International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Hybrid Approach for Integrating Genetic Algorithms with SVM for Classification and Modelling Higher Education Data

Kamiya Malviya, Prof. Anurag Jain

Bhopal, Madhya Pradesh, India

Abstract: ***Higher education institutions are hub of research and future development acting in a competitive scenario, with the basic goal to generate, gather and share knowledge. In this research work my objective is to explore data mining technique like classification, clustering on higher education data, my objective is to integrate genetic algorithm (GA) andsupport vector machines algorithm(SMO),for integration of two classifiers we use ensemble stacking, which is a fusion of classifiers. We present a generalize and powerful hybrid methodology of spectral clustering which originally operates on SMO and genetic algorithm classifiers, and further develop algorithms for classification on the basis of minimum attribute selecting and normalization of dataset. It could be concluded that the proposed GA-SMO classifier approach improves the classification accuracy and gives the better results, than other methods.***

Keywords: SMO, GA, Classification, Ensemble, Feature Extraction

1. Introduction

Educational Data mining (EDM) is an recent developing area, having growing method for studying the special types of data that come from educational system and using those methods to better understand students performance [1]. The classification of higher education data has become an increasingly challenging problem; many institutions do not have sufficient information to give guidance to students, therefore they are not able to give suitable advice to the students.We also observe that there is no perfect grouping of courses to recognize which type of course is most suitable to be offered to which type of student and Classification of this largest amount of data is time consuming and take excessive computational effort, which may not be for predicting the academics performance of students. For this, we develop an approach to pre-processing reducing the size of the training dataset, by removing noise points, outliers and insignificant points, which are not important for classification, Then we classifying the data by Sequential Minimal Optimization (SMO) algorithm is applied on the reduced dataset for optimize the support vector machine parameters, and optimize the results by genetic algorithm (GA). After we compare our work with the traditional SMO technique to show its improvement in terms of classification efficiency and other measure.

2. Methodology

2.1 Data Pre-Processing

Real data is often incomplete, inconsistent, and lacking in certain behaviours and is sometimes contain many errors. Data pre-processing prepares raw data for further processing [2]. Transformation of data includes dimensional reduction techniques like feature selection and feature extraction. Feature Selection is the method of finding the "best" subset of features from the initial 'N' features in the datasets; this reduces the dimensionality of feature sets, removes redundant, irrelevant data. It brings a speeding up a data mining algorithm, improving the data quality [3]. Feature Extraction defines a transformation from pattern space to feature space such that the new feature set gives both better separation of pattern classes and reduces dimensionality of datasets. Thus feature extraction is a kind of feature selection, it is a superset of feature selection; feature selection is a special case of feature extraction [5].

2.2 SMO (sequential minimal optimization)

A support vector machine (SVM), first introduced by Vapnik in 1995. SVMs are a set of supervised learning methods used for classification, regression and outliers detection in both linear and nonlinear data [10].For training SVMs we have three basic algorithms: Chunking, Osuna's algorithm, and SMO [7].

J. C. Platt proposes an algorithm for training support vector machines: Sequential Minimal Optimization (SMO). Training a SVM requires the solution of a very large quadratic programming (QP) optimization problem, SMO breaks large QP problem into a series of smallest possible QP problems, these small QP problems are solved systematically, that avoids using a time-consuming numerical QP optimization as an inner loop. The memory requirement for SMO is linear in the training setsize [7].This is simple, easy to implement, faster, and has better rising properties for difficult SVM problems than the standard SVM training algorithm.

Optimization problems are rapidly solved using SMO algorithm. Consider a binary classification problem with a dataset (x_1, y_1)... (x_n, y_n), where $x_i$is an input vector and y_i, belongs to {-1, +1}, is itscorresponding binary label. Support vector machine helps solve the binary form of quadratic programming problem as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j,$$

Subject to'

$$0 \leq \alpha_i \leq C, \quad \text{for } i = 1, 2, \ldots, n,$$
$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Where C is a Support Vector Machine hyper-parameter and K (x_i, x_j) is the kernel function [6]

2.3 Genetic algorithms

Genetic algorithms imitate the evolution of the living beings, described by GACharles Darwin [8]. GA is an optimization and robust search method based on the principles of genetics and natural selection. It is known as a subset of evolutionary algorithms that model biological processes which is influenced by the environmental factor to solve various numerical optimization problems. A state is evolved under the specified rules that maximises the fitness or minimizes the cost functions. GA allows such population to evolve. This population is composed of many individuals or called chromosomes. This is mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation; good strings are combined by the crossover to generate better offspring; string is altered by mutation, locally, to maintain the genetic diversity of a population of chromosomes from one generation to the next. The population is evaluated and tested for termination of the algorithm, in each generation. The three GA operators drive the population and are then re-evaluated, if the termination criterion is not satisfied. Until the termination criterion is reached, the GA cycle continues.

3. Proposed Methodology

3.1 Data Pre-Processing

The training data set is pre processed by filters, here we use attribute selection filter, which is supervised filter that can be used to select attribute row and column heading that is the numerical such as student enrol no, mobile no, DOB, and remove them. The empty value columns for each row are removed the mean obtained by each column is subtracted with each row of column entries. After this subtraction if the value obtained is between the range +1 and -1, the field is assigned a value '1' and if not, then a value '0' is assigned. Since the data is required to be given as an input to the genetic algorithm, the fields are converted to 1's and 0's. The genetic algorithm works best for binary attributes.

3.2 Fusion of two classifiers

In this section, we describe the proposed fusion of Genetic-SVM system. The aim of this system is to combining multiple classifiers to get the best accuracy. In WEKA the class for combining classifiers is called Stacking.

The first ensemble learning method is Stacking. It is combining process of multiple classifiers generated by different learning algorithms $L_1..L_n$ on a single dataset. In the first phase a set of base level classifiers C_1, C_2.. C_n is generated. In the second phase combining the base level classifier develops a Meta level classifier.

3.2.1SMO setup

In WEKA SVM classifier is called SMO. To implement our proposed approach, this research used the RBF kernel function for the SVM classifier since the RBF kernel function can analyse higher-dimensional data and requires that only two parameters, complexity parameter C and γ be defined.

3.2.2 Genetic set up

The first step in GAs is to define the encoding, we use a vector of (0 and 1) with length of 10 (the number of features) which 0and1 is for the omitted and selected features respectively. At first, randomly we generate 50 chromosomes as a population. We use Roulette Wheel Selection for the cross- over which have probability parameter 0.7 and also we apply Swap mutation which have probability parameter of 0.1.The choice of the fitness function is important because this basis that the Genetic evaluates the goodness of each candidate solution for designing our SVM classification system.

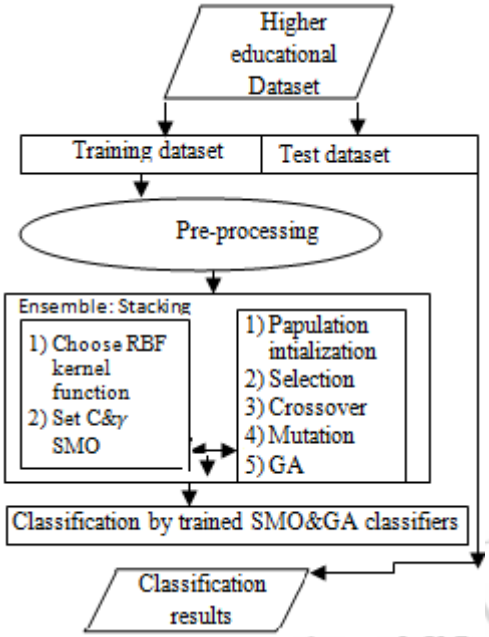
3.2.3. SMO Classification with genetic Algorithm

The procedure describing the proposed SVM classification system is as follows:

1. [Start] Generates randomly an initial population of size n (we take 100).
2. [Training SVM Classifier] In weka this classifier is SMO trained by training set with selected feature subset and value of parameters.
3. [Fitness]. For each chromosomes of the population, train $\frac{n(n-1)}{2}$SVM Classifier for computing fitness of each chromosome (subset of features).
4. [New population] Select individuals from population directly based on fitness values and regenerate new individuals from old ones.
   a. [Genetic Operation] apply Selection, Crossover, Mutation.
5. [End criteria] If the maximum number of iteration is not yet reached, we proceed with the next generation operation. The termination criteria are that the max generation number reached or the fitness function value does not improve during the last 10 generations return to step 2.
6. Select the best fitness as optimal feature subset

Apply the optimal feature to dataset and optimize the classification accuracy.

3.2 System Architecture



4. Dataset and Tool Description

A. Higher educational dataset

We are using datasets of Alpha College from rural area, and Bhopal, which is urban area from Madhya Pradesh state to apply data mining techniques and calculate results. From database, first attributes are required which is selected from these excel sheets. These attributes are as follows: Course, Branch, Genders, Category, Class, Date of Admission, and Minority/Non-Minority

B. Simulation Tool

WEKA i.e., Waikato Environment for Knowledge Analysis (WEKA) is used as a simulation tool that would allow researchers easy access to state-of the art techniques in machine learning [6]. Weka provides three options:

Weka Explorer: The Explorer has several panels, which provides access to the main components of the workbench. It has panels i.e. Pre-process, Classify, cluster, associate. Weka Experimenter: It provides comparison of weka algorithms in systematic way. Weka Knowledge Flow: It provides better representation to the Explorer as a graphical front end to weka's core algorithm.

5. Model Evaluations

A. Cross Validation

It is a technique for showing how the results of a statistical analysis will generalize to an independent data set. 10-fold cross validation is frequently used method.

B. Criteria for Evaluation

To estimate the performance of any model Accuracy, Sensitivity, Specificity, Kappa statistics and correctly classified Instance are employed as major criteria.

a) Accuracy- Accuracy means probability that the algorithms can correctly predict positive and negative examples. Accuracy=TP+TN/ (TP+ FN +TN+ FP)
b) Sensitivity- Sensitivity means probability that the algorithms can correctly predict positive examples. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant. High precision means that an algorithm returned substantially more relevant results Precision = TP/ (TP+FP)
c) Recall - Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Recall=TP/ (TP+FN)

C. Confusion Matrix

A confusion matrix that summarizes the number of instances predicted correctly incorrectly by a classification model.

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

6. Simulation Results

This section describes the experimental results obtained by applying the proposed algorithms on education dataset. In order to validate the prediction results of the comparison of the two classification (SVM, SVM + Genetic) techniques and the 10- fold crossover validation is used. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation.

Table 1: Performance of SMO and proposed SMO+GA approach

Algorithms	Classification accuracy	TP rate	FP rate	Precision	Recall	F-measure
SMO	30.21	0.985	0.999	0.97	0.93	0.947
Proposed SMO+GA	60.55	1	1	0.358	1	0.527

Table 2: Other Performance of SMO and proposed SMO+GA approach

Parameters	*Classifiers*	
	SMO	*Proposed SMO+GA*
Correctly classified instances	160	220
Incorrectly classified instances	385	325
Kappa statics	-0.0008	0
Mean square error	0.3607	0.3654
Root mean square error	0.4552	0.4288
Relative absolute error	99.45	100%
Root relative square error	107.34	100%

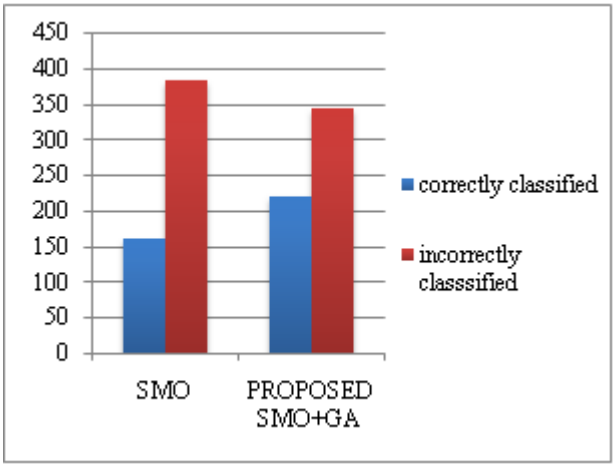


Figure 4: Compares the proposed SMO+GA approach with plain SMO classifier.

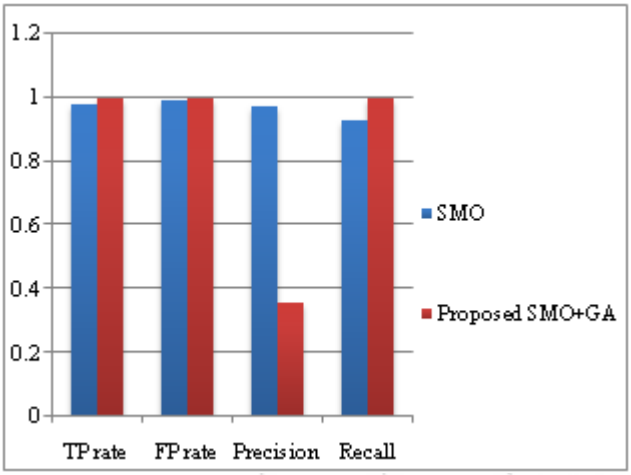


Figure 5: compares the proposed SMO+GA approach with plain SMO classifier

7. Conclusion

Nowadays higher learning institutions are facing problems regarding course marketing, different interests of different students from rural and urban areas. This work will help the institution management to improve the quality of education. Our proposed pre-processing procedure passes quality data to fusion of classifiers on training procedure and it results in increase in accuracy of classification as well other parameter also give better results. Hence, fusion of classifiers improves system ability as well as classification accuracy of classifier with respect to higher educational datasets. A comparison of the proposed algorithm results with Existing SMO approach demonstrates that the proposed method improves the classification accuracy rates. The SMO+GA method was applied to remove insignificant features and effectively find best parameter values.

References

[1] pujashrivastav "Uncovering Hidden Information Within Higher Education Students Data Using Data Mining Techniques" International Journal Of Research In Technology (Ijrt) PhD
[2] P. Kamavisdar, S. Saluja, S. Agrawal "A Survey on Image Classification Approaches and Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013
[3] G.RKumar, G.A.Ramachandra, K.Nagamani" An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets" IJARCSVolume4, Issue 2, February 2014
[4] Jihoon Yang and Vasant Honavar "Feature subset selection using Genetic Algorithm." IEEE Intelligent Systems, 1998.
[5] Lakshmi Devasena C "Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction "international Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 4, April 2014
[6] John C. Platt "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines "Microsoft Research jplatt@microsoft.com Technical ReportMSR-TR-98-14April21
[7] http://www.ro.feri.unimb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic%20algorithm/.ml
[8] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd, 443-491
[9] Vapnik, V.N. Statistical Learning Theory. John Wiley and Sons, New York, USA, 1998.