ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Application of KNN-Genetic Algorithm for Analysing Student Learning in Educational Data Mining Paradigm

Ruchi Jain

Assistant Professor, Department of Computer, Jeev Sewa Sansthan Group of Institutions for Women, Faculty of Management, Bhopal, M.P, India

ABSTRACT: There are number of students who get enrolled for certain courses and various factors are responsible to perceive the knowledge. In order to classify the enormous students into the defined classes some technique must be applied. There are number of techniques introduced in educational data mining for the classification of data. K-Nearest Neighbor, Support Vector Machine, Genetic Algorithm have been discussed in this article. KNN-GA has been introduced to classify the students enrolled in the specified course.

KEYWORDS: K-Nearest Neighbor, Genetic Algorithm, Support Vector Machine, KNN-GA.

I. INTRODUCTION

Educational data mining is a discipline in which the data which is obtained from various educationalsectors. The learners can receive knowledge from different sources such as traditional classrooms, online learning, educational software etc. Ample of methods and techniques have been applied to classify the plethora of data obtainedfor the classification purpose. There are various attributes of the learners which plays a vital role in the analysisand classification of the learning process.

One of the most popular technique used in machine learning is Support vector machine. It produces good accuracy when compared with other data classification. When SVM was compared with KNN-GA then it was proven that the KNN-GA can be much more effective and can accomplishes very good performance over SVM.

Initially, various data mining techniques have been introducedfor the classification of the students' data in educational data mining paradigm. K-nearest neighbor (K-nearest neighbor) has been proven to be the most simple and effective classification technique.K-NN is meant for classifying the unknown instances by relating it to the known instances with the help of some distance measure or similar measure. Additionally, one more process known as genetic algorithm came into existence which is meant for randomized searching and optimizing technique which is led by the concepts of evolution and natural genetics. KNN classification have certain limitations which can be overcome by combining it with genetic algorithm and thus named as KNN-GA.

II. BACKGROUND

K-Nearest neighbor works on the concept of classification of the instances which are closely situated to the specified class are likely belong to the same class rather than the instance which are apart from the specified class or the distance between those instances is greater. To determine the closeness of the instance some distance formula can be used such as Euclidean distance,Minkowski distance and Manhattan distancefor continuous variables and for categorical variables Hammingdistance is used. In this paper, the Euclidean distance is preferred as the distance measure.

Limitations of K-NN are[1]:

(i) Initially, the number of neighbors must be known (i.e. the value of k).
(ii) At each prediction, the complexity of calculation is increased as it uses the whole training set.
(iii) The training samples are alike, as there is no weight difference.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Initially, the major issue was to choose the value of k for the application of K-NN algorithm in classification process. The value of k plays a vital role in the development of robust models and when the value of K found to be greater than 1, then the outcome of class labeling will rely on the neighbors' majority votes. The value of K must be as higher as it can be so that a result obtained must be less sensitive and a smoother function.

Another challenge was to select the appropriate features from the set of various number of features. There can be two stages for the application of neighbourhood classification: first, to select the optimal feature space which is reduced in number than the original dataset and second, is to apply the neighbourhood classification. The feature

selection can be done on the basis of gain, which can be estimated with the help of entropy of the dataset.

In this paper K-NN can be combined with GA in order to overcome the limitations of the K-NN classification process. This algorithm is termed as KNN-GA. In order to apply the K-NN, the distance between the test data and training data has to be calculated and the neighbors with greater distances will be considered for the classification. In the proposed method, at each iteration, the k number of samples are drawn. The fitness function can be calculated by the classification accuracy.

III. SUPPORT VECTOR MACHINE (SVM)

It is derived from the statistical learning theory which is meant for classifying the data points into two disjoint half spaces. The main objective of SVM is tomaximize the geometric error and minimize the empirical classification error, thus it is often called as Maximum Margin classifiers. It works on the principle of Structural risk minimization.

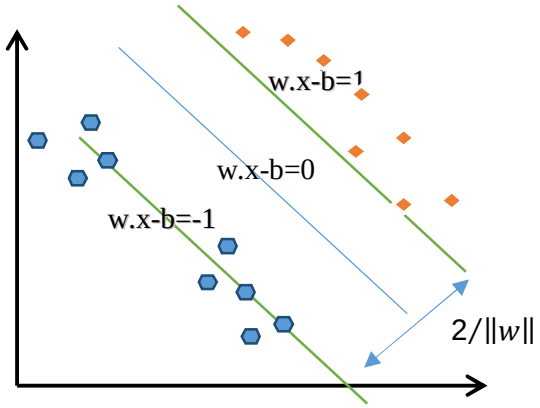

Figure 1: SVM- Maximum margin hyperplanes with samples from two classes

In this technique, two parallel hyperplanes are constructed enclosing the hyperplane which separates the data. The objective of constructing these two hyperplanes is to maximize the distance between separatinghyperplane. It is considered that the larger distance between these hyperplanes the better the generalization of the classifier will be. In these hyperplanes equation b is a scalar value and w is a supporting vector whose value has to be determined. b is an offset parameter which is responsible for the increase in margin whereas w is the vector which is the perpendicular to the separating hyperplane. The separating hyperplane must have the largest margin which can be defined by $2/\|w\|$. The equation for SVM can be expressed as follows:

$L_{(w,h,\alpha)}=1/2\ w^T w\ .\ \sum_{i=1}^{n} \alpha_i\ (y_i(w^T x_i + b) - 1)$

The dual form of the above equation can be represented as:

$L(\alpha)= \Sigma\ \alpha_i - 1/2 \sum_{i=1}^{n} \alpha_i\ \alpha_j\ y_i y_j x_i^T x_j$

Where α is the Lagrangian multiplier, which is to be minimized with respect to w and b and has to be maximized with respect to non-negative α_i.The superscript T demotes the data in training dataset.[2]

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

IV. FEATURE SELECTION

Feature selection is one of the important step in order to enhance the performance of any classification algorithm and also improves the efficacy of training dataset [5]. There can be three types of features such as – relevant, redundant and irrelevant. The irrelevant features does not contribute to the learning process and redundant does not add any additional information to the procedure whereas the relevant features leads to the best performance.
Feature Selection algorithm comprises of wrapper, filters or embedded methods. The wrapper methods make use of predefined classifier which interacts with feature search component to generate the set of features. The feature evaluation, with the help of classifier, estimates the performance. The filters methods ranks the features according to some discrimination measures and without making use of any learning algorithm, the features having higher ranksare selected. The embedded methods apply the feature selection procedure with the learning process.
In this research, the filter method have been used to select the features. The information gain has been evaluated with the help of entropy of the dataset. The higher the gain of the attribute the more valuable feature it will be considered. Once the features selection has been done then the classification process can be implemented.

V. MODIFIED FORM OF K-NN USING GA

In this method, the K-NN is implemented for the estimation of the training dataset.
The volume is considered around the observation point that encompasses k patterns of the training set. It fixes k, considers the nearest patterns from the observation point. The volumes are hyperspheres so the use of Euclidean distance formula has been made to measure the distance between the observation point and the volume. The class having the highest number of valuesin k is selected.
The predefined classes have been introduced on the basis of the grade of the students scored of one subject:

Table 1.1: Categorization of Grades

Classes	Grade
G0 and G1	GH1
G5-G9	GH2
G10-G14	GH3
G15-G19	GH4

The Concept of GA is applied to the subset. Initially, the various points from search space are considered to generate a random population. The points in population are known as strings (or chromosomes). For evaluationofthe goodness of the string the objective and fitness of the function is associated with each string. Each chromosome is encoded with real numbers by following certain patterns[1].By following the principle of survival of the fittest, some of the strings are selected and each of them are copied to go into the mating pool. The genetic operators such as crossover and mutation are applied on these strings in order to generate new population. The process of selection, crossover and mutation is iterated till a termination condition is satisfied.
Reproduction (Selection) –Considering the concept of survival of fittest, the chromosomes are selected from the matting pool. Roulette wheel selection has been used in this research for the implementation of proportional selection strategy. The proportional selection strategy follows a process of generating the number of copies of the chromosomes which is proportional to the fitness in the population for further genetic operations.
Crossover – It is one of the genetic operator which generate the two child chromosomes by exchanging the strings amongst two parent chromosomes. In this research single point crossover has been used.
Mutation – Each chromosome considered can undergo the mutation process with a fixed probability. The random position inchromosome has been selected and that specific bit can be flipped.
In the initial population, the chromosome having the highest fitness value is termed as global maxima. After the application of genetic operator, then the chromosome having the highest fitness value is chosen and stored it as local

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

maxima. The local maxima is compared with the global maxima and if the local maxima is found to be greater than the global maxima then the global maxima is replaced by the local maxima and the process is iterated with the new population generated. The cluster points will be rearranged according to the chromosome having global maxima else if the local maxima is found to be smaller than the global maxima then the iteration is proceeded with the old population. This process is iterated for N number of times.
The algorithm can be summarized as follows:
Step1: Initialize population, the normalized dataset.
Step2: Apply genetic search to the selected population.
Step3: Apply KNN classifier to each category of classified or misclassified data.
Step4: Every attribute will be assigned the specific rank according to the evaluated gain.
Step5: Highest rank attribute will be proceeded to next iteration.
Step6:KNN-GA classification with optimization process is applied to improve classification rate of each classes.
Step7: If (knn-ga-classify (class_knn)>knn-classify (class_knn))
//when more than normal knn classifier
class-data = class_knn-ga;
else
class-data = class_knn;
Step8: Apply reproduction process followed by crossover operation.
Step9: Mutatethe information and generate new population which should be storedseparately.
Step10: Find local maxima for every classes.
Step11: Repeat through step 2 to step 11 if iteration are not satisfied.
Step12: For each test of new mutated population, apply trained base models proceeded by prediction of result and separate the classified/misclassified result by optimized KNN-GA,
Step 13: Evaluate accuracy and error rate of classified data.

VI. RESULTS AND DISCUSSIONS

The classification done by SVM and KNN-GA is compared and it was found that KNN-GA produces better results. The performance parameter considered were accuracy, sensitivity, specificity and error rate. The components which are helpful to estimate the value of these parameters are expressed in the form of confusion matrix:

Table 2: Confusion Matrix

	Predicted values	
Actual values	A: Hits	B: Misses
	C: False Alarms	D: Correct Rejections

Sensitivity determines the proportion of actual hits which are correctly identified and can be expressed as:
Sensitivity = $\frac{A}{A+C}$
Specificity are correct rejections which are correctly identified and can be evaluated using:
Specificity = $\frac{D}{B+D}$
The error rate are the misclassified instances in the whole dataset and can be determined as:
Error Rate = $\frac{B+C}{A+B+C+D}$
Accuracy is the total number of predictions that were correct. Thus it can be expressed as:
Accuracy = $\frac{A+D}{A+B+C+D}$

The values of accuracy and error rate for different classes estimated by existing method (SVM) and Proposed method (KNN-GA) can be tabulated as follows:

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 3: Calculation of Accuracy by SVM and KNN-GA

Classes	Existing method	Proposed Method
GH1	97.53	98.46
GH2	87.04	92.76
GH3	64.51	84.28
GH4	79.94	90.29

Table 4: Calculation of Error rate by SVM and KNN-GA

Classes	Existing method	Proposed Method
GH1	0.024	0.015
GH2	0.129	0.072
GH3	0.354	0.157
GH4	0.200	0.097

By comparing these values, it can be easily drawn that the accuracy of existing method is far better than the existing method for each individual classes. The error rate found in proposed method was significantly low as compared to existing method. So it can be concluded that proposed method is better than the existing method.

VII. CONCLUSION

The traditional methods such as K-NN, SVM, GA discussed for the classification process have certain advantages and disadvantages. The more optimized technique KNN-GA have been introduced for the classification of the learners. It was found that the proposed method (KNN-GA) is more accurate than the existing method (SVM) with less error rate. The main drawback of the proposed method is that it is bit time consuming.

REFERENCES

[1]Suguna,N. and Dr. Thanushkodi, K.; "An Improved k-Nearest Neighbor Classification UsingGenetic Algorithm";IJCSI International Journal of Computer Science Issues; Vol. 7, Issue 4, No 2, July 2010.
[2]Srivastava,Durgesh K, Bhambhu,Lekha; "Data Classification Using Support Vector Machine"; Journal of Theoretical and Applied Information Technology; Vol. 12,pg1-7;2005 – 2009.
[3] Gunavathi C., Premalatha K.; "Performance Analysis of Genetic Algorithm with kNN
and SVM for Feature Selection in TumorClassification";International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, Issue No:8, 2014
[4] Gil-Pita R. and Yao X.; "Using a Genetic Algorithm for Editing k-Nearest Neighbor Classifiers";Springer-Verlag Berlin Heidelberg,pp. 1141–1150, 2007.
[5] ThiHoai An Le, Le Hoai Minh, DinhTaoPham;"Feature Selection in Machine Learning- an exact penalty approach using a difference of Convex Function algorithm"; Springer; pp- 2-3; 2014.