https://doi.org/10.48047/AFJBS.6.15.2024.4860-4870

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

An Efficient Approach for Detecting Outliers in data Using Advance Segment Intelligence Chronicle Data Detection

N. Sivashanmugam, Assistant Professor, Tamilnadu Open University.

Volume 6, Issue 15, Sep 2024

Received: 15 July 2024

Accepted: 25 Aug 2024

Published: 05 Sep 2024

doi: 10.48047/AFJBS.6.15.2024.4860-4870

ABSTRACT

The modern world all of them using internet and searching data according to their needs but the result of the data is not accurate some added impurities added in the data that impurities has removing is the important concept of the data mining in the dataset. In our new approach we are going to eliminate the impurities in the data. Those impurities are called as the outliers. We are going to use the Advance Segment Intelligence Chronicle Data Detection algorithm to eliminate the outliers in the searching data's and give the best result to the user's need in this methodology the data searching is the best performance and it going give the exact result to the user. During the process of cluster the data has comes like bunch of the data that data we are going to eliminate the impurities. Basically the clustering two type of clustering first one data mining and spatial mining these are the two types. In that we are going to use the data mining and eliminate the impurities like fraud detection and removing the unwanted data's. Here we are going to use the three values to find the outliers. First value tendency this tendency value base on the local behaviour and second value has used in the kernel k clustering the type of local outlier factor based method and third value has the threshold. It used to give the better performance of the data clustering.

Key Words: Data Mining Outline Detection, Data Uncertainty.

1. INTRODUCTION

To day every one surfing the internet and searching the any data from the database server getting their result in these days every searching the data has comes like a dataset that data set has give to the user. The user has find the actual data from the data set it is very tedious process to search the data from the data set. In that particular impurities data is called as the outlier data that data has

giving the irritation to the use because user should not get the appropriate results it is very difficult to get their original data. So that we are going to eliminate the outlier data from the data set it is very important and very useful to the user's to search the data. in medical application to find the data to easily retrieve the data in the scanning methodology.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

And also outlier detection has used in the wide range of application fraud detection method in credit cards, insurance and health care methodology so much of the outlier detecting methods detect the existing data. The outlier detection has different types clustering based, density based. In that we are going to develop the algorithm to remove the outlier data' from the original searching data. normally the outlier detection has handle with imperfect labels propose the approach by generating the description on training data set. The SVDD is the algorithm has used to detect the outlier in the various domain fields. Support vector data description is the commonly used algorithm. In that there are different types of methodology have to be the outlier detection technique they are partitioning method , hierarchical method ,density based method, grid based method, model based method and constraint based method.

Here we are going to cluster the data in different application like research, pattern recognition, data analysis and image processing. In existing approach they are using the outlier detection using the two values to calculate the outliers so it will give the result has little bit slow and not in the searched correct output format so we are going to make a new approach to find the outlier in the data mining clustering.

2. RELATED WORK

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web. The data mining contain the some important parts in the data mining data cleaning is used to reduce the noise and inconsistent data, data integration is used to multiple the data , data selection analyse the task are retrieved from the database data transformation performing the aggregation operation data mining extract the data patterns pattern evaluation patterns are evaluated, knowledge presentation knowledge is represented.

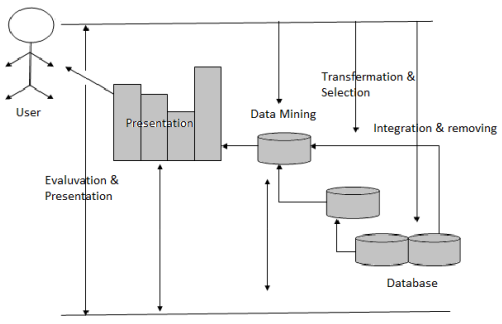


Figure 1

2.1 PARTITIONING METHOD

In that method database has take it as a n objects and method has construct the symbol k of portioned data. Each cluster will represent the data has K<=N each group have one object. In that partition method k has construct the initial portioning. In this technique has improved the partition by moving objects in a single group.

2.2 HIERARCHAL METHOD

Decomposition of the data objects in this methodology we going to use the two methods agglomerative approach method, divisive method the agglomerative method has using the bottom up approach and make it a separate group merging the object in to group using nearby nodes. It continuous up to al the nodes come in to a single object or meet the condition.

2.3 DIVISIVE APPROACH

This approach is directly proposed to agglomerative approach (top-down approach) here all the object has combined into a one single cluster that cluster has convert into to the single or separate cluster is met the condition or terminate the condition once it has split or merge the data objects we can't do the undo.

2.4 DENSITY BASED METHOD

Serially growing the density has in the neighbourhood exceed the threshold this method contain the minimum least number of points in the density. Constrained based method clustering is doing by the incorporation user application oriented constraint gives the proper communication at the time of the cluster process constraint has denoted by the user or application is called as the constrained based methodology.

Text database is the large amount of data documents this data documents contains the several information's like emails web pages etc in the text data contain the so much of unwanted data has like the project document contain the abstract, introduction, reference, propose and so much of unstructured or in appropriate data will be there in the project document so that the data mining is the important and very useful.

2.5 GETTING INFORMATION

Information getting is the process because every text document doesn't contain the only the text data it have the some other information's diagrams and other events will be there in the particular text document. The main problem the data retrieval has comes by the user search query string or information of keywords.

Outlier detection is the method of broad spectrum technique. It is used for detecting the outlier fundamentals. It has many approaches detect the outliers it is the abnormal condition. Outlier detection can find the unwanted dataset factors fault factory protection of normal and abnormal data in the real time. It used by the time series analysed by the statistics [1].

Anomaly detection attempt to provide the reduce the discrete symbolic sequence to identify the these problem normal database according to the normal sequence sub sequence with large sequence in that we have the lot of solution have to find the outlier detection of anomaly discrete function. Normally outliers are identify as the anomalies statistics of the data. Data is retrieved from the different set of locations using many process so that outlier detection is used to identify the abnormal data it will reduce the or impact of the data.[2][3].

Anomaly detection points the problem of detecting structures in data is not Exact to expected data by the user. These undefined structure are denoted as unwanted, outliers, , exceptions, different

application. In these, anomalies, outliers are two terms used in the method of anomaly finding process. [4].

Outlier is stands for a looking that "shows" to be unwanted with none looking in the form of data. Outlier is a less possibility that it develop from the related spatial arrangement as of the looking dataset. Apart from this, an ultimate value is an looking have a low possibility of appearance but not be stable show to original from the data set [5].

Huge amount of data that are stored in the databases, Maximum requirement for competent analysis way to prepare the information having internally in the data. Knowledge discovery databases (KDD) is stands as the unimportant and identifying the actual, useful, and ultimate known knowledge from the data set [6].

A distance-based Method use the wavelength method for analyse each pair of object of the data. Distance-based explanation represents an important tool in the data analysis method. These statements are computationally effective, so distance-based outlier ranks are monotonic non-increasing methods of the section of the database. Now a day, lots of algorithms has been introduced detecting distance based outliers very quickly and efficiently [7].

The research of kind condition is nontrivial factor of huge data set and lot's of level of differentiation proposed at various levels of the research. The statics is further complex by the huge difference that may occur at different possibility used to investigate the similar gene. We are find that, after doing use of the hold information provided by the MM probes, we utilize a set of 21 Hu6800 ever (Hofmann et al., 2002) to enlarge our treatment. This kind of data set in terms of excellence and sample size, of a data set from a particular lab experiment. We have test the methodology to different sets of arrays from different lab and get same results [8][9].

For a Better Classification the initial stage of the data processing is the important. Better processing has the reduce the unwanted data and retain the good quality of the information is possible. Normally we have the more number of objects in the training set basically use the small amount of features most of the procedures should not find the good classification of the finishing data boundaries. The better pre-processing the more number of object per feature has reduce the differentiation problem can be easily solved [10].

A Bulk of new projects has developed unsure databases about particular application needs. For imagine, the control application familiarize query reworking formula to take out to remove and invariable answers from impure data under possible semantics. Functions are also introduced to derive possibilities of impure details. Major aspects of the familiar application is that it allow real time and dynamic data removing such a kind of removing and consistent respond may be attend for sequence[11].

Preparation graphs having a examine graph from a huge graph information is a important process in many graph-based Projects, including intermediate compounds identifying, protein prediction, and pattern recognition. Whatever, graph data processing by these projects is an often unwanted error, uncompleted, and impure because of the way the data is given. we analyse sub graph query on the impure graphs. Normally , we look at the problem of give the threshold-based possibility query on the a huge abnormal graph database with the possibility of semantic world[12][13].

2.6FUNDAMENTAL MEASURING FOR RETRIEVING THE TEXT

We have to check the accuracy of the system has to be retrieve the data based on the users given input if the document contain the relevant query in the document according to users given is called as the {relevant} ∩ {relevant}

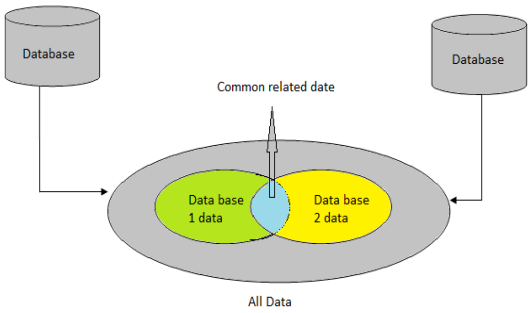


Figure 2

Precision

Percentage of retrieved documents are called as the precision we can defend as

$$Precision=|\{relevant\}\cap\{ retrieve \}|/|retrieve|$$

Recall

Recall is the percentage of document relevant in the user's query or input.

Recall = $Precision=|\{relevant\}\cap\{ retrieve \}|/|Relevant|$

3. PROPOSED DESIGN

Our proposed new Advance Segment Intelligence Chronicle Data Detection algorithm has used to remove the out layer in the data according to the user's search or key word. In this algorithm has we are using the three important functions to work the algorithm they are tendency, kernel and threshold in this three things has to do the performance of our algorithm. In earlier algorithm has detecting the values likelihood and kernel this is very difficult to calculate the likelihood values so that we are going to deal with a problem through the tendency value

Tendency value has calculated using the two ways single tendency and multiple tendencies.

3.1 SINGLE TENDENCY

Tendency calculated by the user query and it convert into the degree of the user query. E.g. user query is take it as x(t) the degree of the converted value has been m[x(t)]

3.2 MULTIPLE TENDENCIES

The multiple tendency has calculated by the user input and converted in to series of the covariance degree that value has represented by (x(i), mt(xi),x(j) mn(xj) in that equation has the xi is represent the relation of the user input keyword.

In this two types of the tendency value has calculating and generating the tendency value for each input data.

4. ADVANCE SEGMENT INTELLIGENCE CHRONICLE DATA DETECTION ALGORITHM

Input: Input Value or key word

1: Initialize input key word S, S: = {nr'1 ….. nr'l};

2: compute t, t: = Tendency value

3: for checking s=t do

4: compute k where k: = Kernel value

5: check s=k do

6: compute th , th= threshold

7: check s= th; then do

S(t),S(k),S(th)= s(t,k,th);

S=(t,k,th)

8: return s;

9: else

10: check S

4.1 ARCHITECTURE DIAGRAM

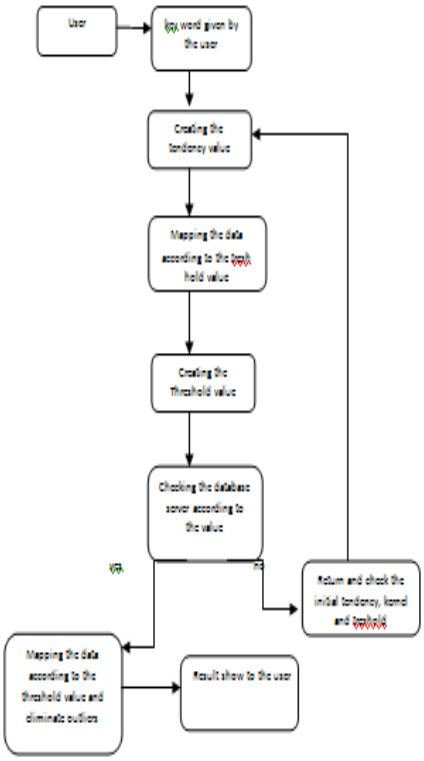


Figure 3

5. EXPERIMENTS AND RESULT ANALYSIS

Return and check the initial tendency, kernel and treshold

The existing method has many drawbacks to the data clustering because it find the kernel only so the data accuracy has not to estimated leve algorithm has give the best result compare to the existing method

5.1.1 CONFUSION MATRIX

		Target class	Negative class
Predicted table	Target class	True Positive	False Positive
	Negative class	False Positive	True Positive

Table 1

In this confusion matrix has contain the target class and negative class and the target class has contain the value of the true positive and false positive has been vice versa in the negative class

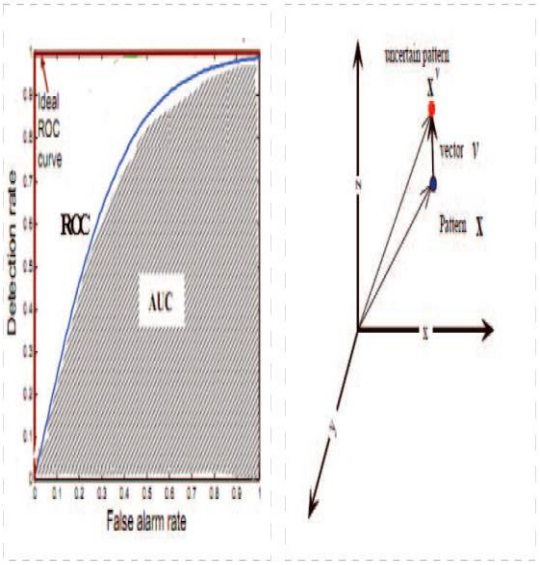


Figure 4

5.1.2 DATA SET DESCRIPTION

Data set	Description	Data Set	#of features
Abalone	Classes 1-8 rest	4177	10
Spam base	Other vs spam	4601	57
Thyroid	Class 2 vs rest 3	3428	21
letter	Class 1 vs rest	6238	617

Table 2

These are the existing algorithm result but our algorithm should be work like that

5.2 CONFUSION MATRIX

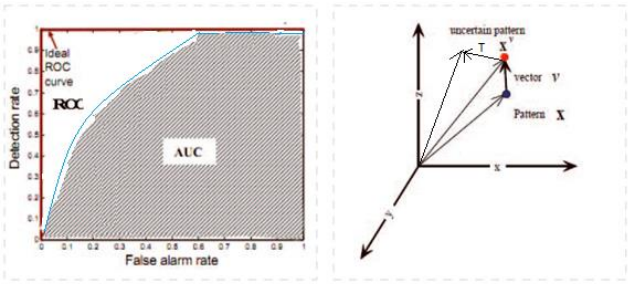
		Target class	Negative class	Threshold
Predicted table	Target class	True Positive	False Positive	Threshold
	Negative class	Threshold	Threshold	True Positive
	Threshold	False Positive	True Positive	False Positive

Table 3

In the above table contain the three values we are using our algorithm the third value is called as the threshold it used find the data with complex method.

5.2.1DATA SET DESCRIPTION

Data set	Description	Data Set	#of features
Abalone	Classes 1-8 rest	5132	36
Spam base	Other vs spam	6695	82
Thyroid	Class 2 vs rest 3	5482	49
letter	Class 1 vs rest	8244	923



(a) (b)

Figure 5

CONCLUSION AND FUTURE WORK

In our algorithm consist of the result has been compare to the earlier algorithm it will be the better and gives the result in 90% of the user given query or key word. So this algorithm has the successful algorithm compare to the earlier outlier detection algorithm if the algorithm contain the three values so that the comparing and cluster and mapping the data set has checking the correct value and it convert to the actual result of the search by the user. The future work of this paper has reduce the complexity of the values is tendency, kernel and threshold value and reduce the timing value of the algorithm and it should gives the ninety percentage of the result but in future we have to obtain the result has more than ninety percentage compare to the our algorithm then the data mining or text mining has is very useful for the further investigation and research in the data or text mining in this domain.

7. REFERENCE

[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] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection for discrete sequences: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.
[5] V. Barnett and T. Lewis, *Outliers in Statistical Data.* Chichester, U.K.: Wiley, 1994.
[6] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2000, pp. 93–104.
[7] S. Y. Jiang and Q. B. An, “Clustering-based outlier detection method,” in *Proc. ICFSKD*, Shandong, China, 2008, pp. 429–433.
[8] C. Li and W. H. Wong, “Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection,” in *Proc. Natl. Acad. Sci. USA*, 2001, pp. 31–36.

[9] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.
[10] D. M. J. Tax, A. Ypma, and R. P. W. Duin, "Support vector data description applied to machine vibration analysis," in *Proc. ASCI*, 1999, pp. 398–405.
[11] 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.
[12] L. Chen and C. Wang, "Continuous subgraph pattern search over certain and uncertain graph streams," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 8, pp. 1093–1109, Aug. 2010.
[13] A. Boukerche, R. B. Machado, K. R. L Juca, J. B. M. Sobral, and
M. S. M. A. Notare, "An agent based and biological inspired real-time intrusion detection and security model for computer network operations," *Comput. Commun.*, vol. 30, no. 16, pp. 49–60, 2007.
[14] A. O. Tarakanov, "Immunocomputing for intelligent intrusion detection," *IEEE Comput. Intell. Mag.*, vol. 3, no. 2, pp. 22–30, May 2008.
[15] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, "Anomaly detection via online over-sampling principal component analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1460–1470, May 2012.
[16] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *Proc. ICML*, San Francisco, CA, USA, 2000, pp. 255–262.
[17] F. Chen, C. T. Lu, and A. P. Boedihardjo, "GLS-SOD: A generalized local statistical approach for spatial outlier detection," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, 2010, pp. 1069–1078.
[18] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," *Knowl. Inform. Syst.*, vol. 26, no. 2, pp. 309–336, 2011.
[19] R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski, "Clustering approaches for anomaly based intrusion detection," in *Proc. Intell. Eng. Syst. Artif. Neural Netw.*, 2002, pp. 579–584.
[20] Y. Shi and L. Zhang, "COID: A cluster-outlier iterative detection
approach to multi-dimensional data analysis," *Knowl. Inform. Syst.*, vol. 28, no. 3, pp. 709–733, 2011.
[21] A. Ghoting, S. Parthasarathy, and M. E. Otey, "Fast mining of distance-based outliers in high-dimensional datasets," *Data Min. Knowl. Discov.*, vol. 16, no. 3, pp. 349–364, 2008.
[22] A. Ghoting, S. Parthasarathy, and M. Otey, "Fast mining of distance-based outliers in high-dimensional datasets," *Data Min. Knowl. Discov.*, vol. 16, no. 3, pp. 349–364, 2008.
[23] F. Angiulli and F. Fassetti, "Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 4, pp. 1–57, 2009.
[24] V. Niennattrakul, E. J. Keogh, and C. A. Ratanamahatana, "Data editing techniques to allow the application of distance-based outlier detection to streams," in *Proc. IEEE ICDM*, Sydney, NSW, USA, 2010, pp. 947–952.
[25] K. Bhaduri, B. L. Matthews, and C. Giannella, "Algorithms for speeding up distance-based outlier detection," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, 2011, pp. 859–867.
[26] E. M. Jordaan and G. F. Smits, "Robust outlier detection using SVM regression," in *Proc. IJCNN*, 2004, pp. 1098–1105.
[27] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *Proc. ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, 2006, pp. 504–509.
[28] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, Dec. 2005.
[29] J. Theller and D.M. Cai, "Resampling approach for anomaly detection in multispectral images," in *Proc. SPIE*, Orlando, FL, USA, 2003, pp. 230–240.
[30] D. Tax and R. Duin, "Outlier detection using classifier instability," in *Proc. Adv. Pattern Recognit.*, London, U.K., 1998, pp. 593–601, LNCS.
[31] 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.
[32] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. IJCAI*, San Francisco, CA, USA, 2001, pp. 973–978.
[33] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[34] M. V. Joshi and V. Kumar, “CREDOS: Classification using ripple down structure (a case for rare classes),” in *Proc. SIAM Conf. Data Min.*, 2004.