Efficient Outlier Detection by Integration of Clustering and Classification

Sarita Tripathy [#1], Dr.Laxman Sahoo [*2]

[#1] Assistant Professor, School of Computer Engineering , KIIT University,Odisha,India
[*2] Professor, School of Computer Engineering,KIIT University,Odisha,India
[1] sarita.tripathyfcs@kiit.ac.in
[2]lsahoofcs@kiit.ac.in

***Abstract*—This paper addresses a new method which integrates the clustering method DBSCAN with the classification method KNN to improve the quality of real data sets by removing noise data. DBSCAN algorithm is a versatile density based clustering algorithm which is employed in this paper.The proposed method consists of first applying DBSCAN algorithm to the data set and secondly KNN algorithm is applied.The experiments conducted proves that the proposed approach is better method and increases the accuracy.**

Keyword- *KNN,DBSCAN,Clustering,Outlier.*

I. INTRODUCTION

Outliers are defined as data present in a data set which are not consistent with the remainder set of data [1] or which is distant from other observations in the data set due to which they seem to be generated by some different mechanism and are considered to be abnormal .If a data set contains erroneous data then it can result in wrong output and will affect the accuracy of the model which is derived from the export rules[2] and data set used.It will also result in additional operational cost.Outlier detection[3] can lead to the discovery of useful information,and it is the primary step in many data mining applications.

One of the crucial part of data cleaning is unwanted data processing in the data set. The types of unwanted data detection methods are : (1) Based on Statistics. (2) Based on Distance[4]. (3) Based on Density. (4) Based on Information. (5) Based on Clustering[5].

Primary objective of the clustering algorithm is to group data,but there are some clustering algorithms which will produce some “by products” which are outliers.Hence,these clustering algorithm can be used for noise detection and are highly intuitive and also human perception of outliers is inline with it [6]. Also the process is simpler as it needs to only traverse over the data set and calculate the distance from the object to the center of the cluster,in the phase testing.As the clustering process is completed it results in faster testing speed and also it does not increase the complexity,the complexity totally depends on the clustering algorithm used[7],[8]. This paper proposes an improved method for noise data detection proposed in which the classical density based clustering algorithm DBSCAN is combined with the classification algorithm used for outlier detection i.e KNN.The algorithm has two stages in the first stage the classical DBSCAN algorithm is applied to the data set, the m clusters obtained after this are labeled as m classes , the DBSCAN algorithm identifies some points as the noise points then in the second stage all the noise points are reconsidered and KNN classifier is applied to check whether the points belong to any of the class or not, if it belongs to any of the classes then it is added to the normal set of points or else it is detected to be a noise point. The organisation of this paper is as follows. Section II describes the motivation and objective of our work, in section III the related work in this field is presented, in section IV DBSCAN algorithm and KNN algorithm are presented. In section V,DBSKNN Noise data detection Algorithm is proposed. Section VI illustrates the practical importance of the work. Finally, section VII Conclusion is given.

II. RELATED WORK

In the year 1996 Zhang et.al. [9] proposed an outlier detection method which was local distance-based outlier detection method.The degree of deviation of an object from its neighborhood is calculated by LDOF. It calculates the extent of deviation of an object from the objects which are in the neighborhood of the object.

In the year 1998 Knorr and Ng [10] introduced outlier detection technique based on distance.An object o is considered as an outlier in a data set if q objects in the data set are present at a distance greater than the distance from p,where q is considered to be a least fraction of objects.This is the definition accepted widely as it generalizes many statistics based outlier test.

In the year 2000 an extended approach of the above mentioned method was proposed by Ramaswamy et.al. [11]. In accordance with the outlier score the ordering of the points was done.If two integers are given as I_1 and I_2 then for o to be considered as the outlier,less than I_2 objects have a value greater for S_k than I_1,here S_k is the distance of the object p from the k^{th} closest neighbor.

In the same year a new method in which for each object in the data set a Local Outlier Factor(LOF), was determined along with indication of degree of outlierness was proposed by Breunig et.al. [12], It followed the process of determining how far a point considered as outlier differ from the other set of points in the data set.It is considered to be more efficient than the instance based scheme.Hence the conclusion is that categorization of outliers and normal data is not done explicitly by the density based scheme.

Subsequently, In the year 2002 a new method which consisted of identification of noisy data by considering all the objects in the neighborhood was proposed by Angiulli and Pizzuti [13]–[15].Sum of distances of k nearest neighbors were considered to rank all the points.

There are other clustering methods helpful in determining unwanted data in a data set for example CLARANS [17], DBSCAN [18], BIRCH [19] and CURE [20].But the primary objective of these methods is to improvise the clustering method and not unwanted data detection.

III. **DBSCAN**

This is a data clustering algorithm. In this algorithm given a set of points grouping of those points is done which are packed together and have many nearby neighboring points.Where as the points having less density i.e whose neighborhood point are at a large distance are marked as noisy data. In the year 2014, this density based clustering algorithm was given the test of time award (an award given to those algorithms which have got high recognition in theory and practice) at the highly recognized conference of data mining KDD[21].The DBSCAN algorithm requires two inputs or parameters,first is Epsilon(eps) which specifies the radius within which the points are to be considered and the second parameter is the least number of points required to form a dense region(minPts).To begin the algorithm,a random point is taken to be the starting point,then the neighborhood of the point within an epsilon radius is checked and the number of points present are retrieved,if the point contains sufficient number of neighborhood points than a cluster is started or else that point is marked as a noise.

Epsilon value can be chosen by plotting a graph known as k-distance graph.The distance is plotted to the k=(minimum points) nearest neighbors. The point where the plot shows a strong bend as shown in figure-1 below is the most appropriate value of epsilon.Choosing the epsilon to be too small results in leaving large number of data with out being clustered and on the other hand high value results in merging of two or more clusters and also putting majority of the objects in the same cluster.However small value is generally preferred.

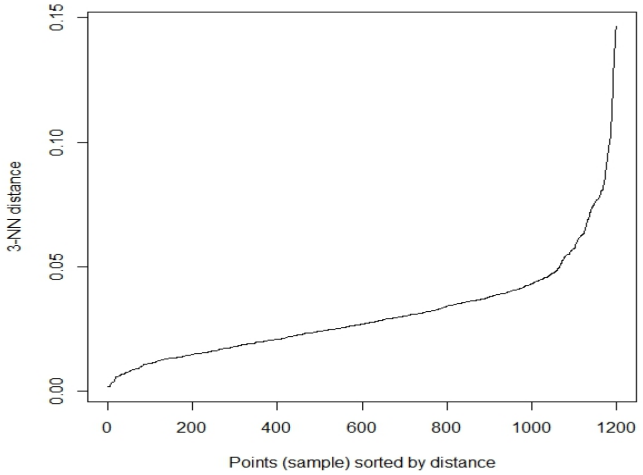

Fig 1. The K-dist plot

A. Advantages of DBSCAN:

1. Unlike the k-means clustering the value of k need not be specified before hand.
2. As opposed by other clustering algorithms like K-means and K-medoids it can find arbitrary shaped clusters and also clusters connected by thin lines and totally surrounded by other clusters.
3. It is also robust to noise or outliers.
4. Ordering of points in the database has no effect on DBSCAN and it just requires two parameters epsilon and minimum points.
5. The design of DBSCAN is such that it can process and run region queries for example,the R* tree.
6. The setting of the parameters can also be done by the domain expert if the data in the data set is easily understandable.

7.

B. Drawbacks of DBSCAN

1. The algorithm is sometimes undeterministic i.e the points on the border can be included in more than one cluster and thus they can belong to either cluster.The order in which the data is processed determines this,however it does not impact much the core and noise points.
2. The accuracy of the algorithm is dependent on the distance measure used,also it becomes very difficult and complex for high dimensional data.
3. Choice of appropriate distance threshold ε is difficult if one has not understood properly the data and scale required.

IV KNN Algorithm

It is one of the classification method used for detection of unwanted data in a data set.In this method a fraction of training examples in feature space is taken as input, and a class name can be obtained as an output.The notion of closest neighbor and k-distance is used here,the set of k-closest data points of a point p is the k-neighborhood of that point.It calculates the k-closest data points of all the points present around it and orders them in sorted i.e decreasing order of these values with the first n data points being considered as outliers.It is very critical to determine the value of k. There will be higher impact of noise on the result for small value of k,on the other hand large value of k is computationally expensive.The most simplest and widely used approach to determine the value of k is to set k=n^(1/2).

A.Advantages of KNN

1. KNN is robust to data with noise especially if we use inverse square of weighted distance as the distance measure.

2. It can be effectively applied to large data set.

B. Drawbacks of KNN

1.Parameter k needs to be determined in advance.

2.The choice of the distance measure and the attribute which is to be used, so that it will give the best result is difficult.

3.Also it is required to determine whether we shall use all the attributes or some attributes.

4.Here we have to compute the distance of each testing sample to all training samples,hence the computation cost is very high.

V Proposed Algorithm

The proposed algorithm combines two techniques of data mining.The first phase consists of application of density based clustering algorithm DBSCAN on the data set,after using DBSCAN the clusters obtained are marked as separate classes and each class is marked as Cl_k (k=1,2,…,m).KNN algorithm is then applied by considering the noise points obtained from the DBSCAN algorithm as the testing set and the clusters obtained as the training set.

A.DBSKNN algorithm

Input: *The original data sets DS, given radius eps, the minimum number of objects in neighborhood mpt, an appropriate value of K is taken according to the data set.*

Output: *The set n_o include all the noise detected from original data sets DS.*

Setp1: *Given the value of eps and mps..*

Setp2: *Select a point randomly from data sets DS, and count the number of points in its neighborhood. If the number is higher than mpts, the point is marked as core object. Else, it is marked as noise.*

Step3: *if the point is a core object, create a cluster with radius eps and the core point. Then add the objects in the cluster into a list container, and check the objects recursively. If the object in the container is a core object, classify it to the same class as the point and add its neighborhood points into the container. Else, mark and delete it from the list container.*

Step4: *repeating Step 2 and Step 3 until the objects in data sets DS are marked as a class or as noise points if it is found not belonging to any class.*

Step5: *mark the m clusters obtained from step 4 as data sets DS_t (t=1,2,…,m).*

Step6: *Take an appropriate value of K.*

Step7: *train and build KNN on DS_t,*

Step8: *Consider randomly a point which is not included in any of the data set, find out the k-nearest neighbors of the point, the following three situations may arise:*

i) If 50% or more than 50% of its k neighbors belong to $class_i$ include the point in $class_i$

ii) If 50% of its neighbors belong to $class_i$ and another 50% belongs to $class_j$ then it can be included to any class.

iii) If there is no class to which at least 50% of the neighbors belong such a point is marked as outlier and included in N.

Step9: *Output the set N which contains the noise data.*

DBSKNN algorithm can effectively reduce error samples of DBSCAN clustering by introducing KNN algorithm, and significantly improve the clustering accuracy.

VI Result Analysis

We have performed experiments to verify the effectiveness of the DBSKNN algorithm and only DBSCAN algorithm without introducing KNN.The algorithms are implemented using r-programming. The UCI data set Iris is used as the experimental data.The data set consists of 50 samples from three species of *Iris flower* which are categorized based on four features they are (i)Length of Sepal(ii)Width of sepal (iii)Length of Petal (iv)Width fo Petal.The class label identified in the dataset is used to test the efficience and effectiveness of DBS-KNN noise data detection algorithm.75% of the data set are taken as training set and the remaining 25% is taken as testing set. The basic flow of the experiment is that conduct clustering on the training set by DBSCAN algorithm at first.Then according to the results of the clustering, train with KNN algorithm on each category which is identified and then get discriminant model for each category. The resulting models are used to classify training and testing sets. DBS-KNN algorithm can be judged by the result of the experiment as shown in table1.

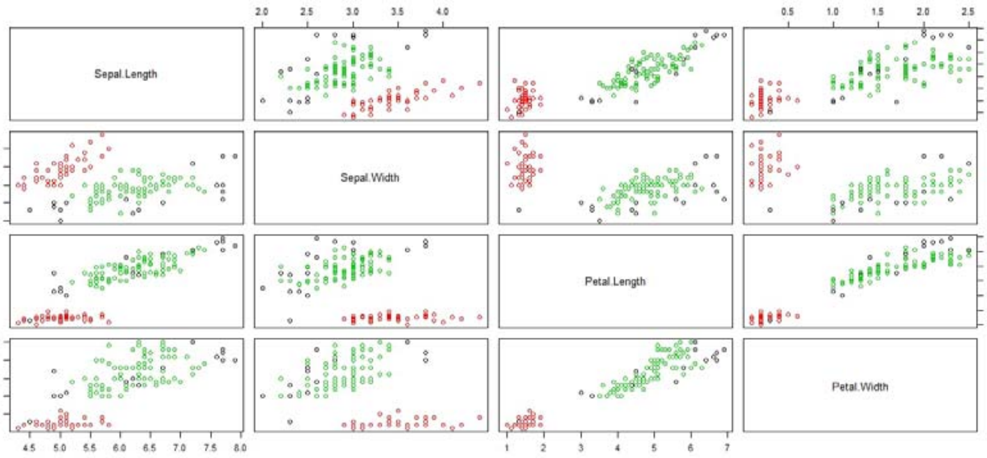

Fig-2 The attributes in Iris dataset

TABLE I

Algorithm	Eps	Minimum points	Number of noise points	Accuracy
DBSCAN	.4	4	25	0.8933
DBS_KNN	.4	4	10	0.9231
DBSCAN	.6	6	11	0.9012
DBS_KNN	.6	6	4	0.9722

It can be observed from the result the accuracy of the algorithm increases with the increase in the Minimum points and the Epsilon value and the the number of noise points also decreases.also the value of k for the KNN algorithm can be varied.

VII Conclusions

The new method based on DBSCAN and KNN proposed in this paper is more efficient as it minimizes the error of DBSCAN algorithm.As evident from the experimental output the algorithm improves the accuracy of the clustering results.The parameter those are required to be selected are

(i) minimum points (ii) Epsilon (iii) the parameter k.Manually we can adjust the parameters according to the application area.Future work will consist of finding out a method for automatic selection of the optimal parameters.

References

[1] R. Erhard, and H. D. Hong, “Data cleaning: problems and current approaches,” IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3-13,2000.
[2] J. Han, and M. Kamber. “Data mining: concepts and techniques,” Morgan Kaufmann Publishers, 2007.
[3] F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. IEEE Transactions on Knowledge and Data Engineering, 17:203–215, 2005.
[4] Y. Li, S. Nitinawarat, and V. Veeravalli, “Universal outlier detection,”Information Theory and Applications Workshop (ITA), 2013: 528-532.
[5] F. Angiulli, S. Basta, and C. Pizzuti. Distance-based detection and prediction of outliers. IEEE Transactions on Knowledge and Data Engineering, 18:145–160, 2006.
[6] P . Poonam, and M. Dutta, “Performance analysis of clustering methods for outlier detection,” 2012 Second International Conference on Advanced Computing & Communication Technologies (ACCT), Rohtak, Haryana, 2012.
[7] M. Ahmed, and A. N. Mahmood, “A novel approach for outlier detection and clustering improvement,” Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013: 577-582.
[8] Ester, M., Kriegel, H.-P., Sander, J., and Xu X. (1996), A density-based algorithm for discovering clusters in large spatial data sets with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR,pp.226-231.
[9] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In PAKDD ’09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,pages 813–822, 2009.
[10] E. M. Knorr and R. T. Ng. Algorithms for mining distance based outliers in large datasets. In Proceedings of the 24th International Conference on Very Large Data Bases, pages 392–403, San Francisco, CA, USA, 1998.
[11] Ramaswamy, S., Rastogi, R., and Shim, K., Efficient Algorithms for Mining Outliers from Large Data Sets,Proc. of ACM SIGMOD Int. Conf. on Management of Data, 2000, pp. 427–438.
[12] M.M.Breunig,H.-P.Kriegel,R.T.Ng,andJ.Sander.Lof: identifying density-based local outliers. SIGMOD Rec.,29(2):93–104, 2000.
[13] F.Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In PKDD ’02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, pages 15–26, 2002.
[14] F. Angiulli, S. Basta, and C. Pizzuti. Distance-based detection and prediction of outliers. IEEE Transactions on Knowledge and Data Engineering, 18:145–160, 2006.
[15] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In VLDB ’99: Proceedings of the 25th International Conference on Very Large Data Bases, pages 211–222, 1999.
[16] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. pages 427–438, 2000.
[17] E. M. Knorr and R. T. Ng. Algorithms for mining distance based outliers in large datasets. In Proceedings of the 24rd International Conference on Very Large Data Bases, pages 392–403, San Francisco, CA, USA, 1998.
[18] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. pages 144–155,1998.
[19] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. SIGMOD Rec., 25(2):103–114, 1996.
[20] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient clustering algorithm for large databases. SIGMOD Rec.,27(2):73–84, 1998.
[21] "2014 SIGKDD Test of Time Award". ACM SIGKDD. 2014-08-18. Retrieved 2016-07-27.

AUTHOR PROFILE

Sarita Tripathy received B.E degree from National Institute of Science and Technology, Berhampur Odisha, in year 2003 and M.Tech degree from College Engineering and Technology Bhubaneswar in Computer Science and Engineering in the year 2008,and is currently pursuing her Ph.D in the field of Data Mining from KIIT University Bhubaneswar Odisha. Moreover, she is presently working as Assistant Professor in KIIT University Bhubaneswar.Her research area includes: Data Mining and Soft computing.

Dr.Laxman Sahoo is currently working as a Professor, in School of Computer Engineering, KIIT University, Bhubaneswar. He was Director at BITS Mesra. His research area includes: Distributed Computing, Distributed database,Soft computing.