International Journal of Engineering & Technology, 7 (4) (2018) 2881-2885

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.13626
Research paper

An efficient and robust cluster based outlying points detection in multivariate data sets

S. Anitha [1] *, Dr. Mary Metilda [2]

[1] Research Scholar,Bharathiar University, Coimbatore Tamil Nadu, India
[2] Asst. Prof. Queen Mary's College, Chennai,Tamil Nadu, India
**Corresponding author E-mail:anitasenthil@gmail.com*

Abstract

Outlier is a data that does not match to the normal points along with the data set. A recent research focused on number of clusters and distance based outlier detection strategies. In this paper, outliers are identified and eliminated in four phases.Feature selection technique using genetic algorithm is applied to the pre processed data to reducedlarge amount of dataset into significant attributes. Data sets are partitioned as clusters after the feature selection process. Multiple outliers are identified by mahalanobis distance based onthe value of median and covariance matrix. Four real life data sets are taken from UCI machine learning repository and rigorous experiments are conducted by the proposed process of GBFS, CLOPD, and IMO for selecting the relevant subsets, clustering and Outliers removal. These three methods are analysed with data sets and results are depicted. It usedforreducing time complexity and improving clustering and classification accuracy.

Keywords:*Classification; Clustering; Euclidean Distance; Genetic Algorithm; Mahalonobis Distance; Outlying Point Detection.*

1. Introduction

Outlier detection is the process of identifying abnormal instances in data set that differ from a collection of normal data.[1],[11],The detection of outlier (anomaly) instances is a crucial task in data mining.The early detection of mistakes is good for diagnostics in the medical field because it can avoid the life losses.Emerging detection of outliers is more essential in all domains. Besides,removing outliers from the original data set are most important for improving classification accuracy.Clustering is one of the widely used methods in data mining.Statistical outlier detection is mainly concerned with finding outliers in multivariate data. In recent, different outlier analysis methods like distance, density, deviation and cluster based detection approaches are used to get rid of outliers from datasets in all domains [12]. Data reduction and feature selection is an eminentpre-processing technique that are use to select relevant data subsets among the original data. Theserelevant subsets are reduced by Genetic Algorithms (GA). Genetic algorithms are well known heuristic search and optimization technique which yields optimal solution for an optimization problem. [4],[16],[18],[23].

Clustering is a most leading domain in Data mining where the instances to be clustered based on cluster compactness and connectedness. Compactness of the cluster is measured by the overall deviation between the objects and their corresponding cluster centers. Compactness of clusters can be stated as

$$\mathbf{Comp(s)} = \sum_{\mathbf{ck\in s}} \quad \sum_{\mathbf{i\in ck}} \mathbf{d(i,\mu k)} \qquad \mathbf{(1)}$$

Where s represents number of clusters, µk is centroid of the cluster and d(i, µk) is the distance between ith object and the cluster center.[5] Connectedness is evaluated with the help of Closest data points of the cluster centers. It is identified as:

$$\mathbf{Conn(s)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \left(\sum_{\mathbf{j=1}}^{\mathbf{L}} \mathbf{xi, nni(j)} \right) \qquad \mathbf{(2)}$$

Where nni is the j th nearest neighbour i and L determines that are number of neighbours connected to one another.[anusha et al 2016]. K-means clusteringis predominant clustering technique in data miningto separate the instances into number of cluster as desire by the user.

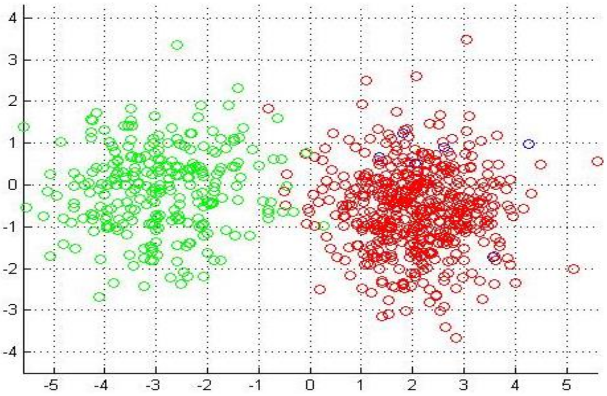


Fig. 1: Clusters with Outliers.

This proposed research work carried out the cluster and distance based outlier detection method which includes three different phases. In first, irrelevant attributes are removed by genetic search method. The selected significant attributes are grouped together by k-means clustering technique in second. In third step, Distance based Outlier detection algorithm has been implemented for discovering outliers. After removing outliers the original multivariate dataset are classified in various classification techniques. Different real lifedata are taken from UCI machine learning repository.

This paper hasorganized as follows: section II exploits pre-processing work which includes data cleaning and normalizing the instance. Section III discusses the feature selectionand clustering techniques. Section IV illustrates proposed outlier detection methodin detail and various classification approaches are used for evaluatingaccuracy.At last, section V concludes our research andthe scope for future work.

2. Background and related work

In recent data mining literature, lot of research efforts directed on outlier detection. outlier is define as data instance which are distinct from rest of data points. And this point contains some useful information on abnormal behaviour.Outlier detection has numerous applications including detecting fraud in network intrusion, business transactional data, medical domain, weather and forecasting and more criminal activities.Many data mining algorithms in their previous work discovered outliers by unsupervised method. [1 - 3], [5 - 7].

Edgar Acuna et al compared the statistical and clustering methods of outlier detection using Mahalanobis distance. Bays algorithm for distance based outliers and LOF(Local Outlier Factor) for density based outlier algorithm. In above works were not discussed about feature selection techniques for clustering and outlier detection. As pointed out in [18] , reducing irrelevant data items by Applying feature selection technique is most advantage for clustering procedures to make clustering accuracy. Charu C. Aggarwal et alproposed new techniques for genetic based outlier detection which find the outliers by studying the behaviour of projections from the data set. Using brute force technique and genetic algorithm, outliers are identified on high dimensional data [1]. Anusha et al proposed an evolutionary clustering algorithm for multi objective optimization to identify more relevant instance based on criterion knowledge from the given data sets andused neighbourhood learning to improve the diversity and efficacy of the algorithm for multi-objective optimization which maximizes the compactness of the cluster and accuracy of the solution through constrained feature selection [4], [5]. Raja, P. Vishnu et al proposed algorithm to detect outliers using genetic algorithm was exceptionally accurate in identifying the outliers the datasetshave tested. The result analysis done on some standard dataset to view accuracy of the algorithm [13].

Chun-Hung Cheng et al explored an interesting techniquepartitioning the dataset with genetic search-based clustering algorithmto achieve high database retrieval performance. By formulating the underlying problem as a travelling salesman problem (TSP) the Genetic algorithm is applied to solve the data-partitioning as Clusters [15] [23]. Hadi. A.S et al proposed the method which exhibits the multiple outliers in multivariatedata sets .This research exploits that the normal and potential outliers are identified with various cut off points using geometric mean value and their results are analysed and depicted [19]. Pachgade.D et al has found the method of outlier detection using clustering and distance based approach depend on the threshold value [21]. Aaron Ceglar et al proposed a method CURIO, used quantisation and implied distance metrics that explicit discovery of outliers. [24]

3. General framework of proposed method

This proposedwork consists of four phases, that are pre-processing the multivariate data instances, feature selection using genetic algorithm (GBFS), clustering (CLOPD) for partitioning the data set. At last outliers are detected andremoved bydistance based outlying points detection (IMO) algorithm. Finally, the resultant instances are implemented in various classification algorithms. This architecture of the proposed work as shown in fig.1

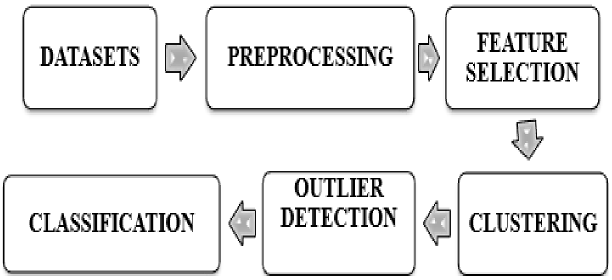


Fig. 2: Architecture of the Proposed System.

3.1. Data pre-processing

In this process, all four raw multivariate data sets were cleaned in order to replace all missing attributes by mean value. In addition, the data are normalized by z-score normalization method.

3.2. Feature selection

Generally, datasets are taken into account for processing in terms of size of the data and number of attribute of the data with class labels. As expected, large amount of data instances were implemented in algorithm was unable to terminate in a stipulated amount of time. Besides, reducing irrelevant attributes arealso increasing the performance of clustering process before identifying outliers. Furthermore feature selection process isperformedwith Correlation coefficient based feature selection (CFS).it is used to determine the significant attribute sunsets and is combined with different types of searchingmethods.In this proposed method Genetic algorithm is implemented for retrieving relevant subsets for further process.It is used to estimate correlation between subset of attributes, class variables and inter correlation between the attributes as well.

Mathematically the Equation for CFS as defined as

$$\mathbf{r\,zc} = \frac{\mathbf{k\bar{r}zi}}{\sqrt{\mathbf{k+k(k-1)\bar{r}ij}}} \qquad \textbf{(3)}$$

Genetic algorithm consists of strings are represented as parameters and theset of parameters depict population.According to Darwin's theory of ''Survival of Fitness'', the strings are randomly selected then induced to crossover and mutation. The genetic algorithm is used to provide a new subpopulation or off-spring. The process of selecting objects from the population and that processcontinuous when required subsets are reached the termination condition or the chosen set of solution attains the fixed number of generation.It is used to classify or cluster the labelled or unlabeled data set. The attribute selection process of proposed methodis given below in algorithm-1.

//Algorithm-1 (Genetic Algorithm Based Feature Selection).

```
Algorithm GBFS ( )
Input:      X: {xi ....xn} be a set of training data points.
P: Initial Population, x1, y1: new population members
Output: Significant subset of Data Points
  1)  P=initial set of population p strings
  2)  Begin (randomly) generating an initial population P.
  3)  If the solution is satisfied then terminate else jump to next
      step
  4)  Evaluate fitness value.
  5)  Initialize number of generation.
  6)  While number generation * 2 ≤ termination condition; do
  7)  Select all the genetic solutions which can transmit to next
      generation (x1,y1)
  8)  Increment number of generation
  9)  Perform crossover operation up to until 50% of bits are
      crossed.
 10)  End while.
 11)  If the solution is efficient then apply mutation.
 12)  Genetic algorithm searches best solution from a large set of
      Solutions
 13)  Go to step2
```

3.3. Clustering

When the attribute reduction is completed, the data size is significantly different from the previous. During the process of GBFS, the irrelevant dataare removed from the dataset andThe resultant data are clustered using partitioned clustering method for separating the data instances in a k no of groups. K is number cluster to be grouped. Interestingly, There are threeclustering approachare used in data mining process, that are raw data based clustering, feature based clustering and model based clustering process. Out ofabove three methods, feature based clustering process is best forproducing more compactness and connectednessfor datapartition. K-means clustering techniques is used for grouping the data as two clusters.as pointed out, the parameter k which id easy to figure out by the user a-priori. When the k value increases, the number of clusters also increases and forms a small subset of clusters with tiny no of instances. In the clustering problem, datasets X = x1, x2......xN^Tε R^{N*D} denotes as matrix, which contains of N samples and each instance vector has D attributes. C = c1, c2.....,cNT ε $[1,M]^{N*1}$ is a vector denoting the cluster assignments. Clustering or partition can be expressed as π = C1, C2, CM of X. An alternative representative element called Cthat is centroid, denotedas Ci, and it can be computed from the instance matrix X (data points). Clustering can done by an Euclidean distance d, between x1 and x2 is formulated as follows

$$\mathbf{D}\ (\mathbf{x_i, x_j})^2 = \| \mathbf{x_i} - \mathbf{x_j} \|^2 = \sum_{k=i}^{D} (xi(k) -, xj(k))2 \qquad (4)$$

D is the distanceof the instances which iscalculated from the centroid point (cluster centerpoint) of cluster depended upon the maximum size of the dataset. Considering the threshold value determined by user which is greater than distance d, that point is Considered as outlier. Based on this strategy, basic outliers are identified from the clusters.After portioning the data set, class labels are added with resultant variables. Next, IMO procedures started with Partitioned datasets with Mahalanobis distance to identify multiple outliers. The Mahalanobis distance is a distance between a point P and a distribution D, It is a multi-dimensional generalization which determines how many standard deviations away P is from the mean of D. Based on the IMO algorithm, and N is size of datasets.

```
Search Method:
Genetic search.
Start set: no attributes
Population size: 20
Number of generations: 20
Probability of crossover: 0.6
Probability of mutation: 0.033
Report frequency: 20
Random number seed: 1
Attribute Subset Evaluator (supervised, Class (nominal): 6 Class):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 1,2,3,4,5 : 5
                    tumor-size
                    inv-nodes
                    node-caps
                    deg-malig
                    irradiat
```

Fig. 3: A Part of Parameters Used in GBFS in Breast Cancer Data Set.

//Algorithm- 2 (CLuster Based Outlying Points Detection).

```
Algorithm CLOPD (DB, d,k,p)
Input:
// DB: x1, x2......,xn be a set of training data points, d:
Distance metric (threshold),
// k: no of nearest neighbours , p: minimum no items needed to accept a cluster, iter : number of iteration
// K: number of clusters, G: assigned groups, C: cluster center point,
Nout: no of outliers,
Output: Number of clusters without outliers Nout
 K←0
For all x € DB do
[o,p]←size (db)
End for
K ← 3
[G, C] ←kmeans(DB,K);
For each iter< p do
Fori =1 to K do
For j =1 to o do
Dist= Ecldistance(xi,Ci) // call Eclidean distance function
Observe the cluster center point Ci
If dist< d then Add ci to cluster centerpoint Ci
Else, label as outliers Nout
End if
End for
End for
Compute log-likelihood for partition obtained.
For all ci consists p then
Include all ci to Ci into group G
End for
End for
Compute mean and standard deviation for the created vector for n rows of DB
Return The Real Data (DB) without outliers (Nout)
```

//Algorithm 3- Identification of Multiple Outliers

Algorithm IMO ( )

Input: X: training dataset with class variables, N: size of X, α: limit of significant.

p: no of variables. S: covariance matrix. , in: inliers, nout: outliers, d: dimentionality.

Output: Data sets without outliers.

1) Initialize N=size(X)
2) For each attribute in dataset X
3) Compute median and covariant for all observations.
4) Calculate Mahalanobis distance for n observation using the value of median and covariant based on p variable.

$$\mathbf{MDi=}\ \sqrt{(\boldsymbol{x}-\boldsymbol{y})\quad {}^{\boldsymbol{T}}\boldsymbol{S}^{-\mathbf{1}}\,(\boldsymbol{x}-\boldsymbol{y})}$$

$$\mathbf{T^2{}_{i\,=}}(\boldsymbol{x}-\boldsymbol{y})\quad {}^{\boldsymbol{T}}\boldsymbol{S}^{-\mathbf{1}}\,(\boldsymbol{x}-\boldsymbol{y})$$

1) if the observation value of data is less than the cut-off (limit of significant)pointreassign the value As ‘o’ as inliers.(in)
2) Else let it is ‘1’ as outliers.
3) Repeat the steps go to step 3 and 4 for rest of the instances
4) After the discordancy test, reject the points are considered as outliers (nout)

4. Empirical results

The proposed methods CLOPD and IMO have been implemented using MATLAB R2013a. WEKA 3.6.13 is used for pre processing the original training data set and its accuracy compares with other data mining algorithms. Multivariate analysis (MVA) is based on the Analysis of two or more variables simultaneously that is why Four different medicalData sets with two class labels are taken from UCI machine learning repository. Then Clusters are formed after irrelevant feature removal and n numbers of outliers are identified using Mahalanobis distance based on median and covariant matrix. Data descriptions are depicted in the table-1. he efficiency of outlier detection techniques can be evaluated by the computational cost, which is known as time & space complexity is graphically rendered as fig-5. There five classifier algorithms are used for evaluating the proposed system, the classifiersare Naive Bayes ,Multilayer Preceptron, Support Vector Machine , Radial Basis Function Network and Random Forest.

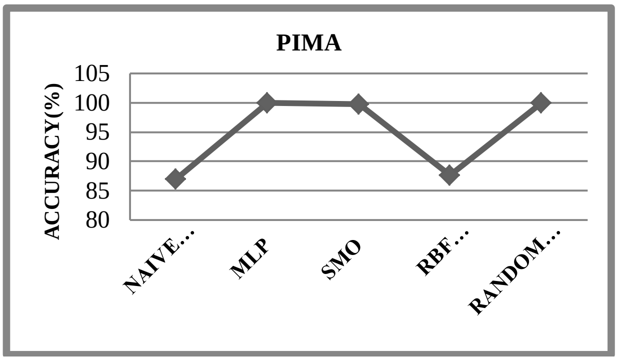
Table 1: Data Descriptions

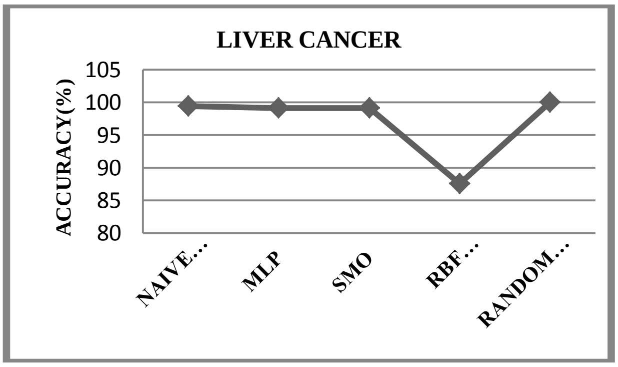
Data set (UCI)	Number of attributes	Number of instances	Class variables
PIMA Indian Diabetes Data	9	768	2
Liver Cancer	7	345	2
ILPD (Indian Liver Patient Dataset)	10	583	2
Brest Cancer (Original)	10	699	2

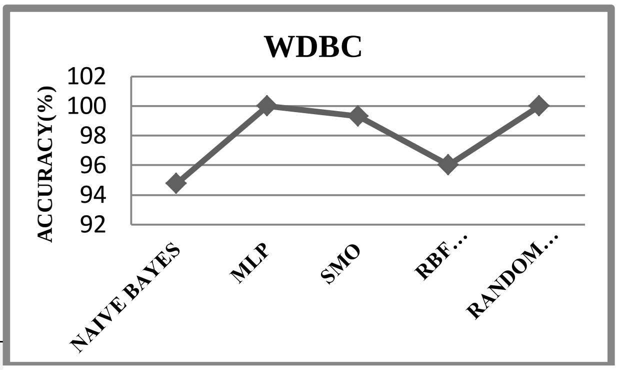
Table 2: Accuracy of Datasets

Datasets	Naive Bayes	MLP	SMO	RBF Network	Random Forest
Wdbc	94.75	100	99.3	96.00	100
Ilpd	96.59	100	99.9	98.96	100
Pima	87.00	100	99.8	87.6	87.5
Liver-cancer	99.4	99.1	99.1	87.5	100

As the result, the performance of classifiers are analysed and their accuracy results were presented in fig-4.







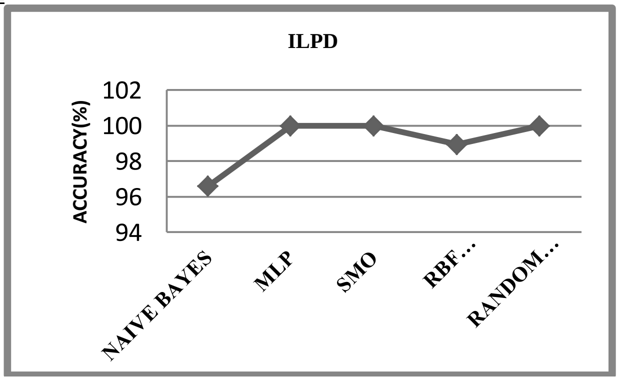


Fig. 4: Classification Accuracy of Four Datasets.

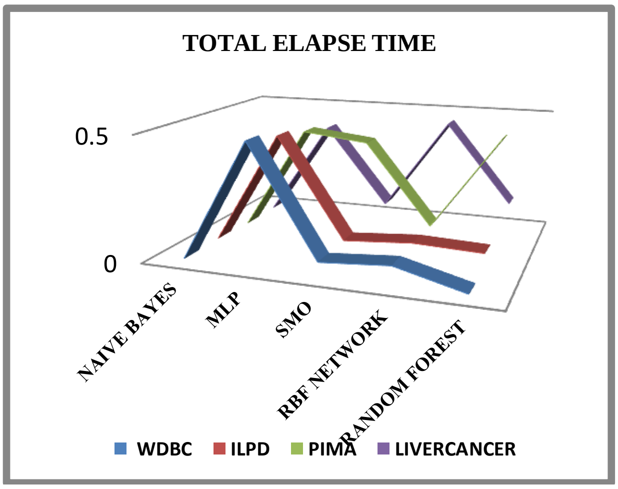


Fig. 5: Total Computation Time.

Evaluation of detected outlierin the data analysis hasessential measuresof Detection Rate, False Alarm Rate and ROC Curves. Intuitively, detection rate gives information about the number of correctly identified outliers, while the false alarm rate represents the number of outliers misclassified as normal data records.Next technique to evaluate accuracy is ROC (Receiver Operating Characteristic) curve can be viewed as figure 8.

Table 3: Outlier Detection Rates of Datasets

DA-TASETS	PIMA		LIVER CANCER		ILPD		WDBC	
Algorithm	Detection Rate (%)	False Alarm Rate (%)	Detection Rate (%)	False Alarm Rate (%)	Detection Rate (%)	False Alarm Rate (%)	Detection Rate (%)	False Alarm Rate (%)
Naive Bayes	87.5	21.6	99.4	33	99.7	33	95.4	8.77
MLP	98.7	27.8	99.1	33.3	99.04	33	99.12	0.2
SMO	87.5	21.6	99.1	33.3	98.36	34	100	4.3
RBF Network	87.5	21.6	87.5	21.6	98.1	32	97.1	8.7
Random Forest	87.5	18.1	100	0	99.87	21	99.2	0.32

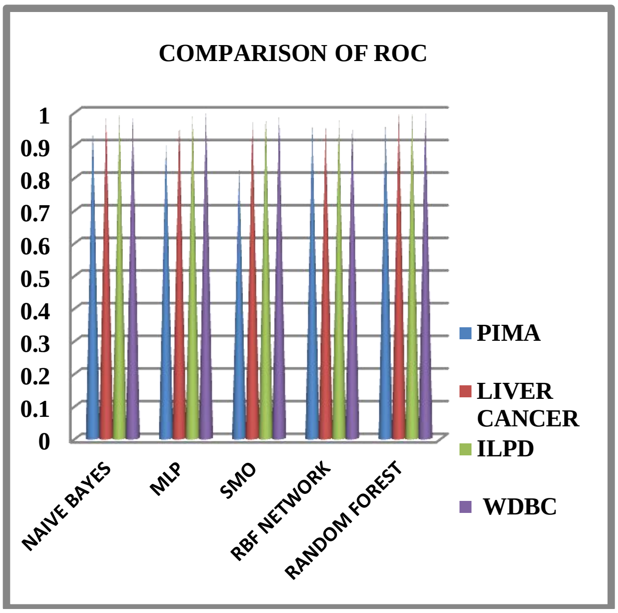


Fig. 6: Comparison of ROC Measures of Data Sets.

5. Conclusion

The proposed methodfocus onfinding outliers in real life datasets during clusters.Outliers are computational errors or misclassified instances that are classified asnormal attributes. Definitely, these points should be removed from the datasets for producing accurate results and decision making in different domains. Interestingly, this proposed robust clustering algorithm, irrelevant features and redundancies also are removed by GBFS algorithm that evaluates the significant subsets of attributes for considering the next phases of cluster analysis. After clustering the datasets, outliers are observed and discarded by robust Mahalanobis distance. The different classification algorithms are applied and explore their results. From the ROC analysis,multilayer perceptron and random forest are providing high accuracy in PIMA Indian diabetes data sets. Besides, on all the five classifiers WDBC dataset having higher accuracy than other data sets. IMO algorithm promotes good observation of outliers during the cluster analysis. Even though the vital outliers are identified, some of critical points could not be observed during the cluster analysis without feature selection. so the dimensionality reduction is an important role in this proposed system and it is significantly simpler than other clustering algorithms with respect to both computational time and rate of error with the highest score of outliers. The future work will be outlier detection on high dimensional dataset and considering the merits and demerits of the proposed system.

References

[1] Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." In *ACM Sigmod Record*, vol. 30, no. 2, pp. 37-46. ACM, 2001.

[2] Anitha, S., and M. Mary Metilda. "A heuristic approach for observing outlying points in diabetes data set." In *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017 IEEE International Conference on*, pp. 199-202. IEEE, 2017.

[3] Anitha.S, Mary Metilda, "A Survey on Cluster Based Outlier Detection Techniques in Data Stream", *International Journal of Data Mining Techniques and Applications (IJDMTA)*, vol. 5(1) pp. 96-101, 2016https://doi.org/10.20894/IJDMTA.102.005.001.023.

[4] Anusha, M., &Sathiaseelan, J. G. R. (2015). An improved K-means genetic algorithm for multiobjective optimization. International Journal of Applied Engineering Research, Special Issue, 10(1), 228–231.

[5] Anusha, M., and J. G. R. Sathiaseelan. "An enhanced K-means ge- *and Computing Research (ICCIC), 2014 IEEE International Conference on*. IEEE, 2014.

[6] Bello-Orgaz, G., & Camacho, D. (2014). Evolutionary clustering algorithm for community detection using graph-based information. In IEEE Congress on Evolutionary Computation (pp. 930–937).

[7] Chatterjee, S., &Mukhopadhyay, A. (2013). Clustering ensemble: A multiobjective genetic algorithm based approach. Procedia Technology, 10, 443–449. https://doi.org/10.1016/j.protcy.2013.12.381.

[8] Halkidi, Maria, YannisBatistakis, and Michalis Vazirgiannis. "Clustering validity checking methods: part II." *ACM Sigmod Record* 31.3 (2002): 19-27.https://doi.org/10.1145/601858.601862.

[9] http://www.ics.uci.edu/mlearn/MLRepository.html.

[10] Jing, L. P., Ng, M. K., & Huang, Z. X. (2007). An entropy weighting k-means algorithm for subspace clustering of high dimensional sparse data. IEEE Transactions on Knowledge and Data Engineering, 19, 1026–1041.https://doi.org/10.1109/TKDE.2007.1048.

[11] Mukhopadhyay, A., Maulik, U., &Bandyopadhyay, S. (2013). An interactive approach to multiobjective clustering of gene expression patterns. IEEE Transactions on Biomedical Engineering, 60(1), 35–41.https://doi.org/10.1109/TBME.2012.2220765.

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Comput. Sure. 41(3), 2009.https://doi.org/10.1145/1541880.1541882.

[13] Raja, P. Vishnu, and V. MuraliBhaskaran. "An effective genetic algorithm for outlier detection." *International Journal of Computer Applications* 38, no. 6 (2012): 30-33.

[14] Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." *Artificial intelligence review* 22, no. 2 (2004): 85-126.https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

[15] Cheng, Chun-Hung, Wing-Kin Lee, and Kam-Fai Wong. "A genetic algorithm-based clustering approach for database partitioning." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 32, no. 3 (2002): 215-230.https://doi.org/10.1109/TSMCC.2002.804444.

[16] Beasley, David, David R. Bull, and Ralph Robert Martin. "An overview of genetic algorithms: Part 1, fundamentals." *University computing* 15, no. 2 (1993): 56-69.

[17] Hawkins, Douglas M. *Identification of outliers*. Vol. 11. London: Chapman and Hall, 1980.https://doi.org/10.1007/978-94-015-3994-4.

[18] Anirudha R C, Kannan R and Patil N, "Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensionaldata", In *IEEE 9th International Conference on Industrial and Information Systems (ICIIS)*, pp. 1-6, 2014.

[19] Hadi, A.S., (1992), 'Identifying multiple outliers in multivariate data', Journal of the Royal Statistical Society. Series B (Methodological), Vol. 54, No. 3(1992), pp. 761-771

[20] Anusha, M., and J. G. R. Sathiaseelan. "Evolutionary clustering algorithm using criterion-knowledge-ranking for multi-objective optimization." *Wireless Personal Communications* 94, no. 4 (2017): 2009-2030.https://doi.org/10.1007/s11277-016-3350-5.

[21] Pachghare, V. K., Parag Kulkarni, and Deven M. Nikam. "Intrusion detection system using self-organizing maps." In *Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on*, pp. 1-5. IEEE, 2009.

[22] Patole, Vivek A., V. K. Pachghare, and Parag Kulkarni. "Self-Organizing Maps to build intrusion detection systems." *Journal of Computer Applications* 1, no. 7 (2010).

[23] Gen, Mitsuo, and Runwei Cheng. *Genetic algorithms and engineering optimization*. Vol. 7. John Wiley & Sons, 2000.

[24] Ceglar, Aaron, John F. Roddick, and David MW Powers. "CURIO: A fast outlier and outlier cluster detection algorithm for large datasets." In *Proceedings of the second international workshop on Integrating artificial intelligence and data mining-Volume 84*, pp. 39-47. Australian Computer Society, Inc., 2007.

[25] Acuna, Edgar, and Caroline Rodriguez. "A meta-analysis study of outlier detection methods in classification." *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez* (2004): 1-25.