International Journal of Advanced Science and Technology
Vol.114 (2018), pp.127-138
http://dx.doi.org/10.14257/ijast.2018.114.12

Performance Evaluation of Data Mining Classification in Educational System using Genetic Algorithm

Navneet Kaur[1] and Jaskaranjit Kaur[2]

[1]*Dept. of Computer Science & IT, LKC, Jalandhar, Punjab, India*
[2]*Dept. of Computer Science & IT, LKC, Jalandhar, Punjab, India*
[1]*saininavneet@gmail.com*, [2]*kaurjaskaranjit@gmail.com*

Abstract

With the development in the field of Information Technology and Computer Science, high capacity of data appears in our lives. Data Mining helps us to find out useful information from large dataset. After retrieving information from large dataset, we have applied Genetic Algorithms to optimize the classified information. This paper provides a concise and representative review for classifying students in order to predict their performance on the basis of features extracted from the data logged in an Education System. We have discussed number of classifiers like 1-NN, K-NN, Naïve Bayes and Decision Tree (C4.5, C5.0, CART). The performance evaluation of these classifiers on students dataset is done using various attributes and we found that CART is the best data mining classification technique among the 6 classifiers when we use two classes and K-NN is the best data mining classification technique among the 6 classifiers when we use three classes.

Keywords: *Data Mining, Classifiers, 1-NN, K-NN, Naïve Bayes, Decision Tree, Genetic Algorithms*

1. Introduction

Data mining also identified as ‘‘knowledge discovery in databases” (KDD), is the way of finding out meaningful patterns from huge databases. It is the process of converting the low-level data into high-level knowledge. Data mining is a technique of analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. It is a technology which is used with large potential to help companies and big enterprises to find out their customer's behaviors [11, 12]. Classification is one of the most well-known and most successful data mining techniques used to classify and predict values. There are several different classification methods and techniques used in Knowledge Discovery process and data mining. Every data mining technique or method has its own advantages and disadvantages. Thus, this paper uses multiple classification methods to confirm and verify the results with multiple classifiers [13]. Then the results of different classifiers are optimized using GA. In the last, the best results are selected in terms of accuracy and low error rate. The rest of the paper is arranged into 6 sections. Section 2 contains the detail of data mining classifiers process implemented in this study followed by Section 3 in which we discussed about Genetic Algorithm. Section 4 contains dataset attributes & class labels which includes a representation of the collected dataset, an exploration and visualization of the data. In next section optimization of classifiers using GA is done and finally the implementation of the classifiers with GA in MATLAB is done with the final results. In last section the work is concluded and insights about future work are included.

Received (February 5, 2018), Review Result (May 1, 2018), Accepted (May 7, 2018)

ISSN: 2005-4238 IJAST

2. Classification and its Types

Classification is a data mining technique that assigns items in a group to target categories or classes. In Classification models categorical class labels are predicted. The Data Classification process includes two steps [13].

- Building the Classifier
- Using that Classifier for Classification

The major issue in classification is preparing the data for Classification and Prediction. Preparing the data involves the following actions −

- Data Cleaning − Data cleaning means eliminating the noise and treatment of missing values. The noise can be eliminated by using various smoothing
- **Normalization** – The data required for classification can be transformed using normalization. Normalization is scaling technique or a mapping technique or a pre processing technique in which we can find new range from an existing range.
- **Generalization** – we can also transform the data by generalizing it to the higher concept. For this we can use the concept hierarchies.

The various classifiers used in this paper are:

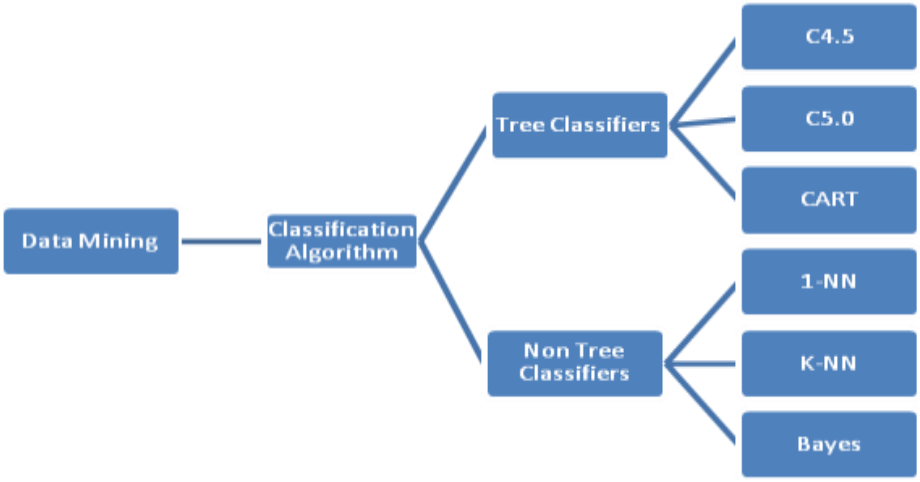

Figure 1. Classifiers Types

2.1. 1-NN for Classification

Let's see how to use KNN for classification. In this classification, we are given with few data points (tuples) for training and also with new data for testing without any label. Our main motive is to assign best class label to the new point. The algorithm has different behavior depending upon the value of k. KNN is a *non parametric lazy learning* algorithm. Non parametric means that it does not make any assumptions on the underlying data distribution. It is a lazy learner in which training dataset is stored. On querying similarity between test data & training dataset, records are calculated to predict the class of test data. The input to this algorithm is the K closest training example and output is the class membership. When K=1 (where k is the number of neighbors) it means object is assigned to the class of single nearest neighbor. It means we are considering first immediate neighbor. This value of k decides how many neighbors (where neighbors are defined based on the distance metric) impact the classification. This is usually a odd number The similarity between test data & training data is mostly calculated using the Euclidean distance.

2.2. K-NN

Nearest-neighbor classifier is based on learning by analogy. This means by comparing a given test data with training data that are similar classification is done. The training data are described by n attributes. Each row of data (tuple) represents a point in an n-dimensional space. When we are given with an unknown data, then k-nearest-neighbor classifier searches the pattern for the k training tuples that are closest to the unknown data. These k training tuples are the k "nearest neighbors" of the unknown tuple. The similarities between the data points are defined in terms of a distance metric, like Euclidean distance.

$$Dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i} - X_{2i})^2} \quad (1)$$

This is the Euclidean distance between two points X1and X2

We normalize the values of each attribute before using the Euclidean distance formula. This is done in order to prevent attributes with initially large ranges from outweighing attributes with smaller ranges. For K-NN classification, the unknown tuple is assigned to the most common class among its k-nearest neighbors [18].

2.3. Naive Bayes

It is a classification technique based on Bayes Theorem. It works with an assumption of independence among predictors which means, a Naive Bayes classifier assumes that the presence of a particular feature in a class is not related to the presence of any other feature. Bayesian classifiers are a popular supervised classification algorithm. It was introduced as text categorizing, the problem of judging documents as belonging to one category or the other with word frequencies as the feature. One important feature of Naive Bayes is that it requires a small amount of training data to estimate the parameters necessary for classification. Naïve Bayes is a conditional probability model [19].

2.4. Decision Tree Algorithm

Decision Tree algorithm comes under supervised learning algorithms.

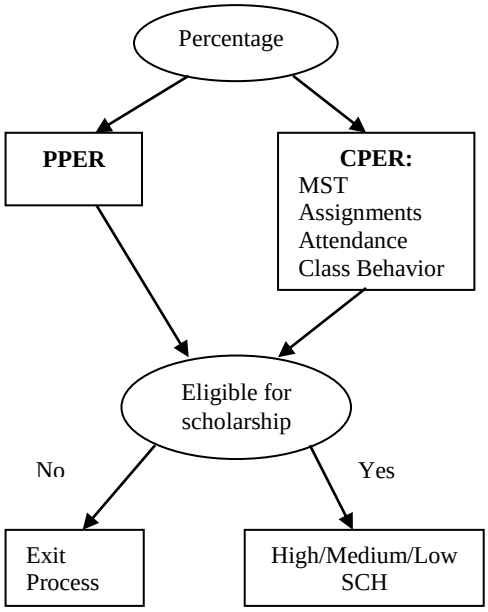

Figure 2. Decision Tree for Percentage-SCH relationship

The basic motive of using Decision Tree is to build a training model which we can use to find out the class of the target variables by **learning decision rules** deduced from prior data (training data). The decision tree algorithm tries to solve the problem, by using tree representation. Each **internal node** of the tree corresponds to an attribute of the dataset, and each **leaf node** corresponds to a class label of the classifier.

Decision Tree Algorithm Pseudo code

1. Position the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be build in a way that each subset consists of data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

Percentage Attribute *i.e.*, PPER and CPER are considered for the eligibility of scholarship. CPER attribute is based on the performance of student in MST, Assignments, Attendance and Class Behavior. Total percentage is based on PPER and CPER attributes. Depending on the value of percentage parameter, we will find the value of SCH. If the student is not eligible for scholarship depending upon the value of PPER+CPER, then it will exit the process and if the student is eligible for scholarship then will assign HIGH/MEDIUM/LOW SCH attribute to student as mentioned in Table 1.

2.4.1 C4.5 Decision Tree Implementation: C4.5 is collection of algorithms for performing classifications in machine learning and data mining. It creates the classification model in the form of decision tree. Decision tree are constructed only using those attributes that are best able to differentiate the concept of learned. C4.5 is categorized into three groups of algorithm: C4.5, C4.5-no-pruning and C4.5-rules. In this paper we use basic C4.5 algorithm. C4.5 implements greedy (i.e., non backtracking) approach in which decision trees are build in a top- down recursive divide and conquer manner. The tree building starts with a training set of tuples and their corresponding class labels. The training set is recursively partitioned into smaller subsets as the tree is being built.

Steps of the System:

1. Select dataset for providing input to the algorithm for processing.

2. Selecting the classifiers

3. Calculate the entropy, information gain, gain ratio of the attributes.

4. According to the defined algorithm of C4.5 data mining process the given input dataset.

6. After this C4.5 processors input the data to the tree generation mechanism.

7. Tree generator generates the tree for C4.5 and improved C4.5 decision tree algorithm [16].

Pseudo Code:

1. Check for base cases.

2. For each attribute calculate the Normalized information gain for splitting an attribute.

3. Out of this select the best attribute which has the highest information gain.

4. Find a decision node that splits the best, as root node.

5. Recurs on the sub lists obtained by splitting on best of a and add those nodes as children node [17].

2.4.2 C5.0: C5.0 algorithm is an extension of C4.5 which follows the rules of C4.5 algorithm. C5.0 is the classification algorithm which we can apply on large data set. C5.0 is superior to C4.5 on the basis of efficiency and the memory. C5.0 works by splitting the sample based on the field that provides the maximum information gain. The C5.0 algorithm split samples on basis of the biggest information gain field. The sample subset that we get from the previous split will be split afterward. The process will continue until the sample subset cannot be split further and is usually according to another field. At last we examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be discarded. The splitting in this algorithm is done on the basis of Information Gain. Gain is computed to estimate the gain produced by a split over an attribute [14].

C5.0 algorithm has features like:

1. We can view a large dataset as a set of rules which we can easily understand.
2. In C5.0 algorithm we get the acknowledgement on noise and missing data.
3. C5.0 algorithm solves the problem of over fitting and error pruning.
4. In C5.0 classifier can foresee the relevant and irrelevant data. [15].

2.4.3 CART Algorithm: Classification and Regression Trees (CART). When the value of the target attribute is ordered, it is called regression tree and when the value is discrete, it is called classification tree. This algorithm uses the binary tree to divide the forecast space into certain subsets. Tree's leaf nodes correspond to different division areas which are determined by Splitting Rules relating to each internal node. CART uses GINI Index method to find out in which attribute the branch should be generated. The basic approach is to select the attribute after splitting with minimum GINI index. The decision-tree generated by CART algorithm is a simple structured binary tree [14].

3. Genetic Algorithm and Proposed Model

Genetic algorithm is an adaptive heuristic method which is based on the "survival-of-the-fittest" principle; genetic search proves particularly effective when the search space is very vast for classical search methods to examine efficiently. The genetic algorithms try to find a best or very good solution to the problem by genetically breeding the population of individuals. The genetic algorithm transforms a population of individual candidate, each with its associated fitness value into a new generation of the population using the Darwinian principle of reproduction and survival of the fittest and naturally occurring genetic operations such as crossover and mutation [1]. Each individual in the population represents a possible solution to a given problem [2]. Before running Genetic algorithms, we will define a relevant encoding of chromosome with different attributes to solve a problem, select an objective function for fitness, and construct genetic operators. In order to run Genetic algorithms, we have generated an initial population consisting of chromosomes having different attributes and evaluated these chromosomes using the defined objective function. And then we select two chromosomes randomly and apply crossover and mutate them and replace a low quality chromosome with a new one of high quality. Higher the fitness value, higher will be the chance that it will survive in next generation. As these processes have been repeated, the population will consist of high quality chromosomes.

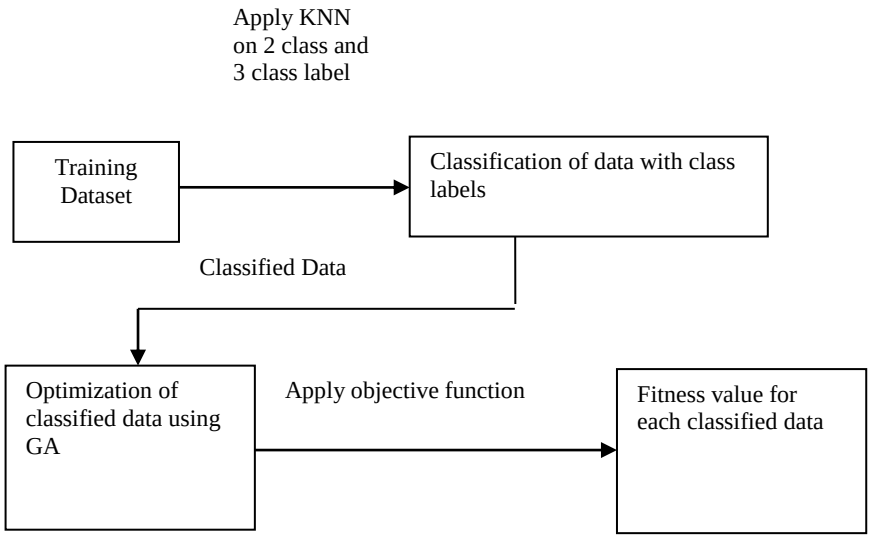

Figure 3. Proposed Approach

Student's data collection: This is the first step of classification process in which we collected the data of various students from University through Questionnaires.

Pre-Processing: After collecting the data it is pre-processed in order to fill in the missing data, remove noise and inconsistent data. The documents prepared for next step in Educational classification are represented by a great amount of features and attributes. The value of 0 is used to fill in the missing values of an attribute in the dataset.

Indexing: The student data representation is one of the pre-processing technique that is used to reduce the complexity of the data and make them easier to handle, the questionnaires have to be transformed from the full text version to excel form. To solve this problem, weighting vector is used to assign appropriate weights to each attributes.

Feature Selection: After pre-processing and indexing the important step of classification is feature selection, which improves the efficiency and accuracy of a classifier. The main idea of Feature Selection (FS) is to select subset of features from the training dataset.

Classification of data with class labels and Optimization with GA: After the completion of above steps then various classifiers are applied to classify the dataset. We have used 1-NN, K-NN, Naive Bayes, C4.5, C5.0 and CART in our paper for classification of data and the classified data is optimized using GA.

Performance Evaluations: This is Last stage of classification, in which the evaluations of classifiers is typically conducted experimentally. And the performance is evaluated using error rate.

3.1. GA Operations

1. *Selection and reproduction operator* copies the individuals with the best fitness value. The roulette wheel reproduction method is used to select individual string for next generation [5].

2. *Crossover* is the genetic operator that mixes two chromosomes together to form new offspring. Purpose of crossover operator is exploration of a new solutions and exploitation of old solutions. GA constructs a better solution by applying crossover operator on strings. Higher fitness value has more chance to be selected than lower ones, so good solution always alive to the next generation. We have used a single point crossover,

exchange the weights of sub-vector between two chromosomes, which are candidate for this process [6]. The crossover and mutation probability is set by the user [8].

3. Mutation is the third operator used in our GA process. Mutation involves the modification of few bits of a chromosome with some mutation probability i.e. flipping the bits from 0 to 1 or vice versa. Chromosome string may be better or poorer than old chromosome string. If they are poorer than old chromosome then they are eliminated in selection step. The objective of mutation is restoring lost and exploring variety of data [6]. In genetic algorithms mutation is randomly applied with low probability, typically in the range 0.001 and 0.01, and it modifies elements in the chromosomes. Mutation probability used in the system is 0.02 [4].

Pseudo code listing of the Genetic Algorithm

Begin

t = 0; Initialize Population POP(t);

Evaluate Population POP(t);

repeat

Selection and Reproduction (t)

Crossover (t)

Mutation (t)

Evaluate Population POP(t);

t = t + 1;

until (termination condition)

3.2. Termination Criteria

In genetic algorithms, the termination criteria for stopping the process are required. Genetic Process has large number of iteration. Termination Criteria is important and vital step. This gives information about when to stop the process.

Two Termination Criteria's are used according to the requirement.

1. Fitness Convergence: A termination method that stops the genetic process when the fitness is deemed as converged. Fitness is deemed as converged when the difference between average fitness across the current population and previous population is less than the value specified. We have used fitness convergence value of 0.001 for termination criteria.

$$\text{average fitness of current population} - \text{average fitness of previous population} \leq |\, 0.001 \,|$$

2. Generation Number: If above termination criteria is not achieved after large number of iteration *i.e.*, Fitness convergence criteria is not satisfied. Then another termination criterion is used that is Maximum number of generations. Maximum numbers of generation is the termination method that stops the genetic process when the specified numbers of generations have been run [3].

4. Dataset Attributes and Class Labels

Table 1. Training Dataset for Educational System

Dataset Attributes	Description	Possible values
GENDER	Student's gender	{Male, Female}
NCAT	Nationality category	{Indian, NRI }
LANG	First Language	{Hindi, Punjabi, English, Other}
TLANG	Teaching language in the university	{Hindi, Punjabi, English}
PPER	Previous class %age	{Excellent (90% to 100%), Very Good (80% to 89.9%), Good (70% to 79.9%), Average(55% to 69.9%) Poor(below 55)}
CPER	Current semester Percentage	{Excellent (90% to 100%), Very Good (80% to 89.9%), Good (70% to 79.9%), Average (55% to 69.9%) Poor(below 55)}
Parameters for CPER: **Internal(40)** MST(15) Online Assignments(10) Online Attendance(10) Class Behaviour(5) **External Exams:60**		
SCH	Does the student have any scholarship on the basis of **PPER+CPER**	{Yes, No} If yes then SCH= **High**(50% of Admission fee) **Medium** (35% of admission fee) **Low**(25% of Admission fee)

We consider the dataset of educational system consisting of 7 attributes which includes GENDER, NCAT, LANG, TLANG, PPER, CPER, and SCH. Table 1 describes the attributes of the data and their possible values [9]. Training Dataset for Education System has been shown.. Also different parameter for CPER has been shown. This is based upon internal and external assessment. Internal assessment is dependent on MST, online assignments, attendance and class behavior. Also, we labeled the students in relation to their percentage [7, 10]. And group them into three classes, "*high*" representing high level scholarship for the students who scores 80-99%, "*middle*" representing middle level scholarship who scores 60-79.9%, and "*low*" representing low level scholarship who scores less than 60%. We also labeled the students in relation to their percentage and group them into two classes "*yes/no*" depending upon whether the student is eligible for scholarship or not. Performance of both class labels are evaluated using six classifiers *i.e.*, 1-NN, K-NN, Bayes and decision tree (C 4.5, C5.0, CART).

5. Optimization of Classifiers using GA

We have used MATLAB to implement a GA to optimize classifiers performance. Our data is based on online checking of internal in engineering courses. Based on PPER and CPER, we will find the value of SCH. Our goal is to find the probability that who will get position in university, which minimizes the classification error rate. Firstly, the data is classified using classification algorithms and then it is optimized using GA. GA is applied

on 1-NN, K-NN, Naïve Bayes and Decision tree classifiers data. The performance of each student candidate is evaluated. And based on the values of 7 attributes of each string, GA will find their fitness value based on objective function. Higher is the fitness value of string, higher will be the chance that it will survive in next generation. After finding their fitness values, the basic operations of GA have been applied on data that includes reproduction, crossover and mutation operators. As these processes have been repeated, the population will consist of high quality chromosomes [3]. This process terminated when we get minimum error rate between current population and previous population. Fitness Function is optimized using the genetic algorithm. Choosing an appropriate fitness function is very important task. For each individual in the population, Fitness Value is evaluated as: Σ AiWi where i=1 to 7.

A_i denotes attribute vectors. Attribute vectors are in binary form, *i.e.*, 0 or 1 only. For example, if SCH attribute is present then value will be 1 otherwise it will be 0. 'W' denote weight vector corresponding to each attribute vector depicting importance of each attribute. Values of each weight vector ranges from 0 to 1.

Table 2. Example of Attribute Vector and Weight Vector

DataSet	Attributes	Weighted Value
Student	1 0 1 1 1 1 0 0 1	0.2 0.6 0.5 0.8 0.5 0.6 0.4 0.5 0.2

6. Experiment Results

To assess the performance of the genetic algorithm for different classifiers, a educational training dataset was used. In Figure 4, the graphical representation of the best results of our classifiers is depicted. These results are obtained by considering dataset of 150 students for two classes and three classes label. Different data classification created using 1-NN, K-NN, Bayes, C4.5, C5.0 and CART algorithm in WEKA tool. Then GA has been applied on different classification data. Fitness value for each student candidate is evaluated using objective function *i.e.*, sum of product of attributes and their given weightage. Objective function of population is calculated by aggregating all fitness value of candidate solutions. Then divide this value by number of candidate that are given in the population. Then three basic operations are applied on this current population. After that we got new population which is more optimized than initial population. This process continued till we get minimum error rate between current and previous population. We found that for both types of class labels, optimized results are obtained when GA is applied on data classified using CART.

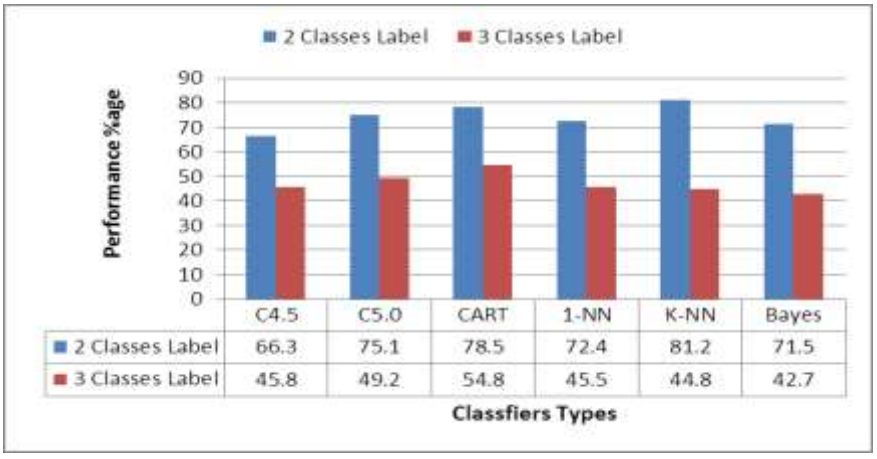

Figure 4. Performance Evaluation of 6 Classifiers based on 2 Class and 3 Class Labels

Table 3 depicts various GA parameters which we have used in our research study. The survey of 150 students have been done and we got the optimized results after 27 and 21 generations with high efficiency and low error rate

Table 3. GA Parameters

GA Parameters	Value	Value
Population Size	150	150
Initial population	112	92
No. of Generations	27	21
Length of Chromosome	7	7
Selection	Roulette Wheel	Roulette Wheel
Cross Over	Single Point	Single Point
Mutation	Low probability 0.02	Low probability 0.02

7. Conclusion & Future Scope

We concluded that the performance of CART algorithm is high when the numbers of classes are less for classification and when we increase the number of classes then K-NN gives best performance in terms of less error rate in educational system. In future we can apply more classifiers like SPRINT, QUEST and collaborate multiple classifiers together. Also we can apply other evolutionary algorithms for best optimization of results.

References

[1] D. H. Kraft, F. E. Petry, B. P. Buckles and T. Sadasivan, "The use of genetic programming to build queries for information retrievals", IEEE, **(1994)**, pp. 468-473.
[2] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning, Reading", MA: Addison-Wesley Publishing Co, **(1989)**.
[3] N. Kaur and J. Singh Budwal, "Search Optimization Using Genetic Algorithms", Fifth International Conference on Neural Networks and Artificial Intelligence' (ICNNAI-08) at Minsk, Belarus, **(2008)**, pp. 302-305.
[4] N. Kaur and J. Singh Budwal, "Intelligent Web Search Optimization with reference to Mutation Operator of Genetic and Cultural Algorithms Framework", 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, **(2014)**, pp. 619-623.
[5] A. Kadar Muhammad Masum, M. Shahjalal, Md. F. Faruque and Md. I. Hasan Sarker, "Solving the Vehicle Routing Problem using Genetic Algorithm", International Journal of Advanced Computer Science and Applications, vol. 2, no. 7, **(2011)**, pp. 126-131.
[6] A. A. A. Radwan, B. A. Abdel Latef, A. Mgeid A. Ali and O. A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science, Engineering and Technology, vol. 17, **(2006)**, pp. 6-12.
[7] B. Minaei-Bidgoli and W. F. Punch III, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System".
[8] C. Romero, S. Ventura, C. de Castro, W. Hall and M. Hong Ng, "Using Genetic Algorithms for Data Mining in Web based Educational Hypermedia Systems".
[9] A. Abu Saa, "Educational Data Mining & Students' Performance Prediction", International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, **(2016)**, pp. 212-220.
[10] S. Natek and M. Zwilling, "Data Mining for Small Student Data Set – Knowledge Management System for Higher Education Teachers", Management, Knowledge and Learning International Conference, **(2013)**, pp. 1379-1389.
[11] G. Singh, J. Kaur and M. D. Yusuf Mulge, "Performance Evaluation of Enhanced Hierarchical and Partitioning Based Clustering Algorithm (EPBCA) in Data Mining", 2015 International Conference on Applied and Theoretical Computing and Communication Technology, **(2015)**.
[12] J. Kaur and G. Kaur, "Clustering Algorithms in Data Mining: A Comprehensive Study", 2015 International Journal of Computer Science and Engineering, vol. 3(X), **(2015)**, pp. 57-61.
[13] I. Bhuvana and C. Yamini, "Survey on classification algorithms for data mining: (comparison and evaluation)", International journal of Advance Research in Science & Engineering, vol. 4, Special Issue(01), **(2015)**.

[14] N. Patil, R. Lathi and Vidya Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 4, **(2012)**.
[15] R. Pandya and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning", International Journal of Computer Applications (0975 8887), vol. 117, no. 16, **(2015)**.
[16] G. L. Agrawal and H. Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008, vol. 3, no. 3, **(2013)** March.
[17] S. Hardikar, A. Shrivastava and V. Choudhary, "Comparison betweenID3 and C4.5 in Contrast to IDS VSRDIJCSIT", vol. 2, no. 7, **(2012)**.
[18] J. Han, M. Kamber and Jian Pei, "Data Mining: Concepts and Techniques",
[19] L. Dey, S. Chakraborty, A. Biswas, B. Bose and S. Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes and K-NN Classifier".

