(IJRMST) 2018, Vol. No. 5, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

LEVERAGING THE SUPPORT VECTOR MACHINE IN CONJUNCTION WITH BACTERIA FORAGING OPTIMISATION IN DETECTING OUTLIERS IN THE MEDICAL DIAGNOSTIC

Himanshu Dahiya

Bachelor of Technology(B.Tech),IT

Manipal University, Jaipur-303007(Rajasthan), India

ABSTRACT

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

I. INTRODUCTION

Data Mining is a quickly evolving area of investigation that is at the connection of several disciplines, including statistics, temporal pattern recognition, and temporal databases. In many data analysis tasks, a large number of variables are being verified or sampled. One of the first steps near obtaining a coherent analysis is the discovery of outlying 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 misspecification, biased parameter approximation, and incorrect consequences. It is, therefore, significant to identify them prior to modeling and analysis.

An exact definition of an outlier often depends on hidden assumptions regarding data construction and the applied discovery method. An outlier is an observation that deviates so much from other observations as to arouse thought that a different mechanism produced it. For example, a scheme event may often reflect the activities of an individual in a particular sequence.

The specificity of the sequence is relevant to classifying the anomalous event. Such anomalies are also mentioned 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.

(IJRMST) 2018, Vol. No. 5, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

II. LITERATURE SURVEY

Dr. S. Vijayarani et al., 2013 defined that Data mining is an extensively studied field of the research area, where most of the work is highlighted over knowledge discovery. Datastream is a 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, a huge quantity of data continuously introduced and enquired; such data has a 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, and financial transactions, and so on.

Zili Li et al., 2015 described that Outlier detection is a basic task in system analysis, which is useful in many submissions such as interruption detection, a 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 described that Wireless sensor networks are fasting more and more consideration these days. They gave us the coincidental of collecting data from a 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 different applications like network traffic analysis, sensor network, internet traffic, etc., which may contain irrelevant attributes called noisy attributes, which causes challenges in stream data mining processes, or it may be animalistic behaviour of the system. 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 come from the single data stream or multiple data streams. In the 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.

- The Notion of time: Timestamp attached with data that give temporal context, based on that temporal context data point is processed.

- The Notion of infinity: Datastream is produced indefinitely from the source, thus at a particular time, the whole dataset is not available, so the summary of data points is used.

- Arrival rate: Data points arrive at different rates, so processing of data points.

(IJRMST) 2018, Vol. No. 5, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

IV. PROPOSED ALGORITHM IN OUTLIER DETECTION

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

A. K-means Clustering: Essentially, it is a technique to order or to gathering your things dependent on characteristics/highlights into K number of gathering. K is a certain digit sum. The gathering is finished by limiting the entirety of squares of separations among information and the comparing group centroid. In this way, the reason of K-mean bunching is to arrange the information. In K-implies bunching If the numeral of data is not exactly the numeral of the group then we dole out every datum as the centroid of the group. Every centroid will have a bunch numeral. In the event that the numeral of information is higher than the quantity of bunch, for every datum, we figure the space to all centroid and get the littlest sum separation. This information is said to have a place with the bunch that has the least good ways from this information.

Fig 2. The k-means calculation is incredibly touchy to anomalies. By evacuating two points (1) and (2), we can acquire a lot more tightly groups (the intense hover almost 3). The goal of this paper is to get a tight bunching and report the anomalies in a programmed manner.

B. BFOA (Optimization): This technique is utilized to find, taking care of, and ingesting the nourishment. All through scavenging, a bacterium can display two distinct activities:

(i) Tumbling or spinning. The tumble action modifies the compass reading of the bacterium. During spinning means the chemotaxis phase, the bacterium will shift in its recent course.

(ii) Chemo taxis development is persistent until a bacterium goes toward positive supplement rise. After a clear number of complete swims, the best parts of the occupants experience the first

(iii) Take out the remainder of the populace. To escape neighbourhood optima,

(iv) An evacuation dispersal occasion is acknowledged out where a few microbes are selling aimlessly with an extremely little shot, and the new substitution is instated at arbitrary areas of the search for space.

C. SVM: “Support Vector Machine” (SVM) is a managed machine learning algorithm that 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 the number of benefits you have) with the value of each feature being the value of a particular co-ordinate. Then, we perform sorting by finding the hyper-plane that differentiates the two classes very well. Support Vectors are the co-ordinates of the individual comment. Support Vector Machine is a frontier that best segregates the two classes.

V. SIMULATION WORK

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

(IJRMST) 2018, Vol. No. 5, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

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

Attributes: 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 2. Proposed Flow chart

Optimization: To reduce the cluster attributes with the help of the 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 the SVM approach.

VI. PROPOSED RESULTS

In this section, I explained in the proposed results with the K-means clustering approach in diabetes detection.

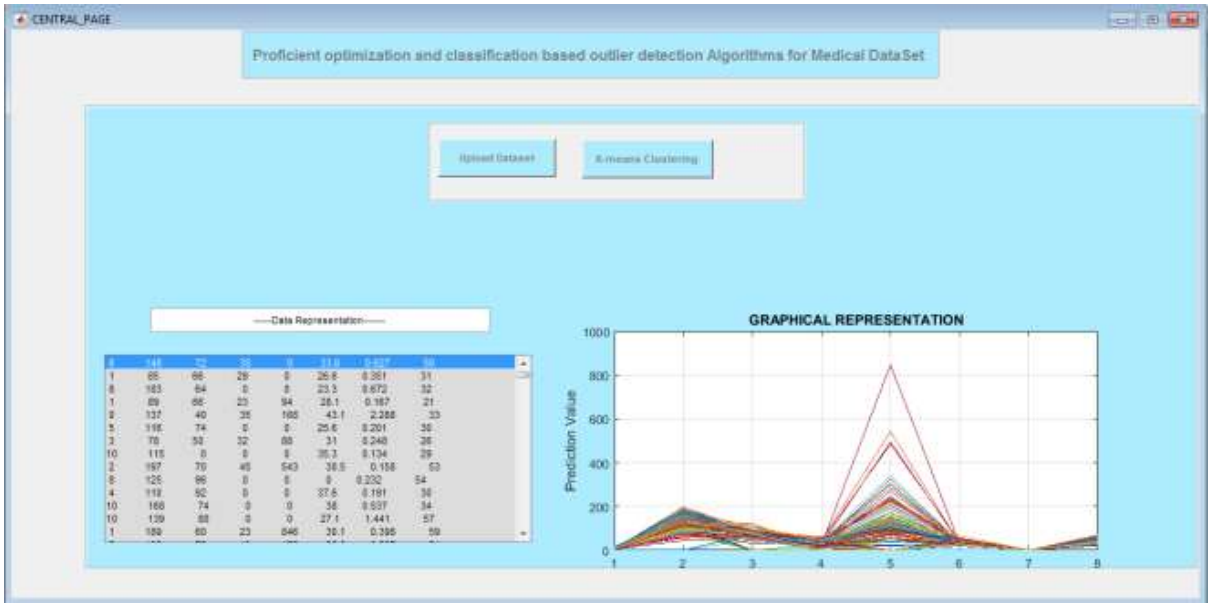

Fig. 1 Uploaded Dataset

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

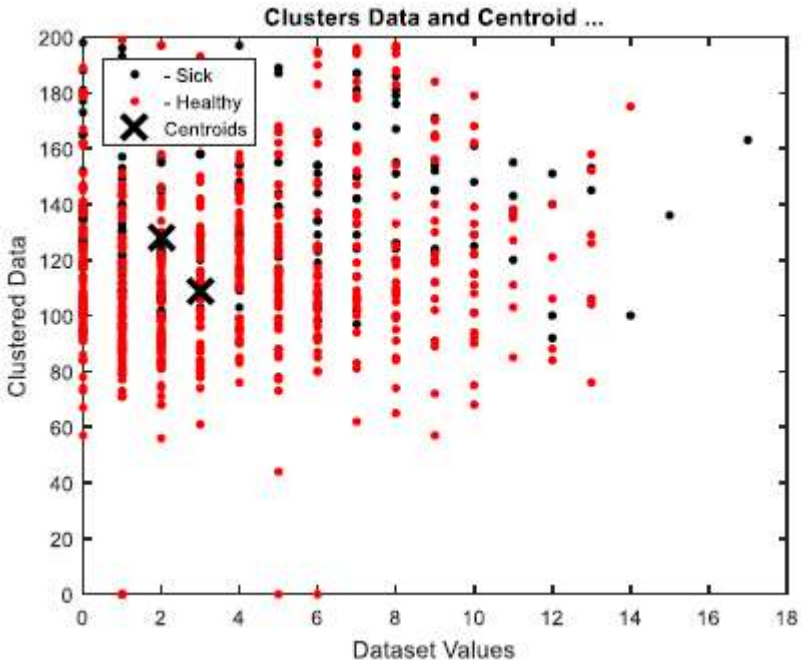

Fig. 2 K-means Clustering Approach

The above figure shows that the k-means clustering approach in two categories. In the k-means clustering approach to identifying the clusters, i.e., Cluster 1 and Cluster 2. It separates the disease attributes in different sections.

VII. CONCLUSION

In this paper, I have proposed a new approach based on optimized Support Vector Machines for anomaly detection in the dataset. Experiments with the DIABETES dataset 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 issue of the unbalanced nature of normal and intrusive training examples for discriminative anomaly detection approaches.