International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 8 (2015) pp. 20313-20325

http://www.ripublication.com

Towards Better Classification Using Improved Genetic Algorithm and Decision Tree For Dengue Datasets

B. Renuka Devi[1], Dr. K. Nageswara Rao[2], Dr. S. Pallam Setty[3]

[1]Research Scholar, JNT University, Kakinada &
Associate Professor, Dept. of CSE, VLITS, Vadlamudi, Guntur.
Mail id: brenukadevi.2009@gmail.com
[2]Professor, Dept. Of CSE, PSCMR College of Engineering, Vijayawada.
[3]Professor, Dept. Of CS & SE, Andhra University, Visakhapatnam

Abstract

This paper presents a novel methodology based on Genetic Algorithm to model a new classification system. The applications of Genetic Algorithm to Data Mining techniques currently evolving so as to enrich the efficiency of the traditional classification techniques. Genetic Algorithm is one of the Soft Computing technique that is proposed in this paper to obtain the most optimal and relevant features. Feature selection is one of the Data Dimensionality Reduction technique employed to minimize the number of attributes to further classify the datasets and maintains an acceptable classification accuracy.Thus, in this paper the Genetic Algorithm is used as a feature selection technique to select significant attributes in order to obtain most relavant and important attributes that are necessary for classification. This paper used the Seperability and Correlation Measures between the attributes as the evaluation function to the Genetic Algorithm. The performance of the proposed approach is observed by applying the technique to Clinical Dengue Datasets and retrieved relative and relevant features. The obtained results shows the accuracy and validity of the approach. The analysed dengue data set is downloaded from http://www.ncbi.nlm.nih.gov/gds.
Keywords: Data Mining, Classification, Dimensionality Reduction, Genetic Algorithm, Decision Tree.

1 Introduction

The presence of large volume of information in huge databases and its analysis is becoming extensively expensive and inefficient. The increasing power of processors and their low cost compelled with the essential to evaluate massive data sets that allows the expansion of novel techniques, depending on the capacity to explore a large solution space. The term Data Mining is defined as the process of searching

through an enormous volume of data, stored in a database, to determine remarkable and valuable information that are previously unidentified [23]. The intelligent analysis of these huge databases by Data Mining is extremely useful, as it is possible to construct computational models that give support to professionals during decision making [4, 5].The key point in data mining is the application of these methods to common real problems in business, biology, telecommunication in a fashion that makes these technologies accessible to the experienced knowledge worker as well as the trained statistics professional [26].

Classification is one of the technique in Data Mining that are employed to extract models to define significant class labels for the purpose of data analysis and to forecast future developments. It represents a learning paradigm that segments the data by allocating it to groups, or classes. A wide research on the classification techniques are proposed in various disciplines such as Numerical Taxonomy, Machine Learning etc. Many techniques generated in these arenas are being distributed to resolve the complications of classification and rule generation in Data Mining and Knowledge Discovery in large databases. The classification model envisage the previously unknown object as either it belongs to the class or not [27].The objective of classification is to accurately predict the target class for each case in the data.Classification Accuracy is measured as the proportion of exact likelihoods considering positive and negative inputs. The Accuracy for the classifier is vastly reliant on the database distribution that effortlessly leads to erroneous assumptions about the system performance.

1.1 Feature Selection

Feature Selection is one of the preprocessing step in classification, through which the complexity of the problem is diminished by the eradicating irrelevant features without diminishing the performance rate of classification task and to reduce the price and running time of the method [15]. Usually data mining technologies perform better with lower-dimensional compared to higher-dimensional data. Thus large diverse technologies has been developed for choosing an optimum subset of attributes from a higher set of conceivable features. The feature selection techniques categorized mainly into two main strategies. In the first category precise strategies are established depending on the domain knowledge in order to diminish the number of features used to a controllable size. The second category is used when the domain knowledge is inaccessible or expensive to achieve. In this case, generic heuristics, principally greedy algorithms, are applied to select a subset “d” of the existing “m” features.

The feature selection algorithm is estimated using certain standards. An optimum subset selected using one principle may not be ideal according to the alternative principle. An evaluation criterion can be largely characterised into two groups [1]. An independent criterion that tries to estimate a feature subset by features of the training data without considering any mining algorithm. Some popular independent criteria are distance measures[6], information measures, dependency measures, and consistency measures [2] [3]. The second group is dependent criterion that entails a pre-scheduled mining algorithm in feature selection and uses the performance of the mining

algorithm functioned on the particular subset to define which features are selected. This proposed paper used a dependent correlation measure criteria for feature selection. Fig.1 represents the general feature selection structure. The objective of feature selection is to attain a feature space with low dimensionality, retaining of adequate data, and improving the seperability in feature space and comparability of features among other features in the same category [28].

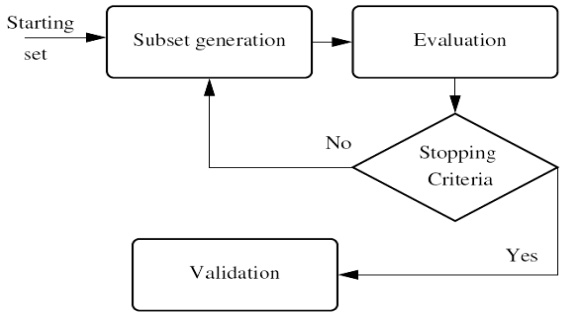

Fig. 1: General Feature Selection Structure

In order to improve the classification accuracy and error rates, this paper presents a novel methodology that is mainly interested in Pre-processing of large Dengue Database by using an improved Genetic Algorithm and Classifying the dengue datasets accordingly. The feature selection technique based on the Improved Genetic Algorithm is used to extract the correlated and relevant attributes using a dependent criterion in order to Pre-process the datasets. Then a Decision Tree Classification Technique is applied on the retrieved relevant feature for the classification of dengue disease.

1.2 Organization of the Paper

A brief introduction to Data Mining Classification and Feature Selection given in this section. The Section 2 discusses the existing methodologies for Classification and Feature Selection techniques using the Genetic Algorithms. The proposed novel methodology for better classification is described in the section 3. Section 4 gives the Experimental Result and its analysis. Finally section 5 concludes the proposed algorithm with improved classification accuracy and error rate.

2 Existing Methodologies

A first generation of GA's proposed for data mining have been built for classification tasks. REGAL [14] like SIAO [11] learns first order rules describing classes. COGM [13] addresses multi-class problem domains. GAMMER [12] is developed for running GA's on large scale parallel data mining. Applications of Genetic Algorithm in Data Mining problems were optimizations of learning tools like artificial neural networks. Since they maintain a population of individuals that may represent patterns such as

rules, they are used as pattern mining techniques as well. Classification techniques are well defined as a foremost category of data analysis in genetic algorithm which usually uses rule based approach[16]. The genetic algorithms have been accompanied using greedy methods in [17] [18] [19] [20] [21]. The latest ten years of study of evolutionary algorithms described in [22].

Numerous methodologies with Genetic Algorithm to resolve the feature selection have been proposed in the literature [30]. In 2005, some enhanced and incorporated forms of feature selection algorithms [34, 38] were recommended and authenticated for their exactness.In late twenties, some astonishing pains have been put in the field of Genetic Algorithm based feature selection methodologies. Feng Tan, Xuezheng Fu, Yanqing Zhang,and Anu G.Bourgeois in 2007 [35] suggested a mechanism where numerous prevailing feature selection methods are applied on a database. [36] Presents, a novel wrapper feature selection methodology for the hyperspectral data, which assimilates the Genetic Algorithm and the SVM classifier through an appropriate designed chromosome and fitness function.

One of the newest procedure for GA centred attribute selection is given in [37]. In this procedure WEKA [36] GA is used a random search method with four diverse classifiers, namely decision tree (DT), C4.5, Naïve Bayes, Bayes networks and Radial basis function as the induction method wrapped with GA. C. H. Tsang proposed ant colony clustering and feature extraction for anomaly intrusion detection [9]. In [10], Bazi and Melgani proposed a support vector machine (SVM) classification system that allows for detecting the most distinctive features and estimating the SVM parameters by using a GA.

3 Proposed Approach

In this paper, a novel methodology is proposed to find the best optimal features from the large database. In this approach, the Genetic Algorithm is used as random selection algorithm to efficiently explore a vast search space which is frequently required in case of attribute selection and choosing attributes to maximize the probability of desired classification. The proposed methodology is broadly implemented in two phases as shown in Fig.2. They are:

Phase 1: Feature Selection using Genetic Algorithm (GA) and Seperability Correlation Measure (SCM)

Phase 2: Classification of selected attributes using Decision Tree (DT) algorithm

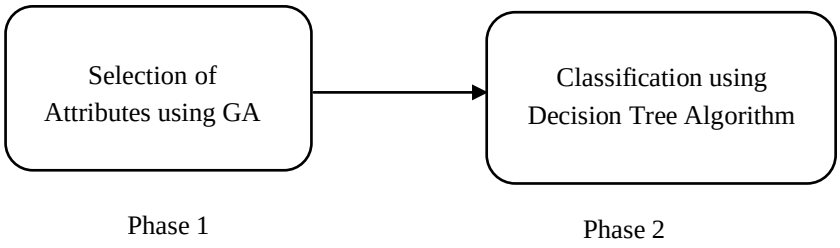

Fig. 2: Structure of Proposed Methodology

3.1 Feature Selection Using Genetic Algorithm

Feature Selection aims to decrease the computational cost of feature measurement, improve classifier efficiency, and permits higher classification accuracy depending on the procedure of deriving novel features from the original features. The best subset is selected on the basis that its value is greatest for a criterion function which signifies the notion of data sets.The two techniques used for feature selection in the proposed methodology are Genetic Algorithm (GA) and Seperability Correlation Measure (SCM)

3.1.1 Genetic Algorithm

Genetic algorithms (GA) are optimized Machine Learning algorithms depends roughly on procedures of biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. GA is motivated by the genetic procedure of biological organisms. It consists of several solutions called chromosomes or individuals which determines the attributes for each individual [30]. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. This type of representation is relative to the position.A set of the chromosomes is made to form a population. The merit of each chromosome is evaluated by using a fitness function.

Genetic Algorithms have two important features. The first is the employment of an algorithmic equivalent of natural selection. When chromosomes are chosen as parents during the reproduction process, the probability that a given chromosome will be chosen is biased according to its fitness. Thus, fittest chromosomes will tend to have more children than the less fit ones. The second feature is the use of mutation and crossover operators during reproduction. The production of high-performance chromosomes can be greatly speeded up with crossover that works to combine subparts of good solutions from multiple parents to a single child.

Parameters of Genetic Algorithm: Several aspects from GA's must be determined when they are used in Feature Selection problem.

A. *Representation of Chromosomes*: Different types of attributes like binary, nominal and numerical attributes can be encoded to chromosomes.
B. *Length of the Chromosomes*: The length is primarily selected by the number of features of the samples as each gene in the chromosome signifies the value of the feature occupying that position.
C. *Size of Population*: There should also be a relation between the number of elements in the population and the population size depends on the way of initialization of the population. Although in most of the approaches the population is initialized randomly.
D. *Genetic Operations*: The standard genetic operators like selection, crossover, and mutation are the most widely used.
E. *Fitness Function*: The merits of each chromosome can be evaluated using this function. In this paper, correlation based function that measures the relevancy between the attributes are used as a fitness function.

3.1.2 Seperability Correlation Measure (SCM) as Fitness Function in the Proposed Methodology

The Seperability-Correlation Measure (SCM), which was first proposed in [7] for determining the significance of the original attributes. The SCM comprises of two parts, the intra class distance to interclass distance ratio and an attribute-class correlation measure [25]. The attribute-class correlation measure is applied to estimate the power of every attribute affecting the class label for each pattern [31]. The greater the correlation factor, the more essential the attribute is for defining the class labels of patterns [24]. The ratio of intra class distance and the interclass distance replicates the class Seperability. Consequently, to recognize a subset of features that maximizes the Seperability among the classes is necessary objective of feature selection. The average pairwise distance between patterns of the two classes replicates the Seperability of the two classes, i.e., the superior the average pairwise distance, the healthieris the Seperability of the two classes [8]. When the number of patterns are high, the cost of the pairwise-distance calculation is high.The comparative significance of a feature is given by its relative magnitude of the SCM.

$$R_k = \chi S_k + (1 - \chi) C_k$$

Where $S_k = {S_{wk}}/{S_{bk}}$ And

$$S_k = {(S_k - \min(S_k))}/{(\max(S_k) - \min(S_k))}$$

is the normalization of S_k $\max(S_k)$ and $\min(S_k)$ are the maximum and minimum of all S_k, respectively. $k = 1,2, \ldots . n$ are the number of attributes. S_{wk} And S_{bk} are intraclass and interclass distances computed with the k^{th} attribute removed from each pattern, respectively.

$$C_k = {(C_k - \min(C_k))}/{(\max(C_k) - \min(C_k))}$$

is the normalization of C_k .χ is a weight parameter where $1 \geq \chi \geq 0$ and χ is determined empirically: the best choice of χ should lead to a subset of attributes which consequence in the highest classification accuracy. The prominence stage of attributes are ranked using the values of R_k. The greater the magnitude of R_k, the more prominent is the k^{th} attribute.

In the proposed methodology, the attributes of the training dataset are the initial population that is encoded to numerical string chromosomes and employed in genetic algorithm. The fitness function used in this methodology is the Seperability Correlation Coefficient Measure. The fitness value for the chromosomes are determined using this function. The genetic operations i.e. crossover and mutation are implemented by selecting a pair of fittest attributes, then the new attributes are replaced with the selected attributes. The termination for the algorithm will attain if maximum number of generations has been reached or if there is no changes to the population best fitness for specified number of times of generations. Algorithm 3.3

and Fig.3 described the genetic algorithm for feature selection using Seperability Correlation Measure. Only those operations fits to perform genetic operations that are having lower correlation coefficient, that is lower the correlation coefficient the higher is the fitness value.

3.2 Classification of Selected Attributes Using Decision Tree (CART) Algorithm

The Decision Tree algorithm fundamentally approximate a suitable features for separation of objects representing dissimilar classes. Classification of patterns may depend on a very few of the most significant attributes, determining which attributes should be retained for the original notion of data is essential in feature selection techniques [29].

The traditional Decision Tree classification technique is performed on the obtained optimized attributes on each sample which generates decision tree rules for the attributes of training data[32][33].Then the generated rules are given to testing data for classification. The classification accuracy is calculated for the testing data. This decision tree recursively partitions a data set into smaller subdivisions on the basis of tests applied to one or more attributes at each node of the tree.

3.3 Proposed Algorithm

1. A dataset with M number of samples and N number of attributes in every samples are considered.
2. The N number of attributes or features of each sample are encoded into numeric chromosomes.
3. The initial population by selecting the parental chromosomes from the dataset is generated.
4. The fitness value for each chromosomes calculated using the Seperability Correlation measure given by
$$R_k = \chi S_k + (1 - \chi) C_k$$
And fitness value = 1- R_k
5. If termination condition is reached go to step 9 else go to step 6.
6. The selection operation on chromosomes performs to select the pair of fittest attributes.
7. Then the genetic operations (crossover and mutation) performed on the pair of selected attributes.
8. Now replace the previous selected attributes with the new attributes computed and go to step 4
9. Rank of all the selected and fittest attributes of each sample obtained from genetic operations.
10. Select top K number of optimized attributes of each sample from the ranked ones and give this as input to the Classification and Regression Tree Algorithm
11. Decision Tree classifier (CART) classifies all the samples into classes depending on the optimized attributes as targets.

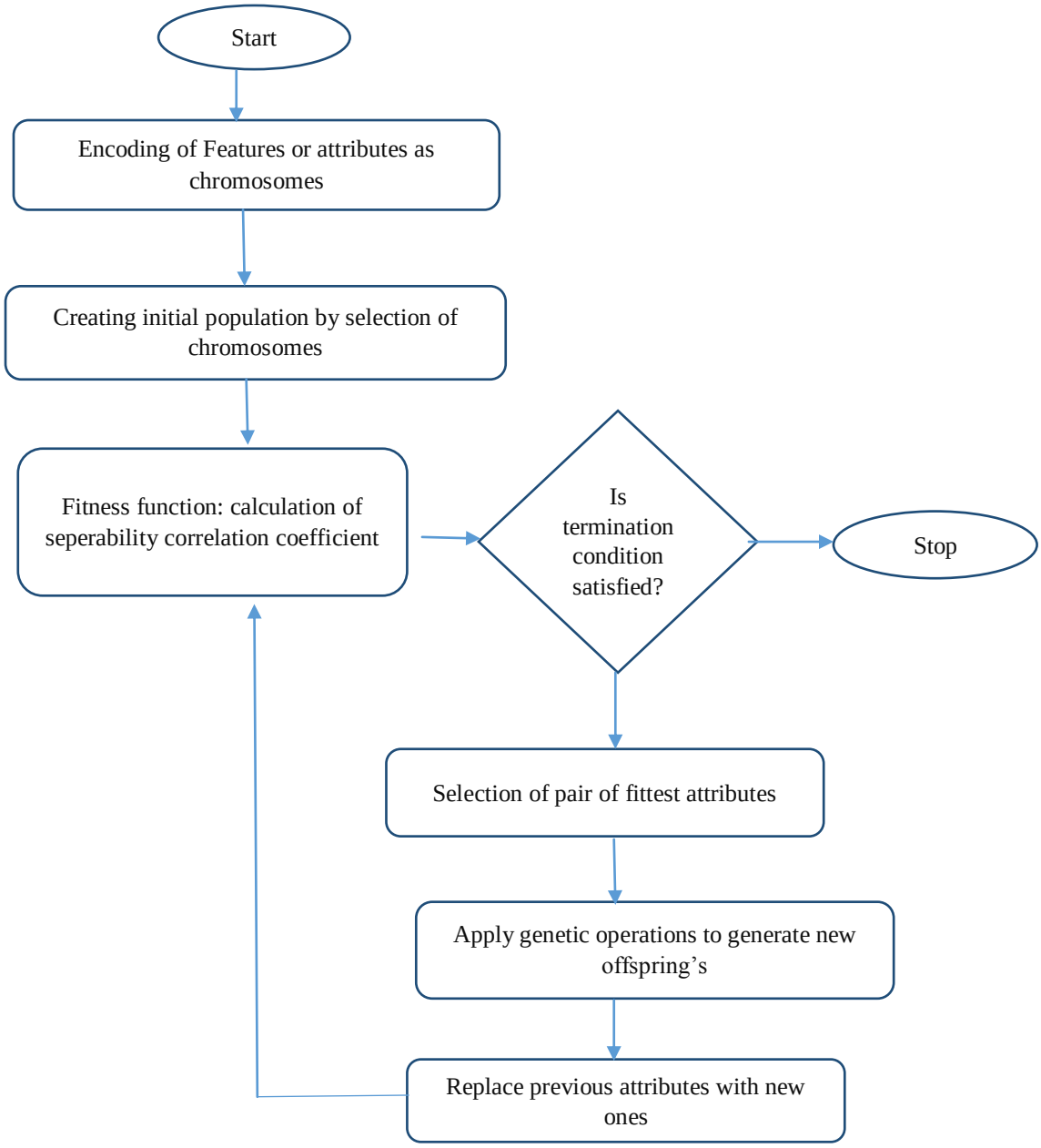

Fig. 3: Flow Chart For The Proposed Methodology

4 Experimental Result and Analysis

The Experimental analysis for the proposed approach is carried out using the datasets available at http://www.ncbi.nlm.nih.gov/gds. In this paper the experiment is performed on the total of 1,275 patient's records, each having 17 attributes along with the target class.

Fig 4 represents the interface for running the proposed methodology where the information about the number of features, sample testing, training information, and the number of optimized features needed which are taken from the user.

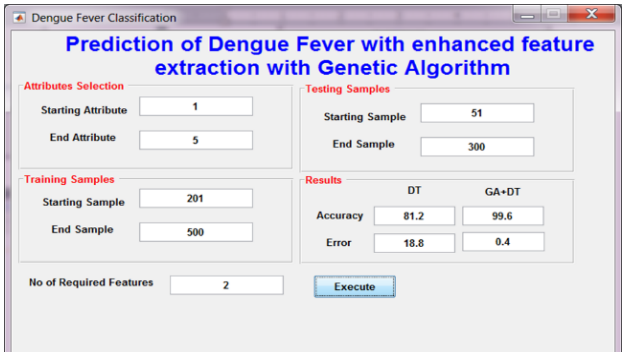

Figure 4: Interface designed to run proposed algorithm

Table 1: Decision Tree and Proposed Methodology Classification accuracy and Error Rate

Attribute Selection		Training Samples		Testing Samples		Features		DT Results		GA+DT Results	
Start	End	Start	End	Start	End	Only DT	GA+DT	Accuracy	Error	Accuracy	Error
1	3	201	500	51	500	3	2	88.00	12.00	98.89	1.11
1	5	201	500	51	500	5	4	88.00	12.00	99.56	0.44
1	5	201	500	51	300	5	2	81.20	18.80	99.60	0.40

The classification accuracy and error rates are obtained for traditional Decision Tree (CART) algorithm and the proposed Seperability Correlation Measure based fitness function in Genetic algorithm. Table 1 and Fig 5 represents the accuracy and error rate of Traditional Decision Tree and Genetic Algorithm combined with Decision Tree for different optimized features selected by the user that vary with the obtained fitness values.The Table 1 clearly represents that the classification accuracy for the proposed methodology is higher when compared to the traditional approach and the error rate is less for the proposed approach compared to traditional approach. Fig. 6 and Fig. 7 represents the decision trees for traditional and proposed approach respectively.

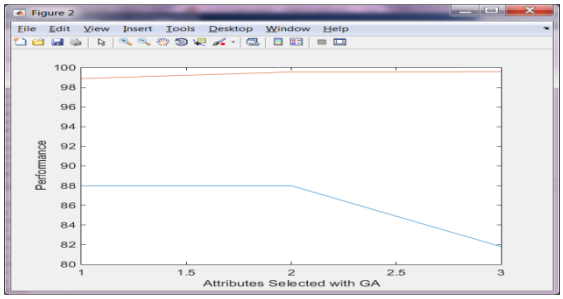

Fig. 5: Performance Vs Selected Attribute for Decision Tree Classifier and Proposed Methodology

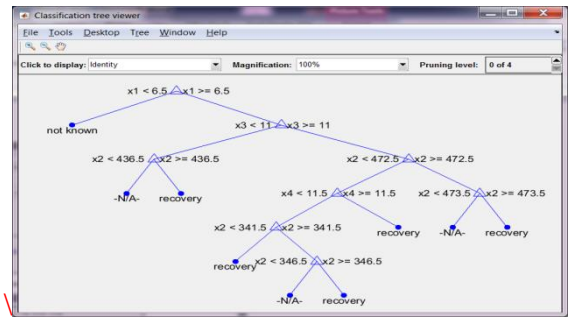

Fig. 6: Decision Tree for 5 attributes

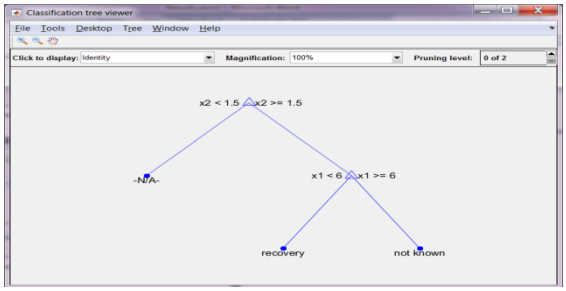

Fig. 7: Decision Tree for 3 attributes selected by GA selected out of 5 supplied attributes

5 Conclusions

The Genetic Algorithm is one of the Soft Computing Technique employed in this paper for Optimized Feature Selection. The classification of the Patients information is performed using the proposed classification system that considered the Seperability Correlation Coefficient measure for the correlating the attributes. The experimental results of the proposed paper showed a better classification accuracy that is obtained and also showed that there is also a considerable decrease in the error rate compared to the existing classification system. From the experimental analysis it is also inferred that

the proposed methodology is effective and efficient with respect to the number of correctly classified patterns.

6 References

[1] Huan Liu, Lei Yu,"Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, pp. 491-502, April-2005.

[2] M.A. Hall,"Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning", Proc. 17th Int'l conf. Machine Learning, pp. 359-366, 2000.

[3] H. Liu, H. Motoda,"Feature Selection for Knowledge Discovery and Data Mining", Boston: Kluwer Academic, 1998.

[4] Usama M.Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and RamasamyUthurusamy, "Advances in Knowledge Discovery and Data Mining", ed. U.M.Fayyad et al., AAA1 Press, 1996.

[5] J. Han and M. Kamber, "Data mining: concepts and techniques", Morgan Kaufmann, 2006.

[6] N. Gnaneswara Rao, T. Sravani, V. Vijaya Kumar, "OCRM: Optimal Cost Based Image Retrieval", International Journal of Multimedia and Ubiquitous Engineering, Vol.9, No.4 (2014), pp.327-342.

[7] Fu, X.J., Wang L.P., "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance", IEEE Transactions on Systems, Man and Cybernetics, Part B. 33, 399–409, 2003.

[8] Devijver P.A., Kittler J,"Pattern Recognition: a Statistical Approach", Prentice-Hall International Inc., London, 1982.

[9] C.H. Tsang, "Ant Colony Clustering and Feature Extraction for Anomaly Intrusion Detection", Swarm Intelligence in Data Mining, Springer Berlin/Heidelberg, 2007, pp. 101-123."

[10] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images",IEEE Trans, Geosci. Remote Sens., vol. 44, no. 11, pp. 3374–3385, Nov. 2006."

[11] S.Augier, G.Venturini, Y Kodratoff, "Learning firstorder logic rules with a genetic algorithm", Proc. Of the lrst Int. Conf on Knowledge Discovery and Data Mining, AAA1 Press, 1995.

[12] I.W. Flockart, N.J.Radcliffe, "GA-Miner:Parallel Data Mining with Hierarchical genetic Algorithms", Final Report,O,University of Edimburgh, Novembre 1995."

[13] D.P Greene and S.F Smith, "Using coverage as a model building constraint in learning classifier systems", Evolutionary computation, Volume:2 , Issue: 1, Page(s): 67 – 91, 19 May 2014.

[14] P.Neri and A. Giordana, "A parallel genetic algorithm for concept learning", Proc. of the 6 Int.Conf. On Genetic Algorithms, Morgan Kaufinan, 1995.

[15] F. Mhamdi & M. Kchouk, "Hierarchical Algorithm for Pattern Extraction from Biological Sequences", Proceedings of the 6th International Conference on Bioinformatics and Computational Biology, Las Vegas, USA (2014)."

[16] Liangxiao Jiang, Dianhong Wang, ZhihuaCai, and Xuesong Yan, "Survey of Improving Naïve Bayes for Classification", Proceedings of ADMA International conference,Springer, pp. 134–145, 2007.

[17] Alex A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery",Springer-verlag, Newyork, 2003."

[18] J. Eggermont, J. Kok, and W. Kosters, "Genetic programming for data classification: Refining the search space", Proceedings of the 15th Belgium/Netherlands Conference on Artificial Intelligence, pp. 123–130, 2003.

[19] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm",Proceding of the 13th International conference on Machine Learning, pp. 148–146, Morgan Kaufmann, 1996.

[20] L. Hyafil and R. Rivest, "Constructing optimal binary decision trees is NP-complete", Information Processing Letters, vol. 5 No.1, pp.15–17, 1976.

[21] M. Keijzer, J. J. Merelo, G. Romero and M. Schoenauer, "Evolving objects: A general purpose evolutionary computation library", Proceedings of Evolution Artificielle'01, Springer Verlag, vol. 2310, pp. 231–244, 2001.

[22] W. B. Langdon and S. M. Gustafson, "Genetic Programming and Evolvable Machines", ten years of reviews, journal of Genet Program Evolvable Machines, vol 11, pp.321–338, 2010.

[23] Vipinkumar, Michael steinbach and Pang-Ning Tan, "Introduction to Data mining", Third Impression, Pearson Education, 2009.

[24] Rajdev Tiwari, Manu Pratap Singh,"Correlation-based Attribute Selection using Genetic Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010.

[25] Asha Gowda Karegowda, A. S. Manjunath&M.A.Jayaram, "Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 271-277, July-December 2010.

[26] SadjiaBenkhider, Ahmed Riadh Baba-Ali, HabibaDrias, "Evolutionary Approaches for the Extraction of Classification Rules", Journal International Journal of Applied Metaheuristic Computing archive, Volume 5, Issue 1, Pages 1-19, January 2014.

[27] Saraee, M.H, Isfahan Sadjady, R.S., "Optimizing Classification Techniques Using Genetic Programming Approach", Multitopic Conference, IEEE International Date of Conference, ,Page(s):345 – 348, Dec 2008.

[28] Mahdi Esmaeili 1, Fazekas Gabor, "Feature Selection as an Improving Step for Decision Tree Construction", 2009 International Conference on Machine Learning and Computing, vol.3, IACSIT Press, 2011.

[29] Mrs. ShantaRangaswamy, Dr.Shobha G, Sandeep R V, Raj Kiran, "Comparative Study of Decision Tree Classifier with and without GA based feature selection", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.

[30] NidapanChaikla and Yulu Qi, "Genetic Algorithms in Feature Selection", IEEE 1999.

[31] Rajdev Tiwari, Manu Pratap Singh, "Correlation-based Attribute Selection using Genetic Algorithm", International Journal of Computer Applications (0975 – 8887), Volume 4– No.8, August 2010.

[32] Huang Ming, NiuWenying Liang Xu, "An improved Decision Tree classification algorithm based on ID3 and the application in score analysis", Control and Decision Conference, Page(s):1876 – 1879, June 2009.

[33] R.C., São Carlos, Cerri, R. Jaskowiak, P.A. de Carvalho, "A bottom-up oblique decision tree induction algorithm", Barros, Intelligent Systems Design and Applications (ISDA), 2011, 22-24 Nov. 2011, Page(s): 450 – 456, IEEE, 2011.

[34] Noelia Sanchez-Marono, Amparo Alonso-Betanzos and Enrique Castillo, "A New Wrapper Method for Feature Subset Selection", Proceeding of European Symposium on Artificial Neural Networks, Belgium, 2005.

[35] Feng Tan, Xuezheng Fu, Yanqing Zhang,andAnu G. Bourgeois, "A genetic algorithm-based method for feature subset selection", Soft Computing - A Fusion of Foundations, Methodologies and Applications, Volume 12 , Issue 2 , Pages: 111 – 120, Springer-Verlag, 2007."

[36] Li Zhuo , Jing Zheng, Fang Wang, Xia Li, Bin Ai and Junping Qian, "A Genetic Algorithm based Wrapper Feature selection method for Classification of Hyperspectral Images using Support Vector Machine" The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7. Beijing, 2008."

[37] M.A.Jayaram, Asha Gowda Karegowda, A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications (0975 –8887) Volume 1 – No. 7, pages 13-16, 2010.

[38] Huan Liu and Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering" IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, Pages: 491 - 502, 2005."