International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey on Outlier Detection Methods

Rajani S Kadam[1], Prakash R Devale[2]

[1]Bharati Vidyapeeth University, College of Engineering, Dhankwadi, Pune, India

[2]Professor, Bharati Vidyapeeth University, College of Engineering, Dhankwadi, Pune, India

Abstract: ***Outlier detection is the process of finding the data that behave very differently from the normal expected behavior. Outliers may be due to system errors, noise or human intended action. Outlier detection becomes an important task in any application to provide reliable and effective results. Many outlier detection methods are proposed amongst SVDD is most commonly used method. This paper provides a brief survey on the outlier detection methods.***

Keywords: Outliers, Outlier detection, PCA, SVM, SVDD.

1. Introduction

Data mining is the process of extracting the useful data from the given data base or the data set. In the process the unexpected data or not of interest data are called the negative data or outliers or anomaly. Thus outlier is nothing but the data object that violates the predefined behavior. The good and proper definition of outlier is given in[1].The outliers are the data objects that behave very differently from the normal data set giving a suspicion that the data are generated by different machines. Outliers are classified into three types

1) Global outliers -- data deviates significantly from the rest of the data
2) Contextual outlier -- based on specified context, data is labeled as outlier or normal data. For example "today's the temperature is 40^{o} C" this statement can be outlier if it is winter or the normal if it is summer. Depending on the context data can be outlier or normal.
3) Collective outlier -- subset of data objects are considered as outlier if they behave significantly different from rest of the data.

Following diagram shows the outliers

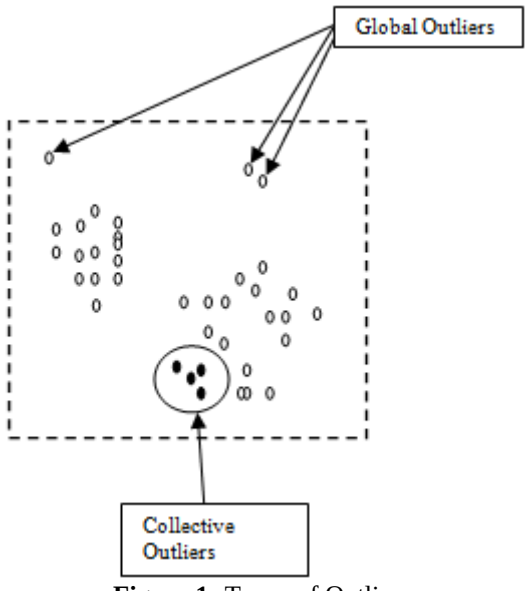

Figure 1: Types of Outliers

These outliers are few in number but can be suspicious. There are many causes for these outliers like

- Change in system behavior
- Errors made by humans
- Intentional act to deceive

Outliers due to error can be just discarded. Outliers of intended action, special care must be taken in using such database or data set. The process of finding the data objects with such a suspicious behavior that are very different from the expectation is known as outlier detection (also known as anomaly detection).

Outlier detection has found many applications like fraud detection, fault analysis in system designs, terrorist attack, medical abnormalities, sensor network surveillance, law enforcement, earth science and many more.

There are many difficulties in outlier detection

1. Outliers are few in number and become difficult to train the classifier accurately with such sparse number.
2. Noise is always present in the data set causing the distortion of data. These may blur the distinction between normal and outlier
3. Sometimes application requires the knowledge of the data behaving as outlier. Finding this reason becomes difficult
4. outlier detection are application specific and hence becomes impossible to develop a universal applicable outlier detection method

There are many outlier detection methods proposed [1],[2],[3],[10],[11]to detect outlier and the normal data. These are classified into four types: statistical based approach, proximity based approach, clustering based approach and model based approach.

Statistical based approach -- make assumptions of the data normality; the data not following the assumption is the outliers. The success of this statistical method purely depends on the assumptions made for the statistical model. The assumptions made may not be sometimes true especially for real and high dimensional data set.

Proximity based approach -- The data objects are treated as outliers if the data object is very far from its neighbor that is

the nearness of the object from its neighbor is significantly large compared to others in the space. The proximity is measured in terms of distance or density. The effectiveness of proximity based methods depend on the proximity metric used. The advantage of proximity based over statistical based is that no prior assumptions are made but these methods have computational complexities. Sometimes it becomes hard in detecting the group of outliers if they are close to each other.

Clustering based approach -- the data objects are grouped into clusters based on the features or behaviors –normal clusters and outlier clusters, normal data objects being huge in number and outliers are small in number and do not belong to any needed clusters. There are different clustering methods but are the costly data mining techniques and thus are not suitable for large data sets. Clustering methods are unsupervised methods, no need for labeled data while training and the performance is limited.

Model based approach -- Predefined models are constructed using training data and then the data objects (test data)are classified as normal or outlier. The main goal is to train the classifier and then recognize the outlier. In some methods normal data are labeled, the objects not matching the model are outliers. Others model the outliers and the data not matching the model are normal data. They are supervised methods, training data set are labeled. Challenges do exist in supervised learning methods like two classes are imbalanced; data may be mislabeled thus restricting the classification accuracy. Different model based approach exist among them SVDD is the most effective and widely used method because of its ability of detecting the outliers in various domains.

2. Overview of the Algorithms SVDD and PCA

2.1. Support Vector Data Description (SVDD): support vector data description is a variant of support vector machine (SVM). SVM is a data mining and a machine learning technique. This technique is most widely used classification method. The algorithm mainly aims at finding the hyper plane in the input space which is used to separate the data objects into classes.

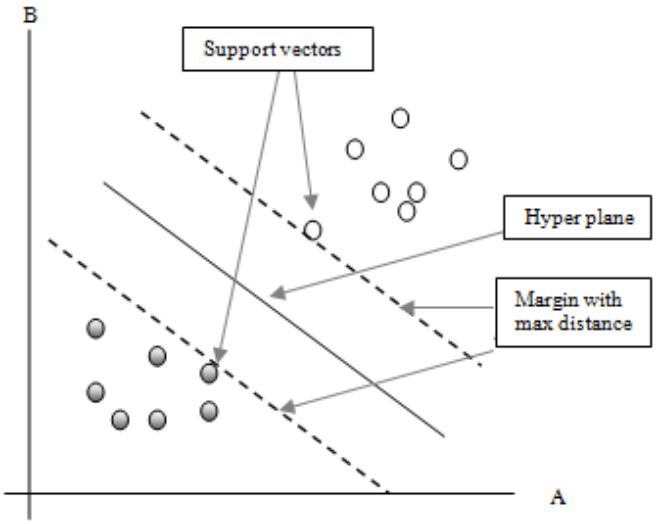

Figure 2: Support Vector Machine

The SVM finds the hyper plane with the help of support vectors and the margin. These margins are marked by support vectors. Larger the distance between these margins there will be the clear separation of data into classes. The method is slow but effective and gives accurate results. Numerous separating lines can be drawn but the algorithm finds the best hyper plane with maximum separation between the margins with minimum error.

Support vector data description is a variant of SVM. SVDD is usually used for one class classification. Only positive data are given and the task is to detect the outliers. The main objective is to find the minimal sphere in space. Given the set of data, SVDD aims at finding the minimal hyper sphere that contains the desired data. The data outside the sphere are nothing but the outliers.

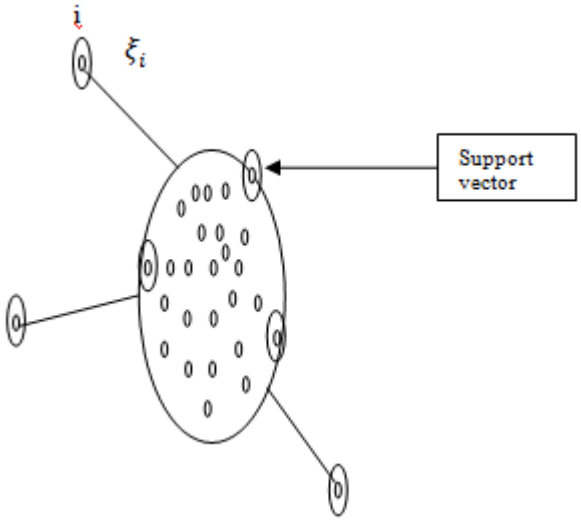

Figure 3: Support Vector Data Description

Minimal hyper sphere is given by

$$R^2 + \frac{1}{\mu n}\sum_{i=1}^{n}\xi_i$$

Such that $\|\phi(i)-b\|^2 \leq R^2+\xi_i$

Where' R' is the radius of the sphere
'b' is the center of the sphere
'$\frac{1}{\mu n}$' *is the constant (sometimes referred as C in the paper)*
ξ_i *are the slack variables with values* ≥ 0

The SVDD is the most effective in detecting the outliers. This SVDD is sensitive to noise in the input data. Many research have been done to handle this problem and to boost the performance of SVDD

2.2. Principal component analysis (PCA)

PCA is the unsupervised method of dimensionality reduction. The data set to be reduced consists of data objects that are represented by n dimensions then the PCA reduces this dimensionality giving the smaller dimension so the data set is represented in small space. First the principal direction of data distribution is determined by constructing data covariance matrix and calculation the Eigen vectors. These

Eigen vectors are informative and are the principal directions.

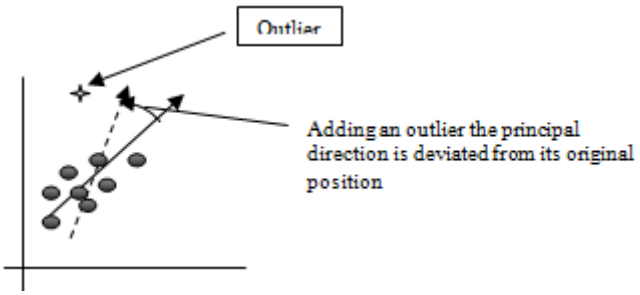

Figure 4: Principal Component Analysis

When data is added to the data set, it is observed that original principal direction deviates. The angle between the resulted principal direction and original principal direction can be used to determine whether the added data is outlier or normal. The angle of deviation is usually small when normal data is added and larger when outlier is added hence outlier detection method. PCA is unsupervised method, data used are not labeled and hence performance is limited

3. A Brief Survey on Outlier Detection Methods

In [5], robust one class SVM is proposed to handle the sensitivity problem of one class SVM. The proposed algorithm aims at modifying the penalty factor$\frac{1}{\mu n}$In optimization model [6], the data point x_i is allowed to locate outside the boundary if slack variable $\xi_i$is non zero. Lesser the penalty factor is, the fact of x_i to locate on the boundary is more. Distance between x_i and the b the center of the sphere are used here to have the control on ξ_i. The outliers are very distinct in space then the normal data with respect to center and hence they have small penalty factor d_i then the normal data. Thus the outlier are more likely to lie outside the boundary. Outliers are lie far enough thus the effect of outliers in training the machine is prevented. Thus the robust OC-SVM handles the sensitivity problem

In [7], Position regularized support vector domain description (PSVDD) algorithm has be proposed by the authors. As defined in [8] Domain description is to describe the data objects that are the normal ones and outliers are rejected. SVDD is the best algorithm for the task as it gives accurate and flexible result with the help of only few number of support vectors. The algorithm is very sensitive in selecting the trade off parameter 'C' which is difficult to calculate. To handle this problem the authors proposed PSVDD, the volume of the feature space is controlled by assigning the weight to every data object based on the position. The performance of SVDD mainly depends on the value of C. The 'C' value for each data of outlier is taken to be same. This is not the true due the different density of each objects and the relation between the data objects. Thus in PSVDD weighting to each point is assigned which is inversely propositional to the distance between feature space image x_i and mean of feature space image $\frac{1}{n}.\sum_{j=1}^{n} \phi(\mathrm{xj})$

In [9], SVDD algorithm has been used to detect outliers on uncertain data. The existing outlier detection methods always make the assumption that the data can either be normal or outlier. This is not true in real world. Data set may contain uncertain data due to errors or partial completeness. Normal data objects can be labeled as outliers due to error or noise in the process. The authors have used the SVDD algorithm to classify these uncertain data. Pseudo-training data set is generated having each data assigned with confidence score using kernel function. This score describes the likelihood value of each data towards the normal class. The data sets along with the confidence score is then used to train SVDD in the training phase to build a classifier. The confidence score of the outlier is smaller as compared to that of the normal data. Thus the data with lowest confidence score can be prevented by influencing the classifier during the training, leading to solution of sensitivity problem of SVDD. The proposed technique out performs as compared with GMM and standard SVDD.

In [4], Authors proposed kernel principal component analysis (KPCA) based mahalanobis distance for detecting outliers in wireless sensory networks. Mahalanobis distance is used to calculate the mapping of data objects to the feature space so as to separate outliers from normal data. WSN produces data highly unreliable and uncertain. Effective processing and analysis of data stream becomes important. Outlier detection in sensor networks is done to improve the performance and quality of the networks. Given the data set the PCA transforms them into set of uncorrelated variables called principal components. The mahalanobis distance is calculated for the data point to decide it is outlier or normal data. The MD-based KPCA gives better and fast result in detecting the outliers effectively with low power consumption.

In [12], Authors proposed anomaly detection via online oversampling principal component analysis. Anomalies are few in number. It becomes difficult for detection of anomalies in large data set as numbers of outliers are very few in number as compared to normal data. These require more memory and high computation. To handle this problem the authors proposed the online oversampling the outliers so the number is increased and becomes easy to train the classifier. Thus giving the better performance in detecting the outliers. This is online oversampling hence can be applied only to online and large scale problems. The unbalanced data problem is solved using leave one out strategy were outlier number is either increased or decreased to affect the principal direction of the resulting data. They should oversampling of data significantly caused the effect of outliers .since the data was used online and not stored during detection process the computational cost and memory consumption was reduced hence providing an effective anomaly detection technique.

4. Conclusion

Outliers are the data objects in the data set that behave very differently from normal data and can be harmful sometimes. Outlier detection becomes important in any application to

provide accurate, reliable results. Many outlier detection methods exist amongst SVDD is more efficient.

References

[1] D. M. Hawkins, Identification of Outliers. Chapman and Hall, Springer, 1980.
[2] M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000
[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, 2006.
[4] Oussama Ghorbel, Walid Ayedi, Hichem Snoussi, and Mohamed Abid, Fast and Efficient Outlier Detection Method in Wireless Sensor Networks,IEEE journal, June 2015
[5] Shen Yin, XiangpingZhu, ChenJing, Fault detection based on a robust one class support vector machine,Elsevier, June 2014
[6] B. Shen,S.X.Ding, Z.Wang, Finite-horizon H1 fault estimation for uncertain linear discretetime-varyingsystemswithknowninputs,IEEETrans.CircuitsSyst .60-I(1)(2013)902–906.
[7] Chang-Dong Wang, JianHuang Lai, Position regularized Support Vector Domain Description ,Elsevier, October 2012.
[8] D.M. Tax, R.P. Duin, Support vector domain description, Pattern Recognition Letters 20 (1999) 1191–1199.
[9] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "Svdd-based outlier detection on uncertain data," Knowl. Inform. Syst., vol. 34, no. 3,pp. 597–618, May 2012
[10] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.
[11] N.L.D. Khoa and S. Chawla, "Robust Outlier Detection Using Commute Time and Eigenspace Embedding," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, 2010.
[12] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang, Anomaly Detection via Online Oversampling Principal Component Analysis, IEEE transactions on knowledge and data engineering, july 2013

Author Profile

Ms. Rajani S. Kadam is a student of M.Tech in Information Technology, Bharati Vidyapeeth Deemed University College of Engg, Pune-43.

Prof. Prakash Devale is Professor in Information Technology Dept., Bharati Vidyapeeth Deemed University College of Engineering, Pune-India. He did his B.E in Computer Science, M.E. from Bharati Vidyapeeth Deemed University College of Engineering, Pune in 2002 and Ph.D.Pursuing in BhartiVidyapeeth Deemed University in the area of Machine Translation .He is having 21yrs of experience in teaching. His Publication in International Journals : 33, Publication in International Conference : 9,Publication in National Conference : 17. He is lifetime member of ISTE.