ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

AN ENHANCED OPTIMIZATION APPROACH FOR IMPROVING CLASSIFICATION ACCURACY IN DATA MINING

Chidambaranathan Krubakaran[1*], Kaliyappan Venkatachalapathy[2]

[1]*Dept. Of Computer Science Engineering, Annamalai University, Annamalai Nagar – 608002, Chidambaram, Tamilnadu, INDIA*

[2]*Dept Of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar - 608002, Chidambaram, Tamilnadu, INDIA*

ABSTRACT

This work presents an enhanced version of JAYA minimization algorithm to solve the issues in data mining classification approached with different objective functions that reflect in accuracy, reliability, cost, and efficiency improvement. The proposed enhanced JAYA algorithm was obtained by modifying the equations that is used in obtaining the best and worst solutions. Formulation of issues with respect to reliability and efficiency is highlighted in this paper to convert the multi-level objective into mono level objective function. To prove the advantages of proposed EJAYA algorithm two different test sets are used and results are obtained for different scenarios. Conventional genetic algorithm is compared with the obtained results to show the superiority of the proposed EJAYA algorithm with different complexities.

KEY WORDS
Data mining, EJAYA, Classification accuracy, Reliability

Received: 5 June 2020
Accepted: 15 July 2020
Published: 17 July 2020

***Corresponding Author**
Email:
kirubabcet2018@gmail.com

INTRODUCTION

Data mining is the process of extract information from the implicit unknown information source through classification and learning process [1]. Using computer programmes the similarities and dissimilarities and its patterns are classified and organized automatically to form a data set. This useful information helps in research to obtain better results which are applicable in many fields such as big data, medical data processing and other applications. Most of the data classification process depends on the learning process to obtain the data automatically. Using general concept learning the concept learning task is obtained in machine learning process. It categorizes the instances into positive class and negative class by train the instances and then groups the information. Using Boolean valued function these two classes are obtained [2]. The general format of concept learning deals with more than two classes of instances to obtain the information from the training instances. Based on the classified results the models are selected. Precisely based on the positive and negative instances the new unknown is compared to identified and grouped into that respective instance. This kind of learning process is given as supervised learning as the class membership of the instances are known. In unsupervised learning the training instances doesn't know the classes so the instances are grouped through data analysis [3]. Unsupervised learning is derived from the supervised learning to make use of information class and the two-step strategy is followed to obtain the class information. Fig. 1 gives an illustration about data mining process [4].

To evaluate the precision of classified data, performance evaluation through classification algorithms is generally used. In this the real data is split into two data sets such as test samples and train samples [5]. In this training samples are used to obtain the learning model and test samples are used to evaluate the accuracy of the designed application [6]. In this process the test samples are given to the model with hidden classes and then it predicts the class labels. Once it predicted the classes then it is compared with respective class labels. This evaluates the prediction accuracy of the designed application. In the comparison if two labels are same then the prediction result is considered as success as positive otherwise it is considered as error or negative [7]. Calculating the error rate is an essential factor in performance evaluation since it defines the proportion of errors obtained for the whole set [8]. This error rate on test samples are meaningful and it is used to evaluate the model similarly the error rate on training samples are useful to know since the model is derived from the same. Fig. 2 depicts the data classification process as two steps which include train data and test data sets [9].

In the unsupervised learning technique data clustering is recognized as most prominent process. In this a given dataset is categorized into likeness and disparity metric to group the useful information from the raw data set [10]. A conventional clustering algorithm requires assumptions that include the cluster structure group and adaptable objective function [11]. The natural paradigm to accommodate the data in the feature space and obtaining the exact number of partitions for a single objective function through clustering algorithm is essential. Also, it is required to estimate a combined solution which is stable and lower sensitive to noise. Similarly, multi objective clustering through multi objective optimization aims to provide several trades-off with numerous objectives [12]. It aims to cluster the data set into comparable groups to obtain the multi objective function. But it has limitations under specific conditions to apply. Conventional and metaheuristic techniques are available to optimize the required functions [13].

IIOAB JOURNAL
ISSN: 0976-3104

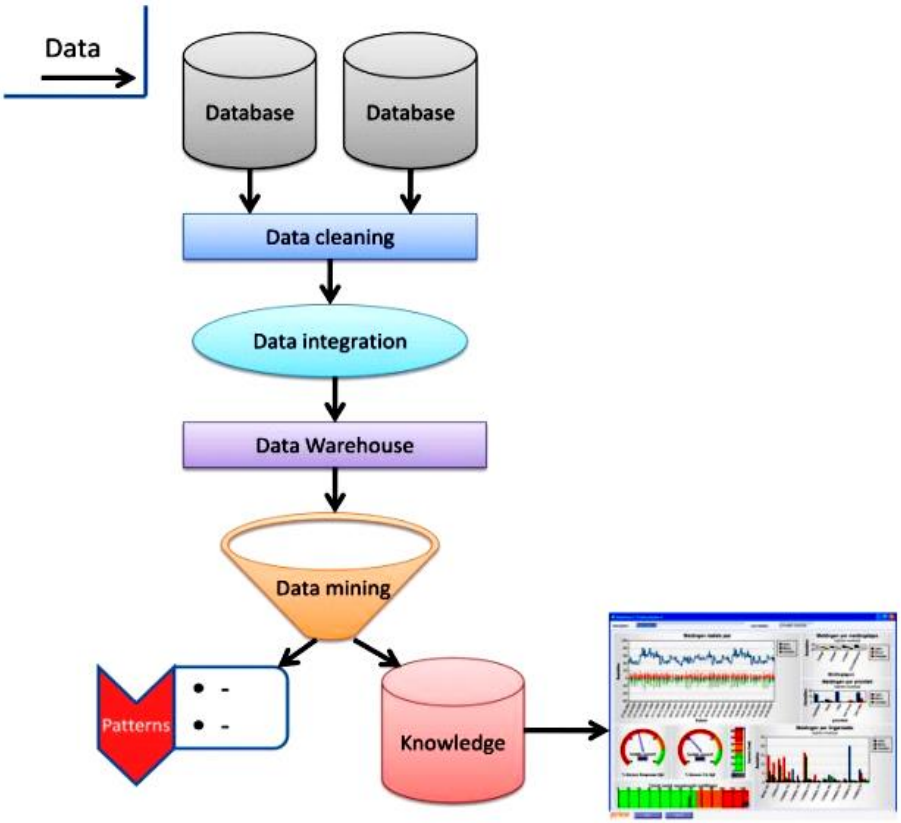

Fig. 1: Data mining process flow

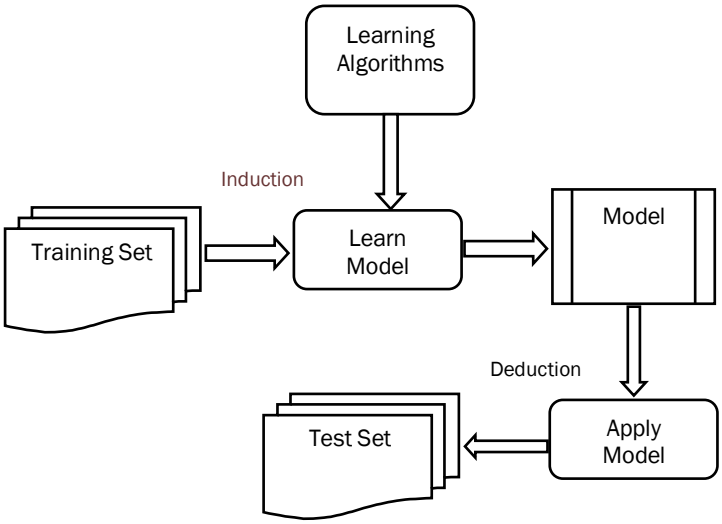

Fig. 2: Data Classification Process

Recent research interest is moved towards to optimization to improve the performance of classification results either it may be supervised or unsupervised model. Majority of the issues require this optimization strategy to improve the multiple objective and computation methodologies. Due to its simplicity and effective calculation to obtain precise results optimization involved in most of the applications now a days. Advantage of this optimization models are based on the problem concept we can modify add and remove the components to achieve better results. Also, these are robust and powerful search procedures which portray the solutions by selecting, segmenting and rearranging the set of various solutions to obtain new solution. This ideal solution increases the research interest in the field of data mining with multi objective optimization which helps in resolving many complex issues. This research article aims to provide an optimized clustering algorithm through Jaya minimization algorithm in an enhanced version by modifying the intra cluster likeness and linter cluster likeness functions. Using large set of data and k determination which is suitable for data sets and cluster validity indices, convergence solution is obtained.

IIOAB JOURNAL
ISSN: 0976-3104

METHODS

Real time data set was created based on the data array observed from four different classes of people which include 50 males and 45 females out of which 25 are elderly in the age group above 50 and 70 are in the age group of 25 – 50. Every two hours the instances for the subjects are observed for activities like sitting, walking, standing, jogging and running. The proposed method has been compared with a sequential model [2] for performance comparison.

The performance of data mining depends on the classification accuracy and clustering parameters. This proposed model provides an enhance version to improve the clustering accuracy and efficiency of the data mining approach. The mathematical model of JAYA minimization is presented in this section for better understanding of optimization in data mining model. JAYA minimization algorithm is developed to resolve optimization issues under constrained and unconstrained problems. It avoids the worst result scenario by obtaining optimal solution for the issues in data mining applications. By achieving accomplishments while obtaining optimal solution it try to attempt long way from the worst solution. This helps to move the genuine solution for the issues. It is mainly application perspective approach and the advantage over conventional optimization model is its unrestricted parameter selection. So that by selecting any two common parameters this algorithm operates. Population size and number of iterations are the two important parameters highlighted in JAYA optimization models.

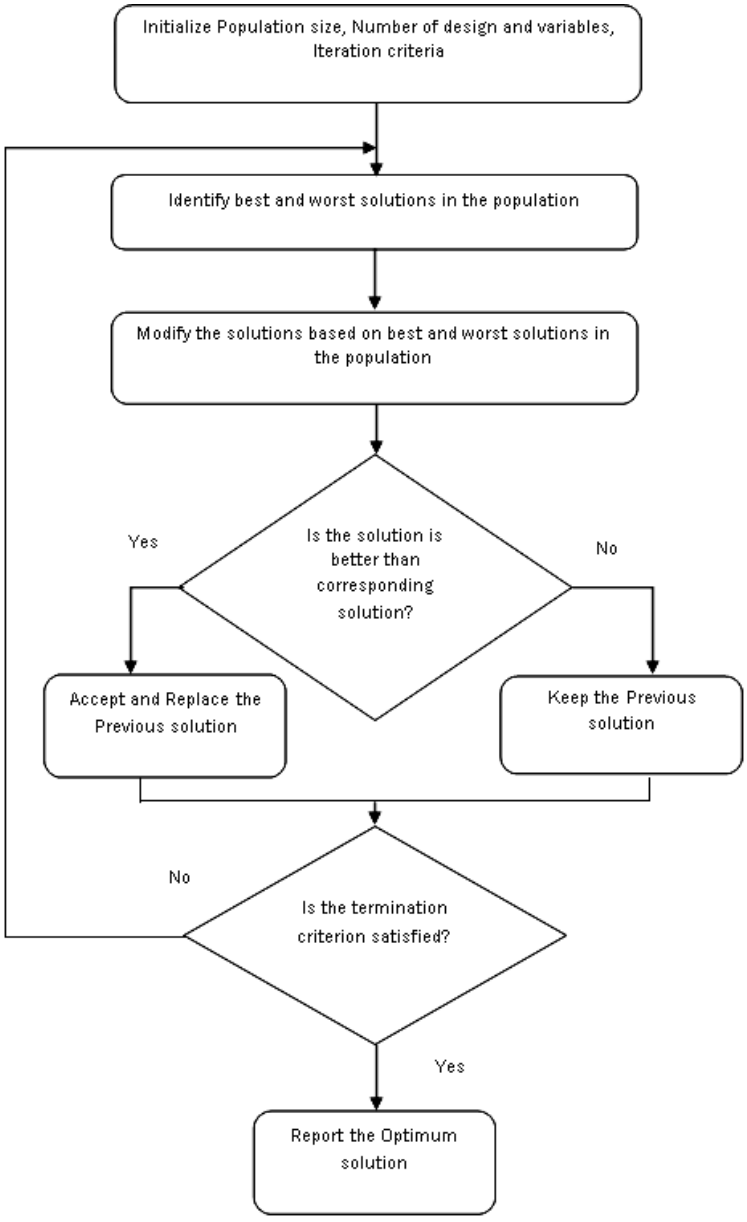

Fig. 3: JAYA minimization algorithm process flow

IIOAB JOURNAL
ISSN: 0976-3104

Since the conventional models differs by selecting different parameters for each application the advantage of JAYA implies in its calculation abilities by ignoring the changing environment constrains and reduces the optimization issues. This helps to implement JAYA minimization model in many real time applications which has complex and large heterogeneous datasets. It doesn't need any specific algorithmic parameters and even it doesn't need tuning parameter before the optimization process. Accepting the best solution and rejecting the worst solution based on the design criteria is main factor in JAYA algorithm. Fig. 3 gives the general illustration of JAYA minimization model flow process which is used in the proposed model.

Let $f(x)$ is the objective function which requires minimization or maximization. Let us assume the iteration as i and the maximum number of variables are n in numbers. The candidate solution is defined as m which is the actual population size. Let the variables be $j = 1,2,3 \dots n$ and the population size be $k = 1,2,3 \dots m$. For obtaining the $best$ solution from the entire data it is given as $f(x)_{best}$ and the worst solution is given as $f(x)_{worst}$. If the $X_{j,k,i}$ is the value of j^{th}variable for the k^{th} set while it performs the i^{th}iteration then the value of modified function is obtained as

$$X'_{j,k,i} = X_{j,k,i} + r_{1j,i}(X_{j,best,i} - |X_{j,k,i}| - r_{2j,i}(X_{j,worst,i} - |X_{j,k,i}|)) \quad (1)$$

Where $X_{j,best,i}$ is the value of $best$ solution for the j variable and $X_{j,worst,i}$ is the worst solution for the j variable. The update value is given as $X'_{j,k,i}$ which includes two random functions for the j^{th} variable when i^{th} iteration is in progress. The range of iteration is [0,1] and the random variables are given as $r_{1j,i}$ and $r_{2j,i}$.

The tendency to obtain the solution which is closer to best solution is obtained from the difference values $r_{1j,i}(X_{j,best,i} - |X_{j,k,i}|)$ and the tendency to avoid the worst solution is obtained from the other values $r_{2j,i}(X_{j,worst,i} - |X_{j,k,i}|)$. Once the best value is obtained the factor $X'_{j,k,i}$ accepts the best solution and this becomes the input for the next iteration.

The absolute value for the solution helps to enhance the ability of the algorithm and this operation considers the value of $X'_{j,k,i}$ to reflect the allowable values. If the value is exceeding the corresponding upper and lower boundaries then the desired operation is represented as

$$X'_{j,k,i} = \begin{cases} 2x'_j - x'_{j,k,i} & if\, x'_{j,k,i} < x'_j \\ 2x''_j - x'_{j,k,i} & if\, x'_{j,k,i} < x''_j \\ x'_{j,k,i} & otherwise \end{cases} \quad (2)$$

Based on the value of objective function the counter past of individual target and the vector is compared to obtain the best functional value. Otherwise the target vector is retained with the same population as

$$X'_{j,k,i} = \begin{cases} x'_j & if\ f(x'_j) \le f(x_j) \\ x_j & otherwise \end{cases} \quad (3)$$

Where, f is the cost function which used to minimize the functional parameters.

The pseudo code for the proposed EJAYA algorithm is given as follows in a summarized manner

---

Initializepopulation size,
Initialize number of designs, variables and meeting criteria,
Initialize number of fitnessfunction evaluations
Analyse the fitness function value for each candidate;
Categorize into best and worst solution
Fitness function =population;
While fitness function <Max_fitness function
do
Select the best candidate xbest and the worst candidate xworst from the population;
For i = 1 to population
do
Select the fitness function value for the updated candidate;
Fitness function = population + 1;
Accept the new solution if it is better than the old one
End for
End while.

IIOAB JOURNAL
ISSN: 0976-3104

RESULTS

The proposed model optimization behavior is experimented and compared with sequential algorithm. Experimentation is performed by implementing C language in GCC v.4.8.5 compiler. The platform is composed of two nodes and each node is comprised of x5660 processors which has processing core frequency of 2.8 Ghz with infinite communication band. The simulation parameters are given in Table 1.

Table 1: Simulation parameters

S.No	Parameter	Value
1	Population	512 and 1024
2	Iteration parameter	25000
3	Number of runs for each node	30 and 40

The speedup function for the proposed model and sequential model is given in Fig. 4 and Fig. 5 with different number of runs. From the results it is observed that proposed model achieves better speedup function in both run values compared to another model. The function values are used to calculate the speedup values and it gradually increases for each run and reaches a maximum of 10 for the last value. While the proposed model achieves more speedup and another model achieves 8.56 for their last function value.

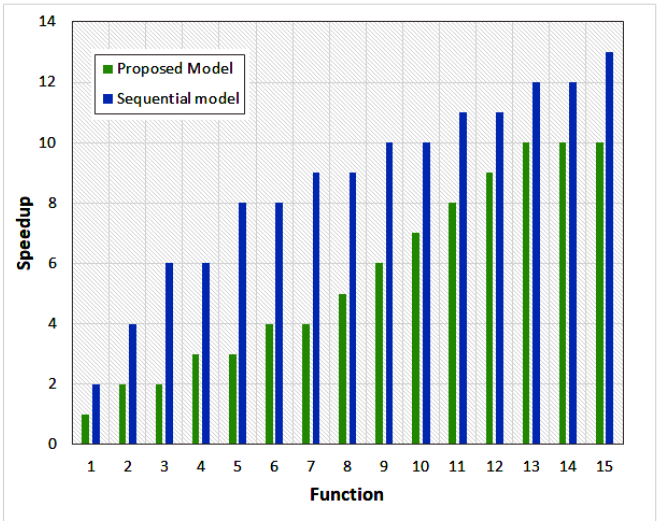

Fig. 4: Speedup comparison for 30 Runs

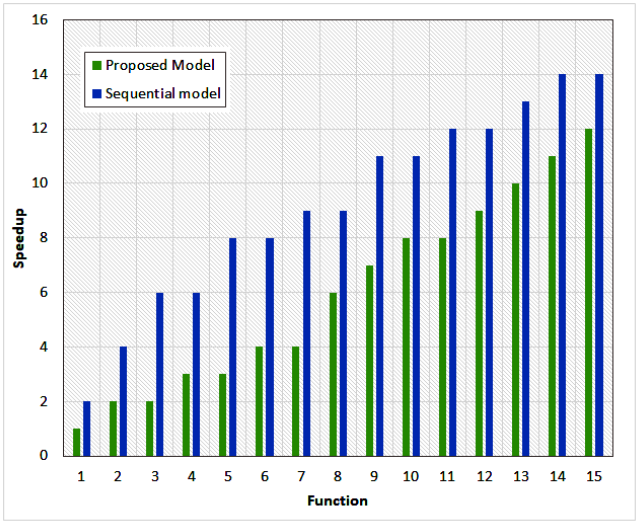

Fig. 5: Speedup comparison for 40 Runs

Fig. 6 gives a comparison of fitness function for proposed model and sequential model through the mean value and obtained value. It is observed that proposed model achieves better fitness function by providing

IIOAB JOURNAL
ISSN: 0976-3104

values near to mean value. While the sequential model deviates in fitness function with respect to mean value. Maximum of 30000 iterations are considered in the comparison process.

The efficiency comparison of each model is given in Fig. 7 and Fig. 8 for different iteration values and population size of 512. It is observed that the number of iterations doesn't affect the performance in proposed model while it is important parameter in sequential model which affects the performance. Fig. 7 gives the efficiency comparison for sequential model and Fig. 8 gives an efficiency comparison of proposed model. 30 runs are used for both population sizes.

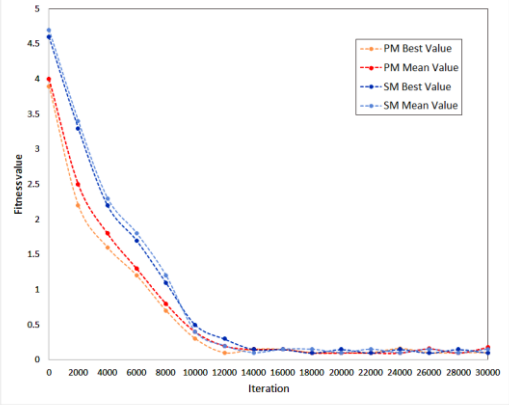

Fig. 6: Fitness function comparison for proposed model and sequential model with best and mean value for 30000 iterations

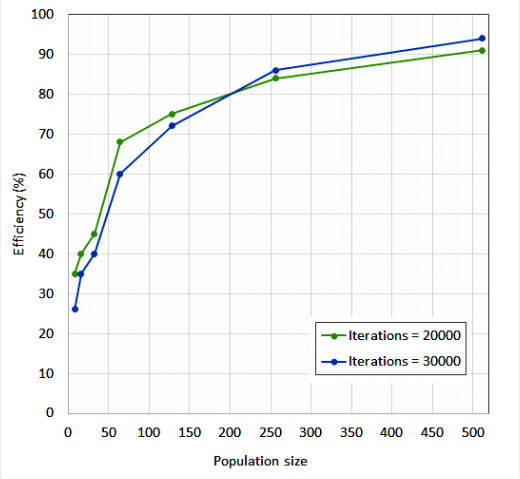

Fig. 7: Efficiency Performance of Sequential model for 30 Runs

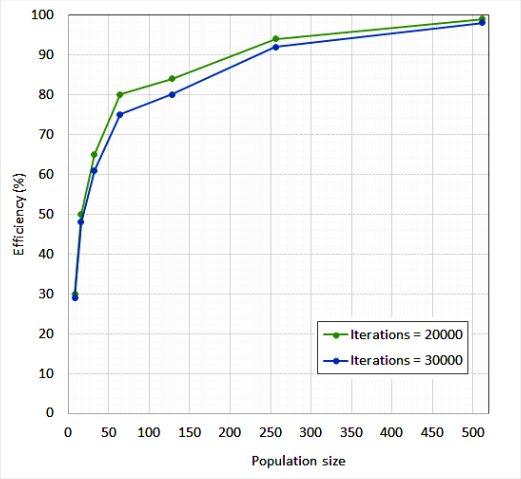

Fig. 8: Efficiency Performance of proposed model for 30 Runs

IIOAB JOURNAL
ISSN: 0976-3104

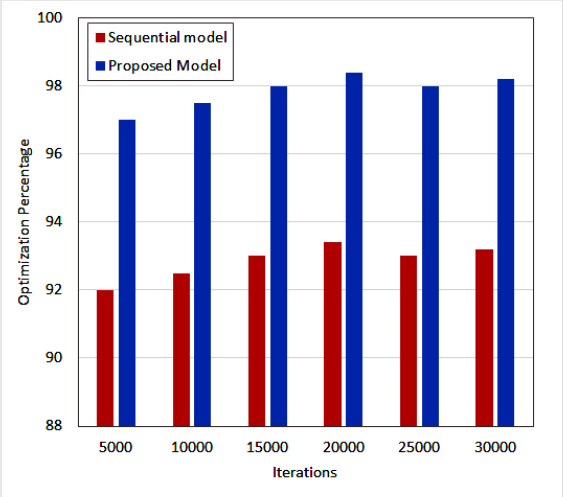

Fig. 9: Optimization Performance Comparison

Fig. 9 gives a comparison of optimization ratio for obtaining best solution with respect to proposed model and sequential model. The percentage of obtaining solution varies with 5% when compared to sequential model which makes a huge difference in data mining approach.

CONCLUSION

This research work provides an enhanced JAYA optimization model to overcome the issues in conventional data mining models and solves the large data set objectives. Using this enhanced optimization model the performance of obtaining best solution was increased. Conventional sequential algorithm is used in experimentation for comparing the proposed model results and it is validated. Through this proposed model we achieve an efficiency of 98% (as observed in Fig-7) and an optimization ratio of 98.2% (as observed in Fig- 9) which are much better than conventional models.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] dos Santos BS, Steiner MT, Fenerich AT, Lima RH. [2019] Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers & Industrial Engineering, 138:106-120.

[2] Lee G, Yun U. [2018] A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Generation Computer Systems, 68:89-110.

[3] Salehi H, Das S, Biswas S, Burgueño R. [2019] Data mining methodology employing artificial intelligence and a probabilistic approach for energy-efficient structural health monitoring with noisy and delayed signals. Expert Systems with Applications, 135:259-72.

[4] Arslan AK, Colak C, Sarihan ME. [2016] Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130:87-92.

[5] Tsai CF, Lin WC, Ke SW. [2016] Big data mining with parallel computing: A comparison of distributed and Map Reduce methodologies. Journal of Systems and Software, 122:83-92.

[6] Apiletti D, Baralis E, Cerquitelli T, Garza P, Pulvirenti F, Michiardi P. [2017] A parallel mapreduce algorithm to efficiently support item set mining on high dimensional data. Big Data Research, 10:53-69.

[7] Gürbüz F, Turna F. [2018] Rule extraction for tram faults via data mining for safe transportation. Transportation research part A: policy and practice, 116:568-79.

[8] Ryang H, Yun U. [2016] High utility pattern mining over data streams with sliding window technique. Expert Systems with Applications, 57:214-31.

[9] Lin HY, Yang SY. [2019] A cloud-based energy data mining information agent system based on big data analysis technology. Microelectronics Reliability. 97:66-78.

[10] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. [2017] Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15:104-16.

[11] Kavakiotis I, Samaras P, Triantafyllidis A, Vlahavas I. [2017] FIFS: A data mining method for informative marker selection in high dimensional population genomic data. Computers in biology and medicine, 90:146-54.

[12] Zhang J, Williams SO, Wang H. [2018] Intelligent computing system based on pattern recognition and data mining algorithms. Sustainable Computing: Informatics and Systems, 20:192-202.

[13] Sekar K, Mohanty NK. [2017] Combined mathematical morphology and data mining based high impedance fault detection. Energy Procedia, 117:417-23.