all of them are distribution based and invariate. It is really difficult to find an attribute of a Normal, Poisson, Gamma, etc. distributions and behavior in multidimensional datasets, respectively.

2) Distance based Method

In order to overcome upon the demerits of statistical distribution method, Knorr introduced the distance based method.

“An object O in a dataset is a outlier (p,D) if at least fraction p of the objects in T lies greater than distance D from O.”

The time required for execution of this method is at least quadratic with respect to the number of objects. Cell based algorithm has a linear complexity with respect to N but exponential with respect to k. It possesses a complexity of $O(k.N^2)$, where k is the dimensionality and N represents number of objects present in dataset.

For instance, a point with few neighbors within a distance can be regarded in some sense as being a stronger outlier than with more neighbors. The weakness with this method is that they are not powerful to cope with certain process with different densities.

3) Density Based Method

The density based method produces the density distribution of the data as output and identify outliers within those low lying density regions. Another idea of this method of error optimization is that it assesses the degree to which an object is an outlier.

4) Deviation Based Method

In order to identify exceptional objects, deviation based method does not use statistical test or distance based measures. The secret of identifying outliers is to examine the prime characteristics of objects in a group. The objects that “deviate” or “free from surface” are considered as outliers of present dataset.

III. PROPOSED WORK

Previous work done shows the impractical type of nature for the several applications under which they have gone. The major reason for the low resultant ratio of low effectiveness to low efficiency was their rich dimensions and poor size of data group. There are also various types of manually defined parameters required which could not be defined without the prior knowledge of complete dataset

Therefore, generation an effective algorithm for error optimization can be come into frame by adding the average density in previous work which contributes for estimation of weighted density.

IV. CONCLUSION

Nowadays, optimizing an error has beamed as an extreme branch in data mining for modernized applications. The underlined word in this paper is the weight calculated with respect to the density of various number of objects present in the dataset. This paper provides an easy way, with future aspect, to evaluate the rate of efficiency.

REFERENCES

[1] J. Han and M. Kamber, DATA MINING: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.
[2] UCI Machine Learning Repository 2012 <http://archive.ics.uci.edu/ml/datasets.html>

[3] A simple and effective outlier detection algorithm for categorical data by Xingwang Zhao, Jiye Liang, Fuyuan Cao in Springer-Verlag Berlin Heidelberg 2013
[4] Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering by Yogita Thakran, Durga Toshniwal in IEEE conference 2012.
[5] Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput Surv 41(3):Article 15.
[6] Z. He, X. Xu, and S. Deng, “Discovering cluster based local outliers,” Pattern Recognition Letters, vol. 2003, pp. 9–10, 2003.
[7] Knorr E, Ng RT, “Algorithms for mining distance-based outliers in large datasets.” In: Proceedings of the 24th VLDB conference, New York, pp 392–403
[8] M. O. Mansur, Mohd. Noor Md. Sap, “Outlier Detection Technique in Data Mining: A Research Perspective” Proceedings of the Postgraduate Annual Research Seminar, 2005
[9] Karanjit Singh and Dr. Shuchita Upadhyaya, “Outlier Detection: Applications And Techniques” International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012
[10] D.Hawkins: “Identification of outliers”. Chapman and Hall, London. 1980.
[11] Amandeep Kaur, Ms. Kamaljit Kaur, “Different Outlier Detection Algorithms in Data Mining”, A Novel Approach to Face Recognition and Expression Analysis using Local Directional Number Pattern