Outlier Detection Techniques Using Evolutionary and Semi Supervised Clustering Methods

Dr. J. RAJESWARI ,

Assistant Professor, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India. rajeswarikrishna82@gmail.com

ABSTRACT - In several data analysis activities a huge amount of sampled variables are recorded. One among the initial pace directed for getting a logical analysis is the identification of outlaying observations. Even thoughoutlaying observations are treated as anfault or noise, they might bring significant evidence. Identified outliers are candidates for aberrant data which might then unfavorably bring about model misspecification, biased parameter estimation and inappropriate outcomes. Outlier detection tries to discover patterns in data, which don't fit inanticipated behavior. It contains broad usage in an extensiverange of applications for instance military surveillance for opponent activities, fraud exposure for credit cards, interruption detection in cyber security, assurance or health care and fault identification in welfare critical systems. Their significance in data is because of the reality that they couldinterpret into illegal information in an extensive range of applications.

As formerly numerous research communities developed various outlier detection methods with numerous of these exactly destined for some applications and others being common naturally. In this paper, proficient methods are presented to identify the outliers exist in the data. The key objective of this research is to discover the innovative methods which could discover the outliers dependent upon the resemblance and the mutual sharing information exist among them.

Keywords: Outliers, Ascent based, Expectation, Optimization, Cluster, GSS, BAT, Dataset.

I. INTRODUCTION

Clustering and outlier identification are two important data mining activities. They are extensively utilized, for instancein bioinformatics, for identifying functionally reliant genes, in marketing, health surveillance, for customer segmentation, anomaly detection, etc. For these methods to perform well, certain type of dependency amongst the objects in a specified data space is needed, that is to say higher resemblance amongst clustered objects as well as higher deviancy among outliers and the residual data distribution. Clustering-based outlier detection techniques couldn'taccuratelyidentify the outliers in noisy data and if not the amount of clusters is well-knownbeforehand. The general tricky with the previoustechniques is the shortage of a properdescription for the outlier detection and doesn't aid for categorical data for bigger dataset.

The major problem found in the existing research methodologies so called clu stering based outlier detection is resolved in this research work by introducing the unsupervised learning methods which can estimate the goodness of the objects accurately and efficiently.

II. LITERATURE SURVEY

OutRank-b, a graph-based outlier detection algorithm is presented by [Moonesinghe, H. D. K., & Tan, P. N 2008]. In this method the graph representation of data is dependent on two methods- the object similarity as well asamount of shared neighbours among objects. Moreoverthis a Markov chain model is constructed on this graph thatallocates an outlier score to every object.

The Multi-Objective Genetic Algorithm (MOGA) [Moradi Koupaie, H et al., 2014] is utilized for searching the outliers from an object space and then the k-means clustering is exploited for developing the model for detecting the outliers. The goal of this work is to present an algorithm for detecting the outlier in stream data by means of clustering technique which are focused on finding the actual outlier in a time period. Hence, this approach tends to be simple and cost-efficient in comparison with the supervised approach.

Spatial fuzzy C-means (SFCM) algorithm [Tehrani, I. O et al., 2015] is proposed by optimizing the SFCM initial point values. In this techniquewith the purpose of enhancing the algorithm speed initially the fairly accurateprimary values are identified by computing the histogram of the real image. Moreover, it has noticeably

 DOI : 10.35291/2454-9150.2019.0453

improved the clustering effect. Hierarchical Clustering (HC) method is proposed thatevades the selection of Natural Language Processing (NLP) tools for instance pos taggers as well as parsers decrease the processing overhead [Bano, S., & Rao, K. (2015)]. Moreover recommend an organization to instantaneously make an extensive corpus interpreted in the company of disease names that could be used to train the probabilistic model. To eliminate undesirable distinct character for gene/protein detection on the datasets, an optimal rule filtering algorithm is used. To produce the amount of clusters and primary centroids for the histogram of the input image, a new algorithm based on empirical mode decomposition algorithm is used[Harikiran, J., et al 2015]. It overwhelms the lack of arbitrary initialization in classical clustering and attains higher computational speed by decreasing the amount of repetitions more accurately than other algorithms.

III. METHODOLIGIES FOR OUTLIER DETECTION

The overall research process is illustrated in fig 1, it clearly shows the research gap and the major contribution of the present research works are

- Novel approach which combines the attributes based Kullback-Leibler divergence (KLD) for attribute weighting process and perform the Ascent-based Monte Carlo expectation–Maximization (AMCEM) methods for outlier detection.
- Expectation Maximization Particle Swarm Optimization Weighted Clustering (EMPWC) outlier detection technique is proposed that needs none of the user-defined constraints for determining whether an object is an outlier
- GSS based BAT (GAABAT) approach is used to improve the performance metrics such as execution time and false alarm rate

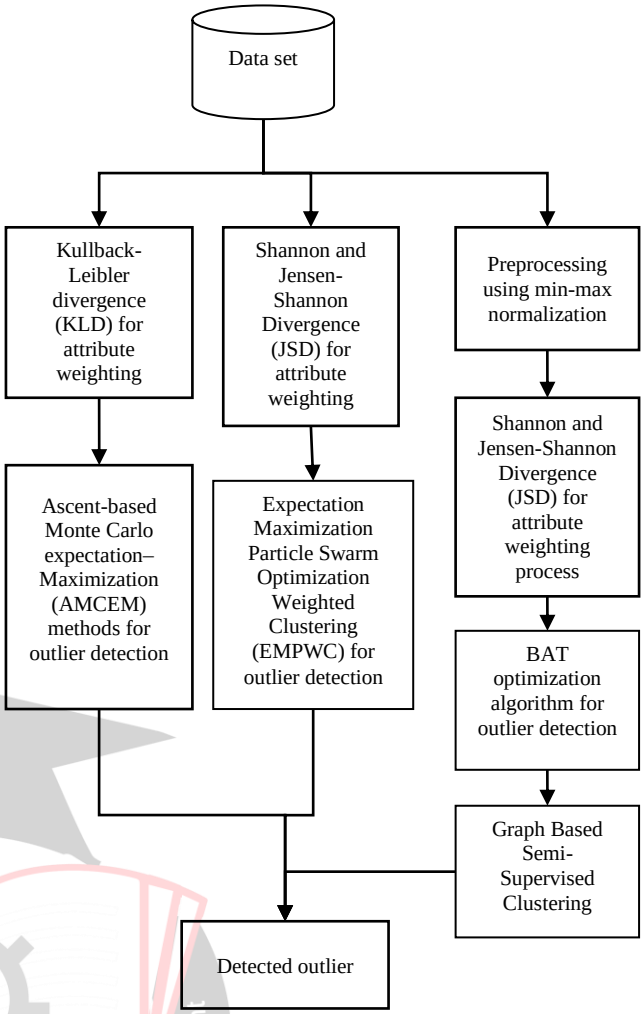

Figure 1 Research Architecture Diagram

3.1 OUTLIER DETECTION USING ASCENT-BASED MONTE CARLO EXPECTATION–MAXIMIZATION (AMCEM)

In this study, present a recognized optimization-based model of categorical outlier detection, for which a novelnotion of Kullback- Leibler divergence thatcaptures the distribution as well as holoentropy with correlation data of a dataset, is presented. With the aim ofresolving the optimization grim, derive a novel outlier factor task from the weighted holoentropy and Kullback- Leibler divergence (KLD) prove that reckoning/apprising of the outlier factor could be carried outdeprived of the necessity to guesstimate the mutualprospect distribution. Then propose an efficient Ascent-based Monte Carlo expectation–Maximization (AMCEM) clustering algorithm for outlier detection. These techniques wantmerely the amount of outliers as an input parameter and totallygive out with the limits for characterizingthe outliers obligatorythroughprevious techniques.

3.1.1 MEASUREMENT FOR OUTLIER DETECTION

The holoentropy $HL_x(Y)$ is such as the summation of the entropy and the total correlation of the random vector Y,

 DOI : 10.35291/2454-9150.2019.0453

and could be denoted by the summation of the entropies on entire characteristics

$$HL_x(Y) = H_x(Y) + C_x(Y) = \sum_{i=1}^{m} H_x(y_i) \quad (1)$$

Holoentropy gives equivalent significance to the whole attributes, while in actual applications, diverse attributes contributein a different way to create the completeorganization of the data set .The presented weighting technique calculates the weights from the data and is enthused by improved efficiency in real-world applications more willingly than by theoretical need.

1.1.2 ASCENT BASED MONTE CARLO EXPECTATION–MAXIMIZATION (AMCEM) FOR OUTLIER DETECTION

The expectation–maximization (EM) technique turn out to be a vastlyesteemed tool for get the most out of probability models in the existence of missing data, outlier's detection from every cluster. Every iteration of an EM contains an E-step and a discrete M-step. The E-step computes a conditional expectation while the M-step increases this anticipation. Over and over again, no less than one among these steps is logicallywillful; numerous researchers recommended that a hard E-step might be overwhelm by estimating the expectation with Monte Carlo approaches.
In the MCEM have also some drawbacks to detect the outlier in the data object cannot discloseindependent sampling as well as Markov chain Monte Carlo (MCMC) methods within a general structure. Secondly, they don'ttry to impersonatecertain primarilypleasing characteristics of the primary EM algorithm. These yieldsfinest outlier detection results while clustering the data object into same group based on the measurement results from MCEM algorithm .In order to overcome these problem in this work use an Ascent-based Monte carlo expectation maximization (AMCEM) for this process first need to define data object samples as (9) . Let KLYrepresent a vector of observed KLD data object results for categorical data and U represent a vector of missing attributes data and considerλbe a vector of unknown categorical data (new samples). Finally, $f_{KLY,U}(kly, u, \lambda)$dusignifiesthe probability model of the whole data to detect the outlier in the data or clustered group (KLY, U).

1.2 OUTLIER DETECTION USING EXPECTATION MAXIMIZATION PARTICLE SWARM OPTIMIZATION WEIGHTED CLUSTERING

The characterization of the Outlier detection techniques for categorical data could be done thru the means of outlier nominees are deliberated with regard to the additionalentities in the data set. This work's goal is twofold. At first, tackle with the absence of a properdescription of the outliers and the prototyping of the problem of outlier detection; secondly, target at proposing efficient and resourceful techniques which can be employed for solving the issue of outlier detection in actual applications. In this section, the means by which entropy, Shannon, Jensen-Shannon Divergence (JSD) and total correlation could be exploited for capturing the likelihood of outlier candidates is looked at. Then the notion of holoentropy is proposed and the outlier exposure problem is formulated.

3.2.1 MEASUREMENT FOR OUTLIER DETECTION

In information theory, Shannon entropy is one among the most essential metrics. Entropy does the measurement of the ambiguity which is related with a random variable.

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = \sum_{i=1}^{n} p(x_i)\, log_b \frac{1}{p(x_i)} \quad (2)$$

Jensen-Shannon Divergence (JSD) gives the mean relative entropy between two distributions and the distribution mean.

$$JS(y_i|y_j) = \frac{1}{2}\sum_i P(y_i) \ln \frac{P(y_i)}{\frac{1}{2}\left(P(y_i) + P(y_j)\right)} + \frac{1}{2}\sum_i P(y_j) \ln \frac{P(y_j)}{\frac{1}{2}\left(P(y_i) + P(y_j)\right)} \quad (3)$$

$$= \frac{1}{2} D(y_i||M) + \frac{1}{2} D(y_j||M) = S(M) - \frac{1}{2} S(y_i) - \frac{1}{2} S(y_j) \quad (4)$$

The equation (5) provides the probability computation formula of every firefly for a set of datagiven.

$$p(y_j) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n} \varphi\left(\frac{y_i - y_m}{h_n}\right) \quad (5)$$

where$\varphi(x)$stands for the window function and n refers to the total number of data objects , $V_n$and $h_n$are the respective volume and edge length of a hypercube. When the JSD is computed, then the weights are computed from the data directly and is influenced by the increase in the effectiveness

$$w_x(y_i) = 2\left(1 - \frac{1}{1 + exp\left(JS(y_i|y_j)\right)}\right) \quad (6)$$

The weighted holoentropy of random vector $W_X(Y)$ is defined to be the summation of the weighted entropy on every attribute of the random vector Y.

 DOI : 10.35291/2454-9150.2019.0453

$$W_X(y) = \sum_{i=1}^{m} w_x(y_i) H_X(y_i) \quad (7)$$

Provided a data set X with n objects and the number o, a subset Out(o) is described to be the fixed outliers in case it reduces J_X (Y; o), described as the weighted holoentropy of X with o objects eliminated

$$J_X(Y, O) = W_{X \setminus set(O)}(Y) \quad (8)$$

here set(O) refers to any subcategory of o objects from X. To be otherwise said

$$Out(O) = argmin J_X(Y, O) \quad (9)$$

Therefore, the formulation of the outlier detection is expressed to be an optimization problem. On behalf of a provided o, the amount of probable nominee sets for the detached function is $C_n^O = \frac{n!}{O!(n-O)!}$, that remains extremely huge. In addition, one mightbe required to decide the best possibleworthof O, i.e., the number of outliers that a data set actually contains. A probable hypotheticaltactic to this issue is searching for a variety of values of O and then deciding over an optimal value of O through the optimization of a particular variational property of $J_X(Y, O)$. Assume this one like a direction proposed in this research work. At present, focus will be on the development of practical solutions for the optimization issue.

3.2.2 OUTLIER DETECTION USING PARTICLE SWARM OPTIMIZATION

In PSO, every particle has analogy to a discrete "fish" that is present in a group of fish. Here, for choosing the most enhancing some variational property of $J_X(Y, O)$analysis and optimization for range of values for O. PSO contains n number of data samples Nwhichmove around a D-dimensional search space for the optimization of a particular variational property of $J_X(Y, O)$. The process of PSO is begin with a population that consists of a number of the data objects as $n(x_1, \ldots x_n)$with every$x_i$with $r_1 \ldots r_i$refer to the attribute numbers selected from 1 to m for each data sample and the optimization appropriately next searches for the best range of values for O through continuously updating the generations. The location of the i^{th} data samples of cluster particle can be referred to by $l = (l_1, , .. l_j)$. The velocity with respect to the i^{th} cluster of data points could be represented as$v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The velocities corresponding to the data points in the cluster are restricted within$[V_{min}, V_{max}]^D$, respectively. The best previous visited location of the i^{th} data points denoted its individual best outlier detection resultslbest $= (lb_{i1}, lb_{i2}, \ldots, lb_{iD})$, a value known as $lbest_i$. The finest value of the whole individual $lbest_i$ values is represented the global best position gbest $= (gb_1, gb_2, \ldots, gb_D)$ and referred to as gbest.

At every generation, the position and the velocity of eachi^{th}data points in the cluster gets revised by $lbest_i$ and gbest present in the swarm. It occurs in the space which data points discrete problem, with the intent of resolving this issue, PSO which is used for discrete binary variables. In binary space, a particle which is a data point in the cluster probably will move to the nearly corners of a hypercube through the flipping of multiple numbers of bits; as a result, the particle velocity on the whole may be defined through the amount of bits modified according to the number of processes.

1.3 OUTLIER DETECTION USING GRAPH BASED SEMI-SUPERVISED CLUSTERING WITH BAT ALGORITHM

According to the proposed system, the categorical, numerical and mixed data are evaluated by using the GSSBAT algorithm more effectively. In this section, the means by which entropy, Shannon, Jensen-Shannon Divergence (JSD) and total correlation could be exploited for capturing the likelihood of outlier candidates is looked at. In this research, the unbalanced dataset is handled and the outliers are detected optimally.

3.3.1 PREPROCESSING

As preprocessing couldprogress the outcome of a clustering algorithm, Itis a very significant step. This sectioncomputes tuples with missing values by means ofdiversechoices such as maximum, minimum, constant, average and standard deviation for identifying missing values tuples previous to employing normalization technique on the dataset.For the unbalanced dataset, the min-max normalization method is used for identifying the missing values effectively. Min-Max normalization couldexplicitlyfitting the data in a pre-defined margin. It carries out a linear transformation on the original data. Min-max normalization plots a value $d\ of\ P\ to\ d'$ in the range $[new\_min(p), new\_max(p)]$.The min-max normalization is computed by using the below formula:

$$d' = \frac{[d - \min(p)] * [new_{\max(p)} - new_{\min(p)}]}{[\max(p) - \min(p)]} \quad (10)$$

Here $\min(p) = minimum\ value\ of\ attribute$

$\max(p) = maximum\ value\ of\ attribute$

By using the above formula, the missing values are identified efficiently. It is used to list out the outliers from the given dataset. An unstable data set is described as one class of data strictlyoutstrips the other class of data. On behalf of the forecast of the data record in data set, classification method could be utilized. It learns the model from previouslytaggedancient data. It utilizesan educated model for foreseeing the class of unseen or unknown data.

This research Synthetic Minority Over-sampling Technique (SMOTE) on the complete class samples,

 DOI : 10.35291/2454-9150.2019.0453

outlier regions might not be definite, as a consequence of sparsely located outlier samples. With the purpose ofevading this class mix in training data distribution, this research utilize extreme outliereradication from the minority class by utilizing k Nearest Neighbor (kNN) notion as a data cleaning technique. The kNNs cardinality value signifiesif the point positioned in a sparse region or in a dense region. It is namedas the points, which are sparsely situated are extreme outliers. By means oferadicating the extreme outliers, it is disregarding the points, which are distant from the minority decision boundary for performing SMOTE.

3.3.2 MEASUREMENT OF OUTLIER

In information theory, Shannon entropy is one among the most essential metrics. Entropy does the measurement of the ambiguity which is related with a random variable. The measurement details of shannon entropy is given in the previous section 3.2.1.

3.3.3 BAT OPTIMIZATION ALGORITHM

Bat Algorithm is enthused by echolocation characteristic of bats. Echolocation is distinctive sonar thatbats use to identify prey and to evadehurdles. These bats produce extremely louder sound and hang onthe echo whichbounds back from the nearby objects. Consequently a bat could calculate how distant they are from an object. Additionally bats coulddiscriminate the dissimilarityamongst a hurdle and a prey eventhough in whole darkness.

BAT consist of n number of data samples N which move around a d-dimensional search space for the optimization of a particular variational property of $J_X(Y, O)$. The process of BAT starts with a population that contains a number of the data objects as $n(x_1, \ldots x_n)$with every$x_i$with $r_1 \ldots r_i$refer to the attribute numbers selected from $1\ to\ m$ for each data sample and the optimization appropriately next searches for the best range of values for O through continuously updating the generations. The location of the ith data samples of cluster bats can be referred to by $l = (l_1, , .. l_j)$. The velocity with respect to the ith cluster of data points can be represented as $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The velocities corresponding to the data points in the cluster are restricted within$[V_{min}, V_{max}]^D$respectively. The frequency rules are updated and global best solution is selected among several bats. By using the best fitness function value, the best solution is ranked.

3.3.4 GRAPH BASED SEMI-SUPERVISED CLUSTERING WITH BAT ALGORITHM (GSSBAT)

According to this research, to improve the clustering result, utilize the graph-based semi-supervised clustering technique that utilizesthe Gaussian random field to perform semi supervised learning. Wherein the mean of the ground is described in regard to choral functions. The technique containsdualkey steps: build the graph and classification.

Let $\chi = \{x_1, \ldots . . x_l, x_{l+1}, \ldots x_n\}$signifies a set of n microarray data objects. The initial$l$opinions$x_i \in X(i \leq l)$ are tagged and the residualopinions$x_u \in X(l+1 \leq u \leq n)$ are untagged. y is known as the class label set.

According to the graph-based semi-supervised learning techniques definedhip [13], the technique defines an undirected graph Won the entire data set. According to the graph W, the knobs are known as data occurrences in the chart and the boundaries are called the power of two data occurrences in the graph. At that moment the techniquebuilds a k nearest neighbors graph with the Gaussian function of Euclidean distance to heft the edges.

IV. RESULT AND DISCUSSION

This section carries out the efficiency and effectiveness tests for analyzing the performance of the novel GSSBAT technique. In order to test effectiveness, the result is compared with the available techniques such as Information-Theory- Based Step-by-Step (ITB-SS) and Information-Theory-Based Single-Pass (ITB-SP) for artificial data sets. For the test of efficiency, evaluations on are conducted over artificial data sets to indicate how execution time sees an increase with the amount of objects, attributes and the outliers. A huge amount of public actual data sets, many of them obtained from UCI [24], are utilized in these experiments, indicating an extensive range of fields in science and the humanities. The data set utilized is the public, categorical “soybean data”, having 47 objects and 35 attributes. This data has an extremely smaller class of 10 objects. As the data doesn’t contain outliers that are clearlyrecognized, it is obvious to have the objects of the smallest class considered as “outliers”.The results that are stated in Table 1 suggest a number of comments. These outcomes are proof of the significance of acquiring attribute weights; it is then compared with the available techniques EMPWC, AMCEM, ITB-SS, ITB-SP with and without weighting. Frequent Pattern Outlier

Factor (FIB), Common-neighbor-based distance (CNB)

Dataset	CNB	FIB	ITB-SP	ITB-SS	AMCEM	EMPWC	GSSBAT
Breast-c	0.99	0.90	0.991	0.993	0.996	0.997	0.998
Credit-a	0.84	0.92	0.985	0.992	0.995	0.996	0.997
Diabetes	0.86	0.88	0.75	0.912	0.945	0.945	0.957
Ecoli	0.89	0.92	0.96	0.99	0.996	0.998	0.999

Table1. AUC Results of Tested Algorithms on the Real dataset

The comparison results are shown in the following figure 2a, 2b, 2c, 2d, 2e

 DOI : 10.35291/2454-9150.2019.0453

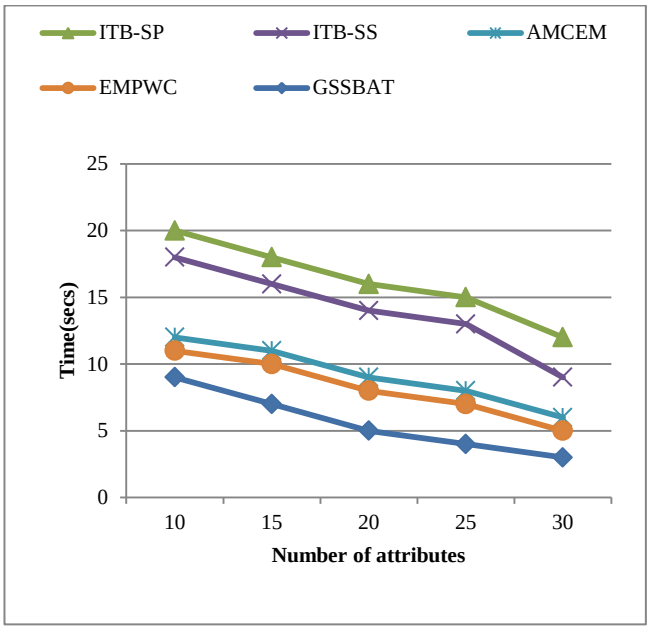

Figure 2a

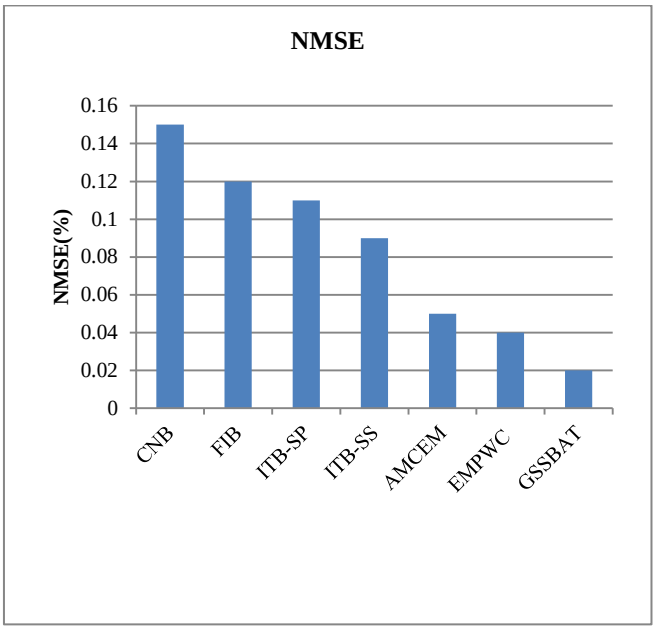

Figure 2d

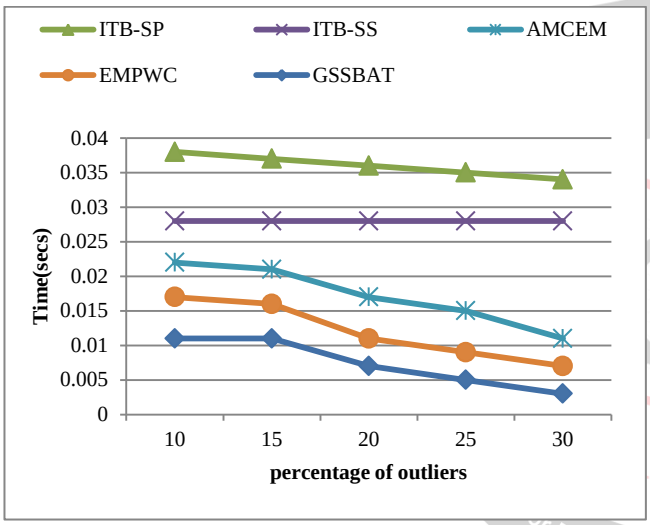

Figure 2b

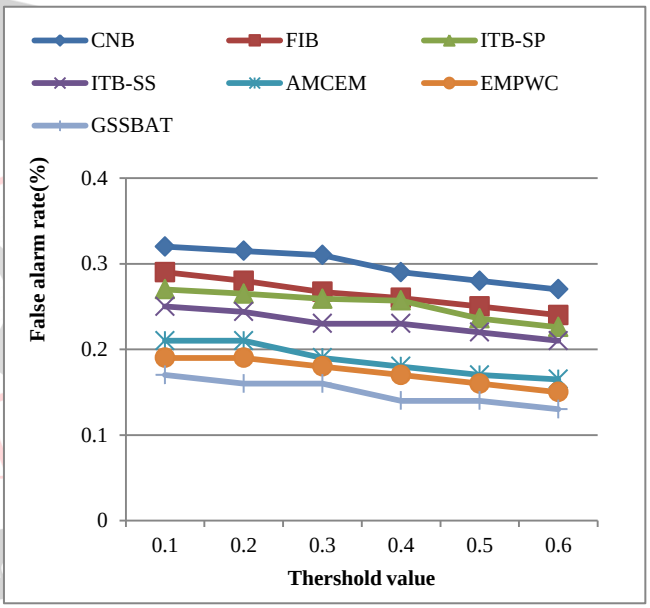

Figure 2e

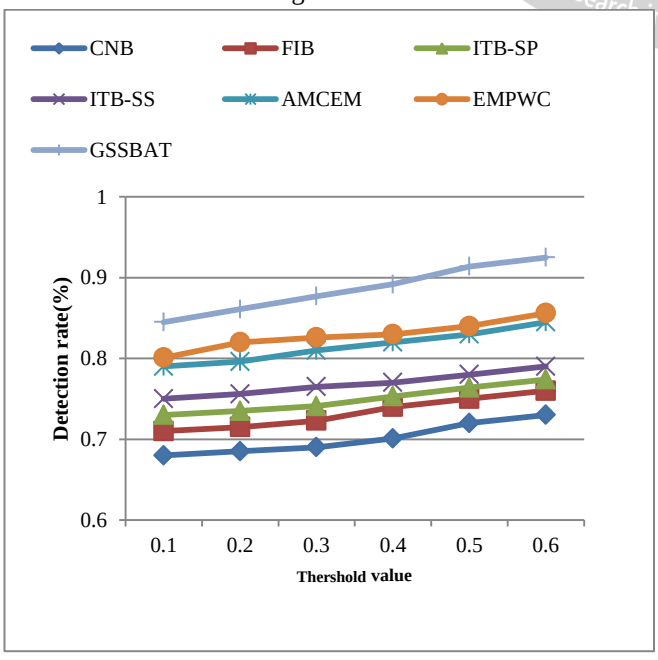

Figure 2c

Figure 2. Comparison graphs for time vs #attribute, time vs %outliers, NMSE rate, Detection rate and false alarm rate

V. CONCLUSION

In this research, the efficient outlier detection methods are proposed to improve the dataset accuracy.In the first research work, Outlier Detection techniques for categorical datasets have employed using hybrid expectation maximization methods which combines the procedure of the ascent Monte Carlo method so it is named as Ascent-Based Monte Carlo Expectation– Maximization(AMCEM) in order to identify those points containing irregular patterns. The efficiency of the second research work EMPWC outlier detection technique needs attribute frequency based outcomes from a novel conception of weighted entropy optimization which takes the data

 DOI : 10.35291/2454-9150.2019.0453

Shannon as well as Jensen-Shannon Divergence (JSD) into consideration for measuring the likelihood of outlier candidates, whereas the effectiveness of techniques proposed is a result from the outlier factor function obtained from the entropy. The efficiency of the third research work GSSBAT outlier detection technique needs attribute frequency based outcomes from a novel conception of weighted entropy optimization which takes the data Shannon as well as Jensen-Shannon Divergence (JSD) into consideration for measuring the probability of outlier candidates, whereas the effectiveness of algorithms proposed is a result from the outlier factor function obtained from the entropy. In this research, the important phases are preprocessing, outlier detection and clustering.

The result proves that the presented GSSBAT approach contains superior performance in regard to execution, NMSE time and false alarm rate than the previous approaches. In future, the kernel based outlier detection method can be developed for distributed mixed arbitrary-type data sets.

REFERENCES

[1] Tamboli, J., & Shukla, M. (2016, October). A survey of outlier detection algorithms for data streams. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 3535-3540). IEEE.

[2] Gupta, M., Gao, J., Aggarwal, C., & Han, J. (2014). Outlier detection for temporal data. Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1), 1-129.

[3] Moonesinghe, H. D. K., & Tan, P. N. (2008). Outrank: a graph-based outlier detection framework using random walk. International Journal on Artificial Intelligence Tools, 17(01), 19-36.

[4] Agrawal, A. (2009, August). Local subspace based outlier detection. In International Conference on Contemporary Computing (pp. 149-157). Springer Berlin Heidelberg.

[5] Moradi Koupaie, H., Ibrahim, S., & Hosseinkhani, J. (2014). Outlier detection in stream data by clustering method. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 25-34.

[6] Tehrani, I. O., Ibrahim, S., & Haron, H. (2015). New Method to Optimize Initial Point Values of Spatial Fuzzy c-means Algorithm. International Journal of Electrical and Computer Engineering, 5(5).

[7] Bano, S., & Rao, K. (2015). Partial context similarity of gene/proteins in leukemia using context rank based hierarchical clustering algorithm. International Journal of Electrical and Computer Engineering, 5(3), 483.

[8] Harikiran, J., Lakshmi, P. V., & Kumar, R. K. (2015). Multiple Feature Fuzzy c-means Clustering Algorithm for Segmentation of Microarray Images. International Journal of Electrical and Computer Engineering, 5(5).

[9] Xu, R and Wunsch, D., 2005. Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3):645-678.

[10] Xue, Z., Shang, Y and Feng, A., 2010. Semi-supervised outlier detection based on fuzzy rough C-means clustering. Mathematics and Computers in simulation, 80(9):1911-1921.

[11] Yang, X.S., 2010. A new metaheuristic bat-inspired algorithm. Nature inspired cooperative strategies for optimization (NICSO 2010): 65-74.

[12] Yang, X.S., 2010. Firefly algorithm, stochastic test functions and design optimisation. International Journal of Bio-Inspired Computation, 2(2):78-84.

[13] Zhao, Y and Karypis, G., 2003. Clustering in life sciences. Functional Genomics: Methods and Protocols, 183-218.

[14] Zhu, X., Ghahramani, Z and Lafferty, J.D., 2003. Semi-supervised learning using gaussian fields and harmonic functions. In 20th International conference on Proceedings of the Machine learning, 912-919.

 DOI : 10.35291/2454-9150.2019.0453