The Proposed in Medical Outlier Detection using Optimized Support Vector Machine

Er. Divya Sharma[1], Er. Ajay Sharma[2]
[1]M.Tech Scholar, [2]Associate Professor
Department of Computer Science& Engineering, Amritsar College of Engineering & Technology, Amritsar, Punjab

Abstract **–** An important application in sensor network in outlier detection like normal and abnormal action detection, animal behave alter etc. It's a hard issue since global data about information regarding data divisions must be called to verify the outiers. In this paper , we discussed with the proposed approach in the research area. In proposed work, We divide the data into in two clusters i.e Cluster1 an Cluster2. We implement the K-means clustering approach to divide the data into two section detected or not detected data. We optimize the outlier data with bacteria Foraging Optimization approach. In BFOA algorithm based on further steps: (i) population Size (ii) Rotation (tumble and swim) (iii) dispersal (iv) reproduction the abnormal data. That means BFOA approach optimize the relevant data. The classification algorithm used to classify the outliers based on training and testing phase. In this technique, to use a optimize the communication cost. Rater grouped data in a single position for centralized processing.

Keywords **-** Outlier Detection, Centralized data, BFOA (bacteria Foraging Optimization Approach) and Classification (SVM).

I. INTRODUCTION

Data Mining is a quickly evolving area of investigation that is at the connection of several disciplines, including statistics, temporal pattern recognition, temporaldatabases [1]. By adding to the growth in the quantity of data, the variation of available data has also increased emails, blogs, transaction data, and billions of web pages create tera-bytes of new data every day[2]. Numerous of these data streams are formless, adding to the difficulty in analysing them. This increase in both the volume and the variety of data requires advances in methodology to automatically understand, process, and summarize the data.In many data analysis tasks a large amount of variables are being verified or sampled[3] . One of the first steps near obtaining a coherent analysis is the discovery of outlaying observations. Though outliers are often measured as an error or noise, they may carry significant information. Detected outliers are candidates for unusual data that may otherwise undesirably lead to model mis-specification, biased parameter approximation and incorrect consequences. It is therefore significant to identify them prior to modeling and analysis [4].

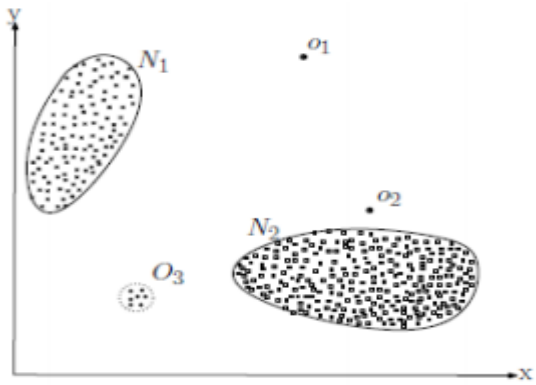

Fig 1. A simple example of outliers in a 2-dimensional data set[4].

An exact definition of an outlier often depends on hidden assumptions regarding the data construction and the applied discovery method.An outlier as an observation that deviates so much from other observations as to arouse thought that it was produced by a different mechanism.

For example, a scheme event may often reflect the activities of an individual in a particular sequence [5].

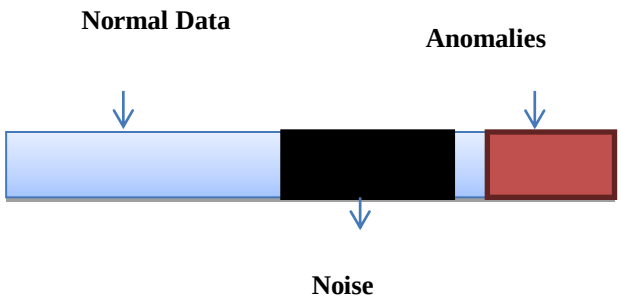

Fig 2.The spectrum from normal data for outliers

The specificity of the sequence is relevant to classifying the anomalous event. Such anomalies are also mentioned to as collective anomalies, because they can only be inferred together from a set or arrangement of data points. Such collective anomalies typically characterize unusual events, which need to be discovered from the data[6].

II. LITERATURE SURVEY

MadhuShukla et al., 2015[12] described that the Data mining is one of the most exciting fields of research for the researcher. As data is getting digitized, systems are getting connected and combined, scope of data group and analytics has increased exponentially. Today, most of the systems generate non-stationary data of huge, size, volume, occurrence speed, fast altering etc. these varieties of data are called data torrents. One of the most recent trend i.e. IOT (Internet Of Things) is also promising lots of expectation of people which will ease the use of day to day happenings and it could also connect organizations and people together. **Dr. S. Vijayarani et al., 2013[13]**defined that the Data mining is extensively studied field of research area, where most of the work is highlighted over knowledgediscovery. Data stream is dynamic research area of data mining. A data stream is an enormous sequence of data elements continuously generated at a debauched rate. In data streams, huge quantity of data continuously introduced and enquired, such data has very large database. The data stream is motivated by emerging applications involving massive data sets for instance, customer click torrents and telephone records, bulky sets of web pages, multimedia data's, and financial transactions and so on. **Christy.A et al, 2015[14]**discussed that Outliers has been studied in a variety of domains including Big Data, High dimensional data, uncertain data, TimeSeries data, Biological data, etc. In majority of the sample datasets available in the repository, at least 10% of the data may be erroneous, missing or not available. In this paper, we utilize the concept of data pre-processing for outlier reduction.**Zili Li et al., 2015[15]**described that Outlier detection is a basic task in system analysis, which is useful in many submissions such as interruption detection, criminal investigation, and information filtering. In this paper, we proposed a hybrid outlier detection approaches in complex systems based on Vertex Dispersed Representation and Local Outlier Factor, with the aim to find abnormal vertexes that are apart from the group or community in complex networks. The proposed outlier detection method based on Vertex Distributed Representation (VDR) and Local Outlier Factor (LOF) is named as VDR-LOF.**Hayfa AYADI et al., 2015[16]**described that Wireless sensor networks are fasting more and more consideration these days. They gave us the coincidental of collecting data from noisy situation. So it becomes conceivable to obtain precise and unceasing checking of different phenomenon. However wirelesses Sensor Network (WSN) is affected by many anomalies that occur due to software or hardware difficulties.

III. ISSUES IN OUTLIER DETECTION

Stream data are produced from the different applicationslike network traffic analysis, sensor network, internet traffic, etc., which may contain attributes that are irrelevant called as noisy attributes which causes challenges in stream data mining process or it may be animalistic behaviour of the system[7]. Outlier analysis is useful in applications like fraud detection, plagiarism, communication network management. For the data stream mining process there are various issues based on the data streams which comes from the single data stream or multiple data streams. In case of single data stream issues involved are discussed below:

- *Transient:* Specific data point is important for a specific amount of time, after it is discarded or archived[8] .
- *The Notion of time:* Timestamp attached with data which give temporal context, based on that temporal context data point is processed.
- *The Notion of infinity:* Data stream is produced indefinitely from the source, thus at particular time whole dataset is not available so summary of data points is used.
- *Arrival rate:* Data points arrive at the different rate, so processing of data points.

IV. PROPOSED ALGORITHM IN OUTLIER DETECTION

In this section, we discussed with the proposed algorithm in the outlier detection data mining.

A. *K-means Clustering:* Simply speaking it is a method to classify or to collection your items based on attributes/features into K number of group. K is positive digit amount. The grouping is complete by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the reason of K-mean clustering is to classify the data. In K-means clustering If the numeral of information is less than the numeral of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster numeral. If the numeral of data is higher than the number of cluster, for each data, we calculate the space to all centroid & get the smallest amount distance. This data is said belong to the cluster that has minimum distance from this data [9].

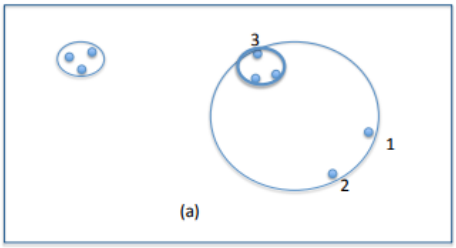

Fig 3. The k-means algorithm is extremely sensitive to outliers. By removing two points (1) and (2), we can obtain

much tighter clusters (the bold circle near 3). The objective of this paper is to obtain a tight clustering and report the outliers in an automatic fashion[9].

B. BFOA (Optimization): This method is used for locate, handling, & ingesting the food. Throughout foraging, a bacterium can exhibit two different actions:
(i) Tumbling or spinning. The tumble action modifies the compass reading of the bacterium. During spinning means the chemo taxis phase, the bacterium will shift in its recent course.
(ii) Chemo taxis movement is continuous until a bacterium goes in the direction of positive nutrient rise. After a definite number of complete swims, the best halves of the inhabitants undergo the original and (iii) Eliminate the rest of the population. In order to escape local optima,
(iv) An removal dispersal event is accepted out where some bacteria are liquidate at random with a very small chance and the new replacement are initialized at random locations of the look for space[10].
C. *Support Vector Machine (SVM) :*"Support Vector Machine" (SVM) is a managed machine learning algorithm which can be used for both classification & regression challenges. However, it is mainly used in classification problems. In this algorithm, we plot each data item as a fact in n-dimensional space (where n is number of benefit you have) with the value of each feature being the value of a particular coordinate. Then, we perform sorting by finding the hyper-plane that differentiate the two classes very well. Support Vectors are basically the co-ordinates of individual comment. Support Vector Machine is a frontier which best segregates the two classes [11].

VI. SIMULATION WORK

In this simulation work, we discussed with the proposed work in outlier detection. We work on medical diabetes data to detect the outlier in two form i.e sick and healthy.

Upload Dataset: upload the dataset in medical diabetes in outlier detection. We search the dataset in UCI machine learning repository in diabetes patients.

Attribitues: To define the li of attributes in this dataset.

Clustering :we implement the clustering approach to separate the attributes in the form of clusters like CLUSTER 1 and CLUSTER 2.

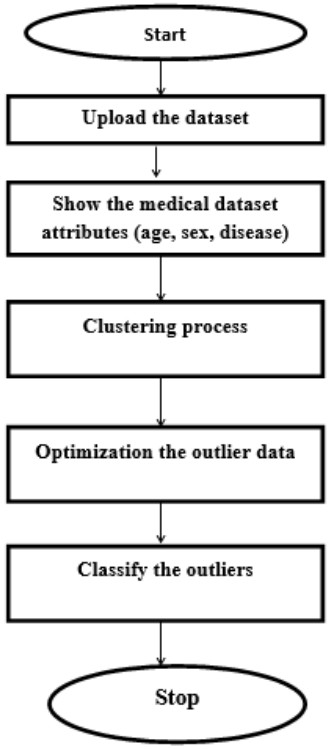

Fig 4. Proposed Flow chart

Optimization: To reduce the cluster attributes with the help of BFOA approach. IN BFOA using the following steps:

(i) Rotation
(ii) Elimination
(iii) Dispersal
(iv) Reproduction.

Classification : In classification approach to classify the training section and testing section . We detect the outlier data with SVM approach.

V. PROPOSED RESULTS

In this section, we explained in proposed results with K-means clustering approach in diabetes detection in data mining.

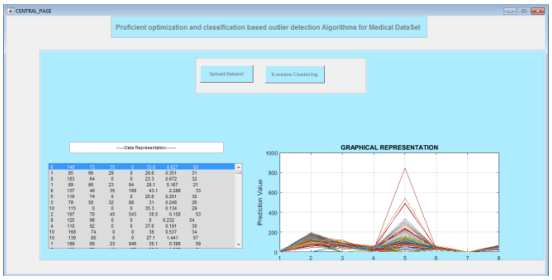

Fig 5. Upload Dataset

The above figure shows that the upload the dataset in .xls file. We select the dataset form UCI machine Learning Repository site in MATLAB simulation Tool. We represent the dataset in list box and graphical representation in axes.

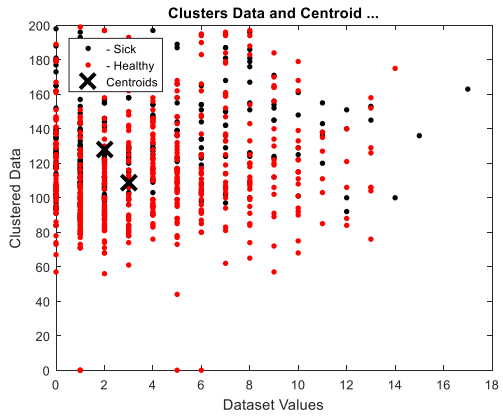

Fig 6. K-means Clustering Approach

The above figure shows that the k-means clustering approach in two categories. In k-means clustering approach to identify the clusters i.e cluster1 and cluster2. It separates the disease attributes in different sections.

VII. CONCLUSION

In this paper, we have proposed a new approach, based on optimized Support Vector Machines, to anomaly detection in computer security. Experiments with the DAIBETESdata set show that k-means clustering can provide good generalization ability and effectively detect outlier in the presence of noise. The running time of k-means clustering can also be significantly reduced as they generate fewer clusters than the conventional K-means clustering approaches.It involves quantitatively measuring the robustness of k-mean clustering over the noisy training data and addressing the fundamental issue of the unbalanced nature between normal and intrusive training examples for discriminative anomaly detection approaches.

VIII. REFERENCES

[1]. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and data Engineering 8, no. 6 (1996): 866-883.

[2]. Keim, Daniel A. "Information visualization and visual data mining." IEEE transactions on Visualization and Computer Graphics 8, no. 1 (2002): 1-8.

[3]. Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok. "A data mining framework for building intrusion detection models." In Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on, pp. 120-132. IEEE, 1999.

[4]. Hall, Mark A., and Geoffrey Holmes. "Benchmarking attribute selection techniques for discrete class data mining." IEEE transactions on knowledge and data engineering 15, no. 6 (2003): 1437-1447.

[5]. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17, no. 3 (1996): 37.

[6]. Agrawal, Rakesh, Johannes Gehrke, DimitriosGunopulos, and PrabhakarRaghavan. Automatic subspace clustering of high dimensional data for data mining applications. Vol. 27, no. 2. ACM, 1998.

[7]. Han, Jiawei, Jian Pei, and MichelineKamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[8]. Berkhin, Pavel. "A survey of clustering data mining techniques." In *Grouping multidimensional data*, pp. 25-71. Springer Berlin Heidelberg, 2006.

[9]. Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31, no. 8 (2010): 651-666.

[10]. Ali, E. S., and S. M. Abd-Elazim. "BFOA based design of PID controller for two area Load Frequency Control with nonlinearities." *International Journal of Electrical Power & Energy Systems* 51 (2013): 224-231.

[11]. Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 32-36. IEEE, 2004.

[12]. Shukla, Madhu, Y. P. Kosta, and PrashantChauhan. "Analysis and evaluation of outlier detection algorithms in data streams." In *Computer, Communication and Control (IC4), 2015 International Conference on*, pp. 1-8. IEEE, 2015.

[13]. Vijayarani, S., and P. Jothi. "An efficient clustering algorithm for outlier detection in data streams." *International Journal of Advanced Research in Computer and Communication Engineering* 2, no. 9 (2013): 3657-3665.

[14]. Christy, A., G. Meera Gandhi, and S. Vaithyasubramanian. "Cluster Based Outlier Detection Algorithm for Healthcare Data." *Procedia Computer Science* 50 (2015): 209-215.

[15]. Li, Zili, and Li Zeng. "A Hybrid Vertex Outlier Detection Method Based on Distributed Representation and Local Outlier Factor." In *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015 IEEE 12th Intl Conf on*, pp. 512-516. IEEE, 2015.

[16]. Ayadi, Hayfa, Ahmed Zouinkhi, BoumedyenBoussaid, and M. NaceurAbdelkrim. "A machine learning methods: Outlier detection in WSN." In *Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2015 16th International Conference on*, pp. 722-727. IEEE, 2015.