Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.164 – 168

RESEARCH ARTICLE

Efficient Classification Technique for Outlier Detection

Ms. Rajani S Kadam[1], Prof. Prakash R. Devale[2]

[1]Department of Information Technology
Bharati Vidyapeeth Deemed University College of Engineering, India
[2]Department of Information Technology
Bharati Vidyapeeth Deemed University College of Engineering, India
[1] rajaniskadam@gmail.com; [2] prdevale@bvucoep.edu.in

Abstract: Outliers are the data objects that clearly differ in their behavior from the normal data. Outlier detection mainly aims at finding these data objects. Outlier detection has become the major area of research in data mining. This plays a crucial role in data mining. Most of the methods used for outlier detection, consider the positive data and their behavior, and then the data violating the behavior are termed as outliers. Most of the time the data may be corrupted making it difficult to identify the data clearly. To handle this problem the paper provides an approach to improve the classification efficiency by generating the likelihood value for each data in the dataset. The kernel K-means clustering is used to compute the likelihood value which defines the membership value towards each class. The data with the value is subjected to classifier thus improving the accuracy in outlier detection.

Key words: Confusion matrix, K-means, Outlier, Outlier detection, SVDD.

I. INTRODUCTION

The snappy growth in information technology results in huge database. These databases become difficult to handle manually. Hence automation is required for these databases. With the growth in size of database, the unnecessary data also increase. This causes a tedious job to search the required information. The data mining plays a crucial role in these areas. The data mining not only fetches the required information but can also be used to find the relation between the data. While mining the required data or extracting the positive samples, some uncertain data are also detected. These data are outliers. They neither are positive nor are they negative samples. Outliers are the data objects the significantly differ from normal or positive data [1], [2]. A proper and well known definition of outlier is given in [3]. These suspicious data can be harmful sometimes. They can be intended actions and hence need to be checked. Outlier detection thus becomes essential in data mining.

Numerous outlier detection methods have been proposed [3], [4], [5], [9], [10] each with their advantage and disadvantages. Outlier detections are used in many areas like intrusion detection, fraud detection, health science, fault detection in systems, military surveillance and many more.

Majority of the methods used for outlier detection mainly construct the model by considering the behaviour of positive samples. These methods are then used to detect the outliers as *the data that fails to behave as defined*.

Sometimes with huge database it may be possible that data is corrupted due to noise or some other condition. This may cause the data to be mislabelled; the positive data may behave as outlier. This makes the classification difficult. Further more if the data is linearly inseparable then special functions are used to separate them.

A. Linear and Non linear data

Linearly separable data are the data objects in the dataset which can be easily separated with the help of Euclidean distance.

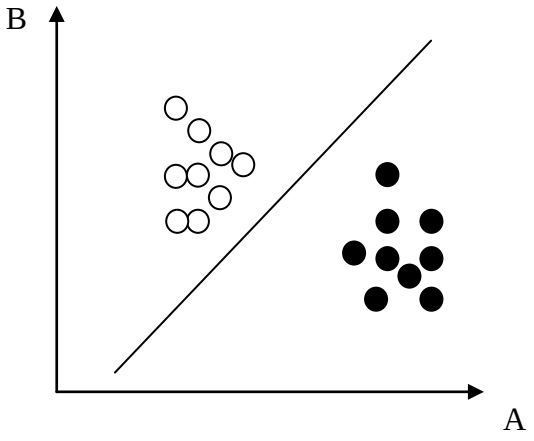


Figure (a)

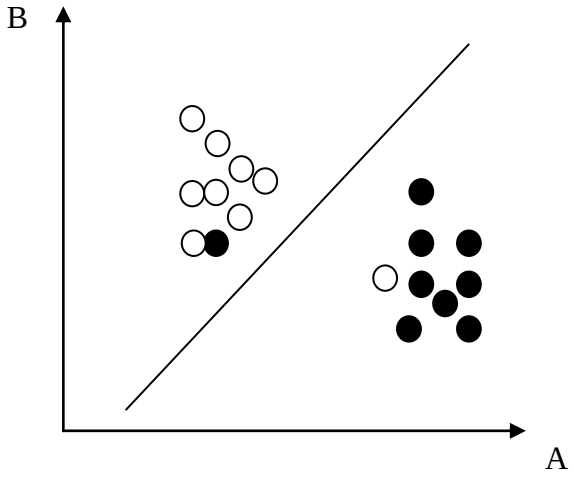
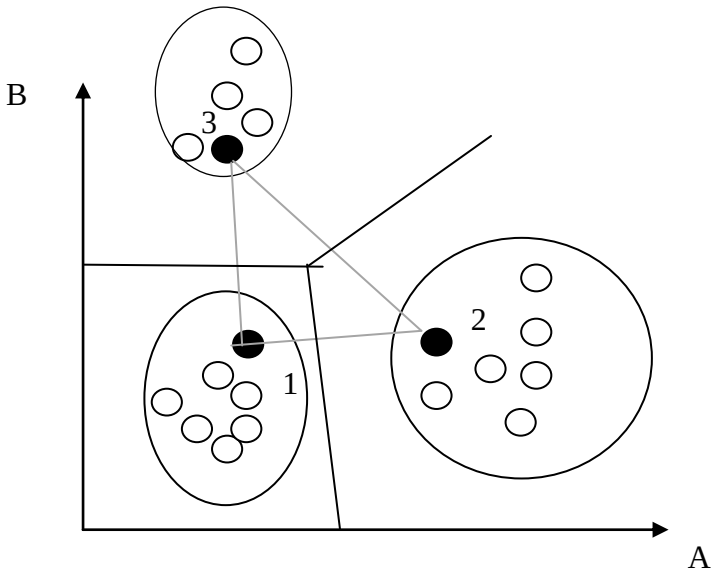


Figure (b)

Consider the figure (a), here we want to separate the dataset into two groups say group 1 and group 2. A simple line can separate the data effectively. Thus set of points are said to be linearly separable if there exists a line that separated to two groups clearly. But in figure (b) the data cannot be separated with the help of a single line in the plane. In such situation where data are linearly inseparable, the original data is transformed to high dimensional space and then linearly separating hyper plane is searched which separates the data.

B. K-means

Clustering is the most important operation in data mining[6].Clustering can be done by many methods[7]:partitioning methods, grid based methods, hierarchical method ,model based methods, density based methods and many more. K-means is a clustering algorithm which follows the partitioning method used in data mining. This method works by computing the Euclidean distance between each point in the dataset and forming K clusters. It works as follows



K-Means clustering

In the figure the distance between point 1 and point 2 is calculated and a perpendicular bisector is found. Similarly between point 2 and point 3, and point 3 and point1.The perpendicular bisectors form the cluster space. The points falling in region 1 with minimum Euclidian distance between the point and point 1 fall into the cluster 1. Similarly cluster 2 and cluster 3.Thus clusters are established.

II. OUR APPROACH

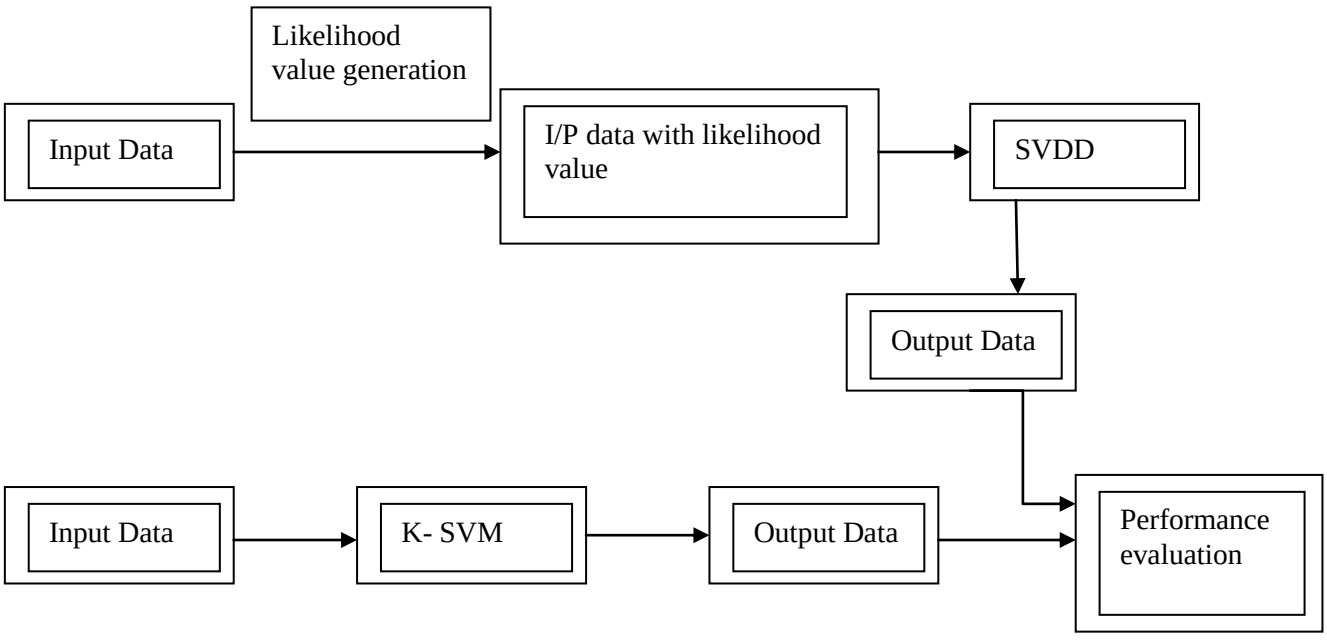
Our work mainly aims at improving the efficiency of outlier detection. The Support Vector Data Description (SVDD)[8] is most commonly used algorithm for outlier detection. The algorithm is a variant of one class SVM mainly aims at constructing the hyper sphere of radius R given by
Minimal hyper sphere 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 centre 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 utmost effective in converting the input data into feature space and detect the outliers. But 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 .Due to presence of noise, data may be misclassified that is sometimes the positive data may be detected as outlier. To handle this many methods have been proposed. Our work mainly aims at this problem. Before the data is classified the likelihood value is generated for each data in the data set [11]. This likelihood value decides its degree of wiliness towards positive or negative class. With the help of this value and SVDD data is classified as positive or outlier and hence improving the performance of classification. To generate the likelihood value Kernel k-means clustering is used. The basic idea is to form n number of small clusters were the data in the same cluster are similar to each other. Thus for any data in the cluster, if most of the data in the same cluster are normal then the change of thee data being normal is more. Similarly if the data does not belong to any cluster, then the possibility of the data being an outlier is more definite. Thus with the likelihood value the data can be clearly classified and the effect of noise on the data can be reduced, boosting the performance of the classifier.



The diagram illustrates the flow of the project. The data is pre-processed and then given to k-means clustering to generate the likelihood value which defines the membership degree. Data with the likelihood value is then given to SVDD algorithm to identify the data as positive, negative or outlier.
Kernel K means clustering is used to generate the likelihood value. When it becomes difficult to separate a data by linear separation non linear classification is used. But sometimes the data are so closely bond towards each other that the separation becomes difficult and so to improve the accuracy we propose the likelihood value generation for each data. Kernel k-means develops the local clusters such that the members in each cluster are similar to each other, the data that does not belong to any of the cluster are probably the outliers with clear cut off from the clusters hence with likelihood value generation the accuracy of outlier detection can be improved. This likelihood value defines the degree of membership towards its own class and towards other class.

A. Evaluation Procedure

Confusion matrix is built for the data. The confusion matrix evaluates the classifier performance. Training data is used to train the classifier and then the trained model, the classifier is used to classify the data. To measure the performance of the classifier the test data is used. Confusion matrix is the tabular representation showing the number actual class and predicted class.

		Predicted Class		
		Positive Class	Negative Class	Total
Actual Class	Positive class	TP=4	FN-2	P=6
	Negative class	FP=3	TN=1	N=4

Confusion Matrix

The confusion matrix above shows the result of two class classifier. The actual positive class is 6 out of which classifier predicted 2 as negative class.

The SVDD classifier performance is evaluated with the help of this confusion matrix using the below five parameters

1. ***Detection rate***: gives information about number of correctly identified outliers

   Detection rate =TP/ (TP+FN)

2. ***False alarm rate***: gives the number of outliers misclassified as normal data tuples

   False alarm rate=FP/ (FP+TN)

3. ***Accuracy***: percentage of test set tuples that are correctly classified by the classifier

   Accuracy=(TP+TN)/(P+N)

4. ***Error rate***: percentage misclassification rate

   Error rate= (FP+FN)/(P+N)

5. ***Precision***: measure of exactness

   Precision=TP/ (TP+FP)

III. CONCLUSION

Outliers are the data that behave very differently from the normal data. SVDD, the algorithm most commonly used to detect outliers is sensitive to noise. The work intents at boosting the performance of SVDD in effectively detect the outliers by generating the likelihood values for the data.

ACKNOWLEDGEMENT

We would like to thank all the reviewers for their reviews, comments and suggestions

REFERENCES

[1]V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM CSUR*, vol. 41, no. 3, Article 15, 2009.

[2] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies,"*Artif. Intell. Rev.*, vol. 22, no. 3, pp. 85–126, 2004.

[3] D. M. Hawkins, Identification of Outliers. Chapman and Hall, Springer, 1980.

[4] 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

[5] 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.

[6] M. R. Anderberg, *Cluster Analysis for Applications*. New York, NY,USA: Academic, 1973.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*.San Mateo, CA, USA: Morgan Kaufmann, 2011.

[8]D. M. J. Tax and R. P. W. Duin, "Support vector data description,"*Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.

[9] V. Barnett and T. Lewis, Outliers in Statistical Data. John Wiley&Sons, 1994.

[10] 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.

[11] S. Y. Jiang and Q. B. An, "Clustering-based outlier detection method," in *Proc. ICFSKD*, Shandong, China, 2008, pp. 429–433.

[12] 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, 2013.

[13] B.Liu, Y. Xiao, Philip S. Yu, Z.Hao, and L. Cao "An Efficient Approach for Outlier Detection with Imperfect Data Labels"IEEE Trans. Knowledge and data engineering, vol. 26, no. 7, July 2014

[14] C. C. Aggarwal and P. S. Yu, "A survey of uncertain data algorithms and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 21,no. 5, pp. 609–623, May 2009.

[15] X. Huang, Y. Ye, and H. Zhang. "Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation "IEEE trans. Neural networks and learning systems, vol. 25, no. 8, August 2014

