Missing Value Aware Optimal Feature Selection Method for Efficient Big Data Mining Process

S. Meera, B. Rosiline Jeetha

Abstract--- Big mining plays a more critical role in the real world environment due to presence of large volume of data with different varieties and type. Handling these data values and predicting the information would be the more difficult task which needs to be concerned more to obtain the useful knowledge. This is achieved in our previous research work by introducing the Enhanced Particle Swarm Optimization with Genetic Algorithm – Modified Artificial Neural Network (EPSOGA -MANN) which can select the optimal features from the big volume of data. However this research work might be reduced in its performance due to presence of missing values in the dataset. And also this method is more complex to perform due to increased computational overhead of ANN algorithm. This is resolved in the proposed research method by introducing the method namely Missing Value concerned Optimal Feature Selection Method (MV-OFSM). In this research method Improved KNN imputation algorithm is introduced to handle the missing values. And then Dynamic clustering method is introduced to cluster the dataset based on closeness measure. Then Anarchies Society Optimization (ASO) based feature selection approach is applied for performing feature selection in the given dataset. Finally a Hybrid ANN-GA classification technique is applied for implementing the classification. The overall performance evaluation of the research method is performed in the matlab simulation environment from which it is proved that the proposed research method leads to provide the better performance than the existing research technique.

***Keywords--- Feature Selection, Missing Value Handling, Preprocessing, Dynamic Clustering, Closeness Measure*.**

I. INTRODUCTION

Feature-selection techniques are an important part of machine learning [1]. Feature selection is often termed as variable selection, attribute selection and variable subset selection. It is the process of reducing input features to the most informative ones for use in model construction. Feature selection should be distinguished from feature extraction. Although, both techniques are used to reduce the number of features in a dataset, feature extraction is reduction technique in dimensionality that creates new combinations of attributes, whereas feature selection includes and excludes the attributes that are present in the data without changing them. Streaming feature selection has recently received attention with regard to real-time applications [2]. Feature selection with streaming data, known as streaming feature selection or online streaming feature selection is a popular technique that uses selection of features that are most informative to reduce streaming data size.

In streaming feature selection, the candidate features arrive sequentially. The size of these features is unknown. Streaming feature selection has a critical role in real time applications, for which the required action must be taken immediately [3]. In applications such as weather forecasting, transportation, stock markets, clinical research, natural disasters, call records, and vital-sign monitoring, streaming feature selection plays a key role in efficiently and effectively preparing big data for the analysis process in real time [4].

In this work feature selection process is analyzed on the big data. This is done by introducing the various methods that can lead to optimal outcome. This research method concentrates on the feature selection process with the presence of missing values in the data set [5]. The missing values in the data set would lead to incorrect decision which needs to be handled with more concern to ensure the accurate and proper feature selection and decision making [6]. In this research method Improved KNN imputation algorithm is introduced to handle the missing values. And then Dynamic clustering method is introduced to cluster the dataset based on closeness measure. Then Anarchies Society Optimization (ASO) based feature selection approach is applied for performing feature selection in the given dataset. Finally a Hybrid ANN-GA classification technique is applied for implementing the classification.

The overall organization of the research work is given as follows: In this section detailed introduction about the feature selection issues and the classification performance degradation due to processing irrelevant features are given. In section 2, varying related research works which are to select the optimal featured from the training data set to increase the classification accuracy is given. In section 3, detailed discussion about the proposed research methodology is given. In section 4, performance evaluation of the proposed research methodology in terms of various existing research method is given. Finally in section 5, overall research of the work is concluded based on the results values obtained in the experimental setup environment.

II. RELATED WORKS

In these section different methodologies proposed by various researchers in terms of performing accurate classification and optimal feature selection is discussed.

Manuscript received September 16, 2019.

S. Meera, Assistant Professor (PG), Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore. T.N, India.(e- mail: meeranarendran@gmail.com)

Dr.B. Rosiline Jeetha, HOD, Professor Dept of Computer Science, Dr.N.G.P College of Arts and Science, Coimbatore. T.N, India.(e- mail: jeethasekar@gmail.com)

Shelly Gupta el al [7] has compared various classification techniques using advanced tool such as WEKA [8], Tanagra & Clementine. All the techniques are evaluated on the basis of accuracy percentage and their respective error rate with 10 fold cross validation. The experiment clearly ranks all the available classification techniques which can be implemented to build an expert system for the patients of diabetes.

H.A. Guvenir et al [9] not only propose an expert system but implement it into a visual tool which can be used for the purpose of diagnosis of complex skin diseases. It uses three classification techniques named nearest neighbor classification, naive bayesian classifier with normal distribution method and the advanced classification technique, voting feature intervals-5.

The nearest neighbor classification determines the closeness of unknown sample to a group of similar patterns. The closeness is determined by calculating the Euclidean distance between them (Jiawei Han, MichelineKamber [10]).

The conditional probability of each feature is calculated in naive bayesian classification technique and in VFI-5, score of each feature is being calculated by giving assigned weight age to them, on the basis of voting, class is determined for each feature. Gilad-Bachrach et al. (2004) [11], introduced a new approach called SIMBA (Iterative Search Margin Based Algorithm), which outperforms RELIF.

This approach introduces the idea of measuring the quality of a set of features by the margin it induces. To overcome the drawback of iterative search, GiladBachrach et al, present A Greedy Feature Flip Algorithm called G-Flip. The G-Flip is a greedy search algorithm for maximizing the margin function

The NSGA-II [12] and SPEA [13] are currently planned for the feature selection in multi objective based methods. To get small set of non-redundant disease related genes by using the multi objective particles swarm optimization from the hybrid multi-objective optimization method [14].

Salcedo-Sanz et al uses multi objective genetic algorithms used to common uniqueness of the samples as feature correlation and search the subset of features by combining different filter approaches criteria for feature selection [15]. The multi-objective approach in hybrid GA and support vector machine classifier (GASVM) is proposed for gene selection and the classification of gene expression data [16]. An archived multi-objective simulated annealing (AMOSA) is developed for predicting miRNA promoters using an SVM with RBF kernel in a feature selection method [17].

A neural network is working to let the use of a representative database to calculate suitability. A multi-objective evolutionary algorithm is projected to solve the difficulty of gene selection in the gene subset size minimization and performance maximization [18].

To establish the benchmark problems by using the optimization of split modified radius-margin model selection criteria [19].The a multi-objective genetic algorithm is used to gene selection in the microarray datasets [20]. This process is consummate by support vector machines. A multi objective genetic algorithm used to increase the classification accuracy rate in the number of features [21].

Simon et al [22] also proved the effectiveness of multi objective genetic algorithm in various fields. Author concludes that genetic algorithm leads to better outcome in the process of edge detection process.

Schwefel et al [23] attempted to utilize the biological behaviour based methods in the various fields and applications. This technique is mainly used to elucidate the reasons after the use of multi objective optimization in each application area and also to direct the possible futures.

Punam et al. [24] introduced the implementation of canopy clustering algorithm based on Twister and Hadoop. This is achieved by applying four major MapReduce modules. In the first module, a list of canopy centers are generated, which are assigned to the points in the next module. However, authors argue in favor of Twister, which offers long running of tasks.

Li and Xi [25] are the first to focus on running DBSCAN in MapReduce. This approach splits data points into clusters within each partition, which are then merged using MapReduce.

The authors stated that their approach could work efficiently in massive data sets. However, they do not provide detailed algorithms to partition the data and merge the clusters. He et al. [26] proposed a parallel density-based clustering algorithm using MapReduce.

The main contribution of this work is to reduce spatial complexity. By this way, this approach is programmed in four-stage MapReduce paradigm. First, the authors proposed a partitioning strategy using grid file, which are executed by only one MR round.

III. MISSING VALUES AWARE BIG DATA MINING

In this research method Improved KNN imputation algorithm is introduced to handle the missing values. And then Dynamic clustering method is introduced to cluster the dataset based on closeness measure. Then Anarchies Society Optimization (ASO) based feature selection approach is applied for performing feature selection in the given dataset. Finally a Hybrid ANN-GA classification technique is applied for implementing the classification.

3.1. Dataset Collection

HIGGS dataset has been acquired from UCI archive1 for simulating the negative impacts of the Big Data in accordance with high dimensions (several features) and massive volume (several instances) for the purpose of experimentation.

Its generation is done using particle detectors in accelerator and is physics dataset. The data is being generated by means of Monte Carlo simulations. Almost 53% of positive samples are used for balancing the dataset. The dataset consists of 11000000 samples. The HIGGS dataset features are described in Table 1.

Table 1: Details of HIGGS dataset

Feature name	Feature details
lepton pT, lepton eta, lepton phi, missing energy magnitude, missing energy phi, jet 1 pt, jet 1 eta, jet 1 phi, jet 1 b-tag, jet 2 pt, jet 2 eta, jet 2 phi, jet 2 b-tag, jet 3 pt, jet 3 eta, jet 3 phi, jet 3 b-tag, jet 4 pt, jet 4 eta, jet 4 phi, jet 4 b-tag	21 low-level features. These form kinematic attributes
m_jj, m_jjj, m_lv, m_jlv, m_bb, m_wbb, m_wwbb	7 high-level features employed to distinguish between the two classes

3.2. Missing Data Handling Using Improved KNN Imputation Algorithm

Missing data present in the data set would lead to inaccurate decision making which needs to be handled with more concern. Missing data value replacement needs to be done accurately by considering the remaining attribute values present in the datasets. The accurate and efficient missing value replacement will lead to enhanced outcome which is done in this work by using kNN algorithm. While using kNN algorithm, after k nearest neighbors are found, several strategies could be taken to predict the category of a test document based on them. But a fixed k value is usually used for all classes in these methods, regardless of their different distributions. Equation (1) and (2) below are two of the widely used strategies of this kind method.

$$y\,(d_i) = \operatorname{argmax}_k \sum_{x_j \in KNN} y\,(x_j, c_k)$$

$$y\,(d_i) = \operatorname{argmax}_k \sum_{x_j \in KNN} \mathrm{Sim}\,(d_i, x_j) y\,(x_j, c_k)$$

Where d_i is a test document, x_j is one of the neighbors in the training set, $y\,(x_j, c_k) \in \{0,1\}$ indicates whether x_j belongs to class c_k, and $\mathrm{Sim}(d_i, x_j)$ is the similarity function for d_i and x_j. Equation (1) means that the predication will be the class that has the largest number of members in the k nearest neighbors; whereas equation (2) means the class with maximal sum of similarity will be the winner. The latter is thought to be better than the former and used more widely. In general, the document distribution of different classes in the training set is uneven. Some classes may have more samples than others. Therefore, it is very likely that a fixed k value will result in a bias on large classes. For example, when using the strategy indicated by equation (2), many tiny similarity values would accumulate to a large one, which may improperly make a large class the final decision. To overcome this problem, we propose a different strategy as follows

When we get the original k nearest neighbors, we compute the probability that one document belongs to a class by using only some top n nearest neighbors for that class, where n is derived from k according to the size of a class cm in the training set. In other words, we use different numbers of nearest neighbors for different classes in our method. For larger classes, we use more nearest neighbors. The dynamic selection is based on the class distribution in the training set. To make the comparison between classes reasonable, we derive the probabilities from the proportion of the similarity sum of neighbors belonging to a class to the total sum of similarities of all selected neighbors for that class. Equation (3) gives the decision function in our improved kNN algorithm.

$$y(d_i) = \operatorname{argmax}_m \frac{\sum_{x_j \in \text{top\_n\_KNN}(c_m)} \mathrm{SIm}(d_i, x_j) y(x_j, c_m)}{\sum_{x_j \in \text{top\_n\_KNN}(c_m)} \mathrm{Sim}(d_i, x_j)}$$

where,

$\mathrm{Top\_n\_KNN}(c_m)$ = {top n neighbours in the original k nearest neighbours KNN | n = $\left\lceil \frac{k \times N(c_m)}{\max\{N(c_j) | j=1..Nc\}} \right\rceil$}

Note that N(cm) denotes the size of the class cm in the training set, and max{N(cj)|j=1..Nc} is the size of the largest class in the same set.

3.3. Dynamic Clustering

After missing value replacement, data clustering is done to divide the large volume of data into multi sub clusters. Thus the feature selection process can be done easily without burden of handling large volume of data at once. The processing overhead of handling large volume of data can be reduce considerably and also can lead to accurate feature selection outcome. In this work dynamic clustering is performed to handle the live streaming data, thus the data integrity can be ensured. The neighborhood reach ability cost and the neighborhood reach ability span are two parts that jointly quantify the difficulty to establish a neighborhood chain between two data points, and the difficulty in establishing the chain can be used to measure the data points' closeness. The CMNC between any two data points A and C in a dataset is defined as:

$$\mathrm{CMNC}\,(A, C) = \frac{1}{\mathrm{NRC}(A, C).\,\mathrm{NRS}(A, C)}$$

A bigger CMNC value means that the chain between the two data points can be more easily established which represents that the two points are closer, while a smaller CMNC represents the opposite. Strictly speaking, CMNC is not a distance metric since it violates the triangle inequality due to the use of the neighborhood relationship. However, using CMNC as a kind of closeness (similarity) measure, we can obtain more intuitive and rational closeness quantifications compared with using traditional closeness metrics based on the geometric distance alone in clustering tasks.

The K-means algorithm finds the predefined number of clusters. In the practical scenario, it is very much essential to find the number of clusters for unknown dataset on the runtime. The fixing of number of clusters may lead to poor quality clustering. The proposed method finds the number of clusters on the run based on the cluster quality output. This method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or by input the minimum number of clusters required. In the former case it works same as K-means algorithm.

In the latter case the algorithm computes the new clusters by incrementing the cluster counter by one in each iteration

until it satisfies the validity of cluster quality threshold. The modified algorithm is as follows:

Input: k: number of clusters (for dynamic clustering initialize k=2) Fixed number of clusters = yes or no (Boolean). D: a data set containing n objects.

Output: A set of k clusters.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat.
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. until no change.
6. If fixed_no_of_clusters =yes goto 12.
7. Compute inter-cluster distance using Eq.2
8. Compute intra-cluster distance using Eq. 3.
9. If new intra-cluster distance <old_intra_cluster distance and new_inter- cluster >old_inter_cluster distance goto 10 else goto 11.
10. k= k + 1 goto step 1.
11. STOP

3.4. Optimal Feature Selection Using Anarchies Society Optimization (ASO)

Feature selection is performed on the clustered data with the concern of handling large volume of data. Here each and every cluster would have data of same kind which would increase the feature selecting accuracy. In this work ASO algorithm is introduced to perform the feature selection task. The algorithm proposed in this section is based on ASO that is a human-inspired optimization method introduced by Ahmadi-Javid (2011). ASO is inspired by a human society whose members behave anarchically and adventurously to find much better situations. The members become more nervous and greedier as the differences among people intensify. Using such members, ASO explores the solution space perfectly and avoids falling into local optimums. The mathematical description of ASO is given in the following. Given the problem of minimizing a function f (x) over theset , $\Omega \subset R^d$ an ASO algorithm tries to solve the problem by using a society of members exploring the solution space Ω to seek the global optimal solution. In the kth iteration of the algorithm, each member of society i (i = 1, 2,…, N) has three characteristics as three dimensional vectors:

- Sit_i (k): The situation of the ith member,
- Dir_i (k): The movement direction associated with the selected movement of the ith member,
- $Best_i$ (k): The best personal previously experienced situation,
- and the society has a characteristic expressed as a d-dimensional vector:
- GBest(k) : The situation of the best member in the society.

Steps of the proposed ASO algorithm are presented below in a nutshell:

Step 1: Initialize the members' situations and movement directions, $Sit_i(1)$, Dir_i (1), randomly, and compute the objective functions for all society members.

Step 2: For member i in iteration k:

- Define the fickleness index FI^k (i), and then determine a movement policy MP^k (i) Current based on vectors Sit_i (k), Dir_i (k −1), and the fickleness index,
- Define the external irregularity index EI^k (i), and then determine a movement policy $MP^k_{Society}$ (i) based on the other society members' situations Sit_j (k), j ≠ i and the external irregularity index,
- Define the internal irregularity index II^k (i), and then determine a movement policy MP^k (i) Past based on the member's past situations $Best_i$ (1),...,$Best_i$ (k −1)i and the internal irregularity index,
- Determine a movement policy based on the three movement policies $MP^k_{current}$ (i) ,$MP^k_{society}$ (i) and MP^k_{past} (i).
- Update the current situation of member i by the determined movement policy.

Step 3: If the termination condition is met, then stop; otherwise, repeat steps (2)–(3).

3.5. Classification Using Hybrid Ann-GA Classification Technique

In this work classification of big data is performed by introducing the hybrid ANN-GA algorithm. This classification is carried out on the selected features. ANN has several disadvantages such as long training time, unwanted convergence to local instead of global optimal solution, and large number of parameters; therefore, there have been attempts to remedy some of these disadvantages by combining ANN with another algorithm that can take care of a specific problem.

An algorithm that has frequently been hybridized with ANN is GA. In 1990, Whitley et al. began to use GA to optimize weighted connections and find a good architecture for neural network connections. In 2006, Kim proposed a hybrid model of ANN with GA that performs instance selection to reduce dimensionality of data.

In 2012, Karimi and Yousefi used GA to find a set of weights for connections to each node in an ANN model and determine correlation of density in nanofluids. Sangwan et al. proposed an integrated ANN and GA for predictive modelling and optimization of turning parameters to minimize surface roughness. Some other successful examples of ANN-GA hybrid applications are network intrusion detection and cancer patient classification.

Inspired by these successes, this study attempted to use GA to solve a feature selection problem—to find effective subsets of input into ANN.The rationale behind our idea of using a hybrid intelligence of ANN and GA was that it should be better to use, first, multiple input variables for each technical indicator based on different past time spans and, second, a small number of effective subsets of input variables that would be imported.

Since the number of subsets of 44 variables is astronomical 244, it would take too much computation time to process them.

GA took care of that. GA is an algorithm that is especially powerful at featureselection, so we used it to find better subsets of input variables. The 10 steps of operation of ANN and GA hybrid intelligence are as follows.

Step 1 (initialization of population). Generate an initial population of chromosomes which are bit strings of randomly generated binary values. The chromosome and population sizes that we used were 44 and 10, respectively.

Step 2 (decoding). Decode chromosomes (bit strings) to find which input variables will be selected.

Step 3 (ANN). Run three-layered feed forward ANN model to make prediction of next-day SET50 index. The parameters in the model that we used were the same as those reported by Inthachot et al.

Step 4 (fitness evaluation). Take the prediction accuracy of each chromosome from ANN as its fitness value for GA.

Step 5 (stopping criterion). Determine whether to continue or exit the loop. The stopping criterion was not more than 10 generations.

Step 6 (selection). Select chromosomes to cross over using tournament selection technique. A tournament selection involves running several tournaments on a few chromosomes chosen at random from the population. The winner of each tournament is selected for crossover.

Step 7 (crossover). Apply an arithmetic crossover operator that defines a linear combination of two chromosomes.

Step 8 (mutation). Inject new genes into the population with uniform mutation operator and generate a random slot number of the crossed-over chromosome as well as flip the binary value in that slot.

Step 9 (replacement). Replace old chromosomes with two best offspring chromosomes for the next generation.

Step 10 (loop). Go to Step 2.

Fitness Evaluation. We used accuracy to determine chromosome selection (subsets of input variables)—chromosomes that would generate the next generation—as well as to measure the performance of the prediction model. Fitness values in GA were taken as the accuracy values that can be calculated as below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP TN} + \text{Fp} + \text{FN}}$$

Where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

IV. RESULTS AND DISCUSSION

The Matlab simulation environment is utilized for implementing the newly introduced research approach. The performance metrics that are taken into consideration are accuracy, recall, precision, f-measure, and time complexity. The comparison is done between the newly introduced MV-OFSM, EPSOGA –MANN scheme and the available techniques such as AABC-ITSVM, APSO-SVM and APSO-NAIVES.

Accuracy

It is defined to be the degree of correct detection. That is less false positive rate. The accuracy is calculated as follows:

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)}$$

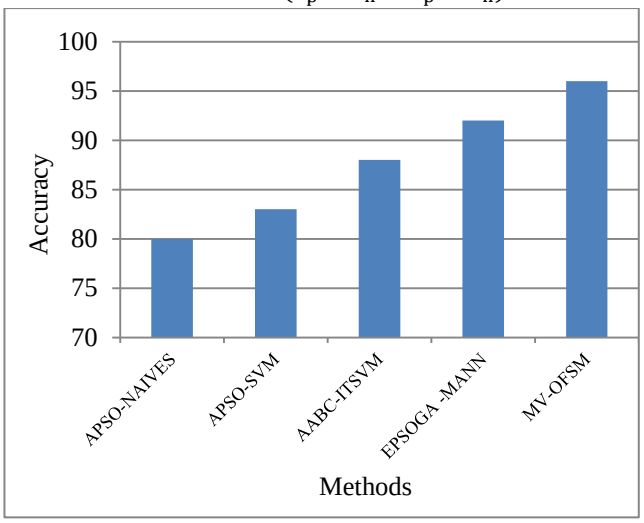

Figure 1: Accuracy

As illustrated in the Fig 1 above, it can be noticed that the comparison metric is assessed employing the available and newly introduced techniques in terms of accuracy. The newly introduced MV-OFSM chooses the best features.

These features are used in modified ANN training and testing stage to generate results with more relevance. The result shows that the newly introduced system achieves superior classification results with MV-OFSM algorithm.

Precision

Precision is described as the ratio of the true positives to both true positives and false positives results for imposition and actualfeatures. It is expressed as below

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

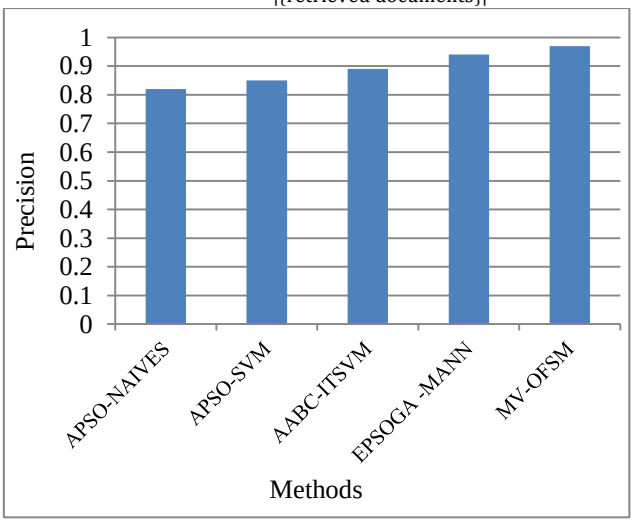

Figure 2: Precision

As illustrated in the above Fig 2, it can be seen that the comparison metric evaluation is done on the already available and newly introduced technique in terms of precision.

The result shows that the newly introduced system achieves superior classification results with MV-OFSM algorithm.

Recall

Recall value is calculated on the root of the data retrieval at true positive forecast, false negative. Generally, it can be computed as

$$Recall = \frac{T_P}{T_{P+F_N}}$$

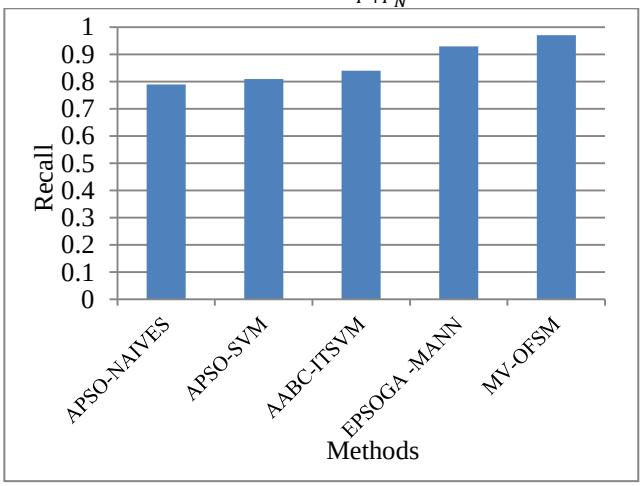

Figure 3: Recall

As seen from the Fig 3 above, it can be noticed that the assessment of their call metric comparison on the already available and newly introduced technique is performed. The result shows that the newly introduced system achieves greater classification results with MV-OFSM algorithm.

F-Measure

It provides the measure of the accuracy of a test. It takes both the precision *p* and the recall *r* of the test into consideration for the score computation.

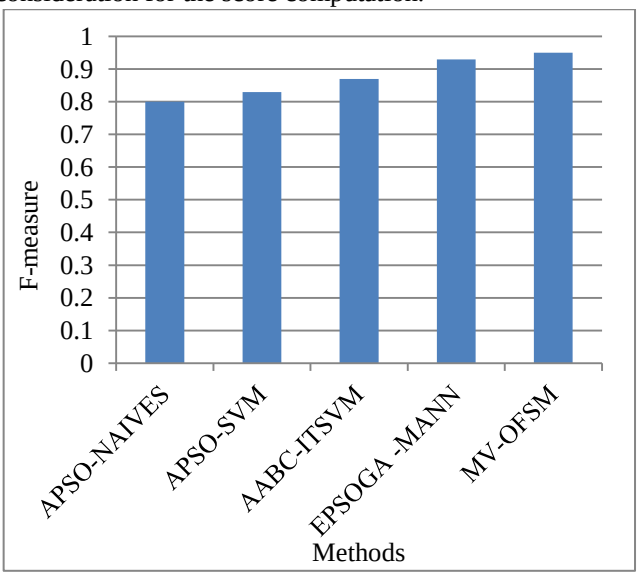

Figure 4: F-measure

As shown in the Fig 4 above, it can be noticed that the comparison metric is assessed on the available and newly introduced technique in terms of f-measure. The result shows that the newly introduced system achieves a greater f-measure result with MV-OFSMalgorithm.

Time Complexity

The system is called better when the algorithm gives minimum time complexity

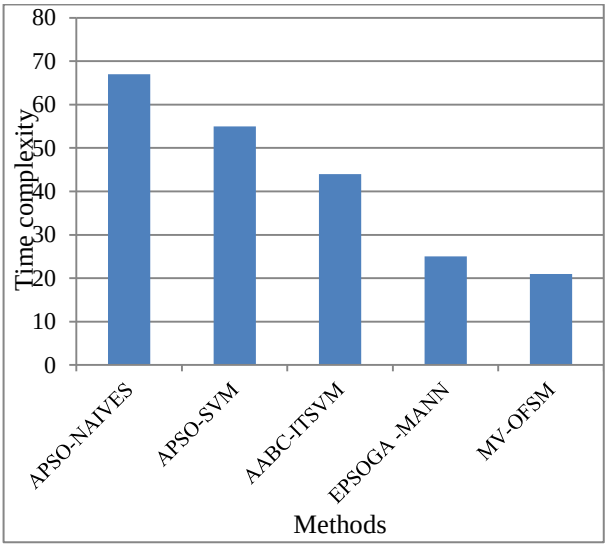

Figure 5: Time complexity

As illustrated in the Fig.5 above, the graph shows the features chosen are assessed on the particular dataset. This way, the result shows that the newly introduced MV-OFSM technique offers superior performance compared with the already available algorithms.

V. CONCLUSION

In this research method Improved KNN imputation algorithm is introduced to handle the missing values. And then Dynamic clustering method is introduced to cluster the dataset based on closeness measure. Then Anarchies Society Optimization (ASO) based feature selection approach is applied for performing feature selection in the given dataset. Finally a Hybrid ANN-GA classification technique is applied for implementing the classification. The overall performance evaluation of the research method is performed in the matlab simulation environment from which it is proved that the proposed research method leads to provide the better performance than the existing research technique. The simulation analysis proved that the proposed MV-OFSM tends to have lesser time complexity where 16% lesser time complexity than the existing research methodologies.

REFERENCES

1. John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think.
2. Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. ICWSM, 14, 505-514.
3. Lynch, C. (2008). Big data: How do your data grow?. Nature, 455(7209), 28-29.
4. Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big data storage. In Big Data (pp. 33-49). *Springer International Publishing*
5. John, G.H., Kohavi, R. and Pfleger, K., Irrelevant features and the subset selection problem. In: *Proceedings of the Eleventh International Conference on Machine Learning*, 121–129, 1994.
6. Koller, D. and Sahami, M., Toward optimal feature selection. In: *Proceedings of International Conference on Machine Learning*, 1996

8. Shelly Gupta, Dharminder Kumar and Anand Sharma, "Performance analysis of various data mining classification techniques on health care data", *International Journal of Computer Science & Information Technology (IJCSIT)*Vol 3, No 4, August 2011.
9. WEKA (2013).An open source free data mining tool for teaching and research, web source:http://www.cs.waikato.ac.nz/ml/wekalast accessed on February, 2014.
10. H.A. Guvenir, N. Emeksiz, "An expert system for the differential diagnosis of erythematosquamous diseases", Expert Systems with Applications 18 (2000) 43–49.
11. Jiawei Han &MichelineKamber, "Data Mining Concepts and Techniques", Second edition, *Morgan Kaufmann Publications*, San Francisco, CA. USA.
12. Gilad-Bachrach, R., Navot, A., &Tishby, N. (2004, July). Margin based feature selection-theory and algorithms. *In Proceedings of the twenty-first international conference on Machine learning* (p. 43). ACM.
13. Xu L, Liang Q. [2012] Zero correlation zone sequence pair sets for MIMO radar. Aerospace and Electronic Systems, *IEEE Transaction*, 48(3):2100–2113.
14. Del Ser J, Gil-Lopez S, Perez-Bellido A, Salcedo-Sanz S, Portilla-Figueras JA. [2011] IEEE 73rd Vehicular Technology Conference (VTC Spring). On the application of a novel hybrid harmony search algorithm to the radar polyphase code design problem (*IEEE Computer Society Budapest, Hungary*, pp. 1–5.
15. Gil-Lopez S, Ser JD, Salcedo-Sanz S, Perez-Bellido AM, Cabero JM and Portilla-Figueras JA. [2012] "A hybrid harmony search algorithm for the spread spectrum radar polyphase codes design problem," *Expert System Application*, 39(12):11089–11093
16. Perez-Bellido AM, Salcedo-Sanz S, Ortiz-Garcia EG, Portilla-Figueras JA, Lopez-Ferreras F. [2008] A comparison of memetic algorithms for the spread spectrum radar polyphase codes design problem, *Engineering Applications of Artificial Intelligence*, 21(8):1233–1238.
17. Karaboga D and Basturk B, [2008] On the performance of artificial bee colony (ABC) algorithm. *Applied soft computing*, 8(1):687–697.
18. Diwold K, Aderhold A, Scheidler A and Middendorf M. [2011] Performance evaluation of artificial bee colony optimization and new selection schemes. *Memetic Computing*, 3(3):149–162
19. Karaboga D and Gorkemli B. [2014] A quick artificial bee colony (QABC) algorithm and its performance on optimization problems. *Applied Soft Computing*, 23:227–238. \
20. Zhang X, Zhang X, Yuen SY, Ho SL, Fu WN. [2013] An improved artificial bee colony algorithm for optimal design of electromagnetic devices. *IEEE Transactions on Magnetics*, 49(8):4811–4816.
21. Zhang X, Wu Z. [2015] Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science, 9140, ed. by Y Tan, Y Shi, F Buarque, A Gelbukh, S Das, and A Engelbrecht. An artificial bee colony algorithm with history-driven scout bees phase, pp. 239–246.
22. Karaboga D, Gorkemli B, Ozturk C and Karaboga N. [2014] A comprehensive survey: artificial bee colony (ABC) algorithm and applications. *Artificial Intelligence Review* 42(1):21–57.
23. Simon D. [2008] Biogeography-based optimization. *IEEE transactions on evolutionary computation*, 12(6):702–713
24. Schwefel HP. [1981] Numerical Optimization of Computer Models (Wiley, Chichester).
25. Ghulia P, Shuklaa A, Kirana R, Jasona S, Shettara R. Multidimensional canopy clustering on iterative MapReduce framework using Elefig tool. *IETE Journal of Research* 2015; 61(1):14–21.
26. Li L, Xi Y. Research on clustering algorithm and its parallelization strategy, in 2011 *International Conference on Computational and Information Sciences*, 2011; 325–328.
27. He Y, Tan H, Luo W, Mao H, Ma D, Feng S, Fan J. MR-DBSCAN: an efficient parallel density-based clustering algorithm using MapReduce, in 2011 IEEE 17th *International Conference on Parallel and Distributed Systems*, 2011; 473–480