ISSN:0975-9646

Divyansh Jain et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6728-6731

Student Profile Assessor for University Prognosis Using Classification Algorithms

Divyansh Jain[#1], Ameya Shringi[*2]

[#]*School for Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*

[*] *School for Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*

***Abstract*— The number of international students seeking admissions in Masters/PhD programs in the United States of America has increased by 7.2% over the past year. [4] In this scenario, a comprehensive rating of universities often do not serve as a reliable parameter to help the prospective students identify a suitable University. While granting admissions, many parameters such as the student's performance in various sections of GRE, TOEFL as well as their academic background and extracurricular activities need to be taken into consideration while selecting universities. In this paper, we propose a solution to the aforementioned problem by modelling the admission criteria based on the profiles of accepted students. This model serves to predict a suitable university for a student based on their profile input.**

***Keywords*—Classification, Support Vector Machine, Random forest, Neural Network.**

I. INTRODUCTION

Attributing to a large number of parameters used as an admission criteria for graduate programs, prospective students rely on university rankings or consultancies for simplifying admission process. The flaw in using University ranks as the decision parameter is that, the methodology used to derive these ranks are not correlated to the admission criteria, but the general standing of the university. Moreover, a myriad of consultancies offering counselling services to prospective students charge a hefty amount and still fall short.

Predictive modelling is the process by which a model is created or chosen to try to best predict the probability of an outcome [1]. The classifiers used in these predictors use the data corresponding to the accepted students of a university as a training set. Some of the common classifying techniques include neural networks, Support Vector Machine (SVM), Naïve Bayes and Random Forest.

In this paper, we propose to design a system with the aid of predictive modelling to recommend apt universities to students. The training phase consists of classification of the accepted students using three different algorithms –neural network, SVM and random forest – to generate three different models, which are then evaluated based on their accuracy. In the next step, the best model is chosen to predict the universities for potential candidates using standardized test scores of GRE, TOEFL, IELTS combined with their extracurricular activities and work experience.

II. BACKGROUND AND RELATED WORK

Previous work in this domain, included an unsupervised two-step clustering of universities based on a student competency score computed by a formula which assigned weights to different parameters such as GRE Total score, TOEFL Score as well academic and extracurricular activities [3]. The clusters formed were assigned class labels as numeric ids and classification was performed using ID3 based on the potential candidates competency score.

Student's profile competency rate
= GRE/GMAT/IELTS SCORE+ TOEFL SCORE+
10*(RND+TE+NGO'S +SPORTS +CC +OTHER)
+ 15*(GPA)

In this paper, a supervised classification of universities is proposed to generate a predictive model from training data comprising of parameters such as University Acceptance/Rejection status, GRE section scores, TOEFL/IELTS, CGPA, Work Experience, Internships as well as Research papers published in International and National Journals. The benefit of this approach is that none of these attributes are pre-assigned any weightage i.e. the weights are computed during the modelling process. Moreover, separate GRE section scores (Quantitative, Verbal and AWA) are accounted for since universities have specific section score criteria while reviewing a candidate's profile. For instance, an engineering Masters/PhD applicant would be expected to score comparatively higher in the quantitative section than a student from some other discipline.

The accuracy is optimized as the three models are compared and the one with most accurate predictive model is further used to predict the university for potential candidate.

III. PROPOSED SYSTEM

The proposed system is divided into three sections-

A. Data Cleaning and Normalization

The training dataset[5,6] of students comprising of numeric (GRE section scores, TOEFL/IELTS, CGP, Number of Research papers in International Journals, Number of Research papers in a National Journal) and binary (University Acceptance/Rejection Status, Work Experience, Internships) attributes has to be cleaned and standardized before performing modelling or applying any classification technique. The binary attributes are converted to numeric form (1- Yes, 0- No) as in **Table 1**.

Table 1. Binary Conversion of attributes to numeric format

University	MIT					ASU
University Status	Accept					Accept
Quantitative Score	170					166
Verbal Score	167					157
AWA	5.5					3.5
TOEFL	117					--
IELTS	--					6
CGPA	9.3					8.3
Work-Experience	No					No
Internships	Yes					Yes
Research papers in International Journals	3					0
Research papers in National Journals	1					1

Since, either TOEFL or IELTS scores are accepted by universities, they must be converted to a standardized value on a common scale.[]

Each value in the dataset is then normalized using the formula:

$$\text{Normalized Value} = \frac{\text{Current Value} - \text{Mean value of Attribute}}{\text{Standard Deviation of the Attribute}}$$

B. Predictive Modelling Using Classification Techniques:

1) Support Vector Machines (SVM): Support Vector Machine (SVM) performs classification by determining hyperplanes in a multidimensional space by assigning different entries to their respective class labels.[2] The dataset in this case calls for a Multiclass SVM as the universities are modelled as class labels. The solution is provided by reducing the single multiclass problem into a multiple binary classification problem. The Classification is done using the one-versus-all strategy meaning that the binary classifier producing the highest output function decides the class of the entry.

In order to classify data that is not linearly separable, a non-negative slack variable ξ_i is introduced to modify the equation:

$$y_i(w.x_i - b) \geq 1 - \xi_i \qquad 1 \leq i \leq n$$

Where,

y_i is the class to which x_i belongs

w is normal to the hyperplane.

$\frac{b}{\|w\|}$ is the perpendicular distance from the hyperplane to the origin

ξ_i : Degree of misclassification of data x_i

While using soft margin SVM, data points that are on the incorrect side of the margin carry a penalty that increases with the distance from it. In order to reduce the misclassifications, the objective function is minimized as follows

$$\underset{w,\xi,b}{\arg\min} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \right\}$$

Where the parameter C controls the trade-off between the slack variable penalty and the size of the margin.

Algorithm

- Create H such that $H_{ij} = y_i y_j x_i . x_j$
- Select C to decide significant misclassifications
- Compute α such that

  $$\max \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha$$

  subjected to

  $$0 \leq \alpha \leq C\ \Upsilon \text{ and } \sum_{i=1}^{L} \alpha_i y_i = 0$$

- Calculate $w = \sum_{i=1}^{L} \alpha_i y_i x_i$
- Determine the Support Vector set S by finding indices such that $0 \leq \alpha \leq C$
- Calculate $b = \frac{1}{N_s}\sum (y_s - \sum \alpha_m y_m x_m . x_s)$
- Every new point x' is classified by $y' = \text{sgn}(w.x' + b)$

2) Random Forest: Random forests are an assemblage of learning algorithms that classifies based on the aggregation of the output from many independent decision trees. Instead of growing a single extremely precise decision tree, Random Forest relies on generation of decision trees by the training data. The training data is generated from the original dataset by randomly sampling the cases present in the dataset with replacement i.e. the original dataset is distorted to a small extent to ensure that each decision trees give distinctive results.

While generating a decision tree from the training data, a small number of attributes are selected at random and the decision tree classifies the entry based on those selected attributes. For instance, some of these trees may have been grown from samples that identified the GRE verbal score as a more important feature in comparison to the GRE quantitative score. Other trees may find that a combination of quantitative score and CGPA are more decisive parameters. Other trees may find completely different features to be relevant. The errors generated from these numerous decision trees will be compensated when the output from them are aggregated leading to a more accurate prediction.

The error is estimated internally as each tree is constructed from a partial dataset i.e. only a part of the original dataset. The classes of the cases that were omitted during the construction of that particular tree are then predicted and the class chosen by the trees most frequently for that particular case in selected. The proportion of times that the class is incorrectly classified is averaged over all cases to determine the 'Out-of-bag' error.

In random forest, a factor called Gini impurity is evaluated. Gini impurity defines the number of times a randomly chosen entry would be incorrectly classified if it were assigned a random label based on the distribution of labels in the subset.

$$I_G(f) = \sum_{i=1}^{m} f_i(1-f_i)$$

where

f_i : fraction of items labelled with value i

Algorithm

- Consider a training set, D of which d tuples are given.
- For generating m decision trees, at each iteration sample a training set D_i of d tuples with replacement from D.
- From the total available attributes (t), select a much smaller number of attributes (k), to split the tree considering these randomly selected k attributes.
- Allow the trees to grow to maximum size and do not prune.
- While classifying a new entry, pool votes from all the m decision trees. The class with maximum votes is assigned as the new class for the entry.

3) *Multilayer Perceptron:* Multilayer Perceptron comprises of an atomic unit that primarily uses a non-linear function as an activation function to map a continuous function of real number to some output interval. The learning algorithm is modelled on a directed weighted graph, wherein the weights are adjusted to reduce the error between the output and the predication of the network. The commonly used activation function are

$$y(x_i) = \tanh(x_i) \text{ and } y(x_i) = 1/(1+e^{-x_i})$$

Where,

x_i is the input vector

y is the predicted variable

The network comprises of three type of layer the input layer, the hidden layer and the output layer. The weights of the interconnection are adjusted to minimise the error. This is done using gradient decent and backpropogation algorithm. The error in an output node is given by:

$$e_j(n) = o_j(n)-y_j(n)$$

Where,

e is the error of output node j of n^{th} data point

o is the target value

y is the value produced by perceptron

The error across all the output node is given by

$$E(n) = 1/2 \, \Sigma \, e_j^2(n)$$

Where,

E is the aggregate output across all nodes and the change in weight using gradient decent is given by

$$\Delta w_{ji}(n) = -\eta \, \partial E(n)/\partial v_j(n) * y_i(n)$$

Where,

y_i the output of the previous neuron

η is the learning rate

Algorithm

- Let the training set comprise of n training example
- Calculate the error for all n training cases
- Calculate the mean square error for the iteration
- Adjust the weight in the backward pass layer wise using the backpropogation algorithm
- Reiterate until the global minima is obtained or maximum iterations have been reached.

C. Comparison of Predictive Models:

1) *Performance of SVM:* The experiments were performed using a radial basis kernel of degree 3. To determine the best classification parameters a grid search was performed with gamma value varying from 0.001 to 0.1 and the misclassification cost value varying from 5 to 500. The best results were observed as on gamma value of 0.001 and misclassification cost of 83.

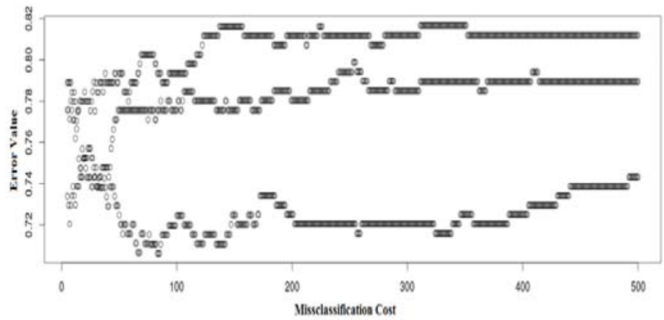

Fig. 1 scatter plot of error against misclassification cost.

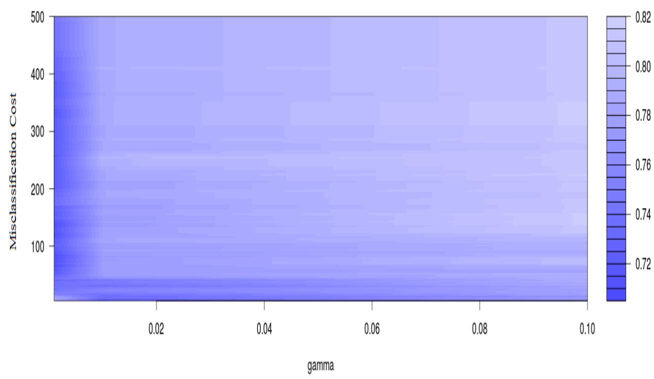

Fig. 2 Performance of SVM as a measure of Cost and Gamma

2) *Performance of Random Forest*: The parameter used to tune the random forest was the number of trees generated. The number of trees were varied from 10 to 500 and the best result was obtained for 80 trees. The error measure used for classification was out of bag error.

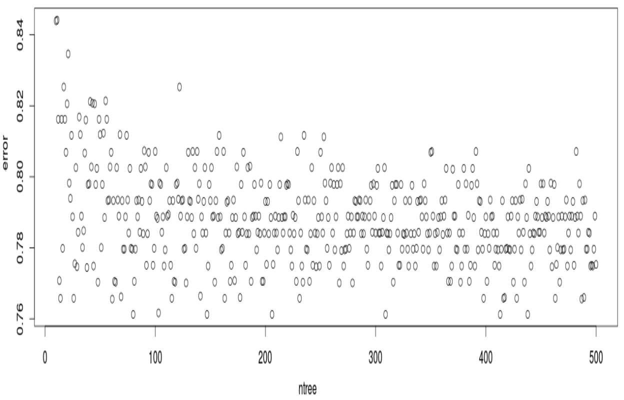

Fig. 3 Scatter Plot of Error as a measure of the number of tree

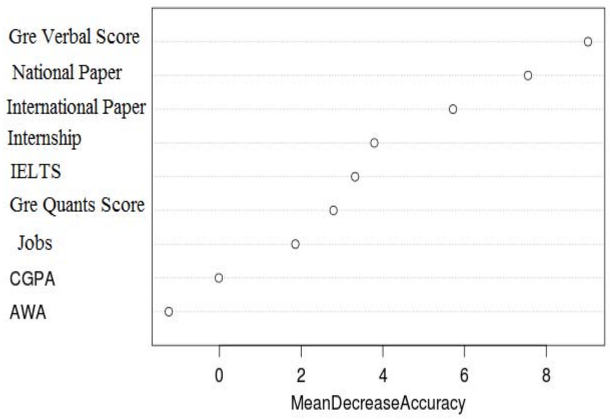

Fig. 4 Measure of Accuracy against classification parameters

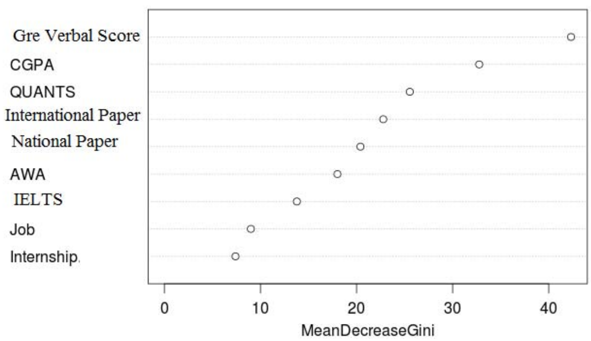

Fig. 5 Measure of Gini against the classification parameter

3) *Performance of Multilayer Perceptron:* The network uses a single hidden layer wherein the number of hidden units are varied. The decay parameter was varied from 0.01 to 0.1, to determine the optimal parameters.

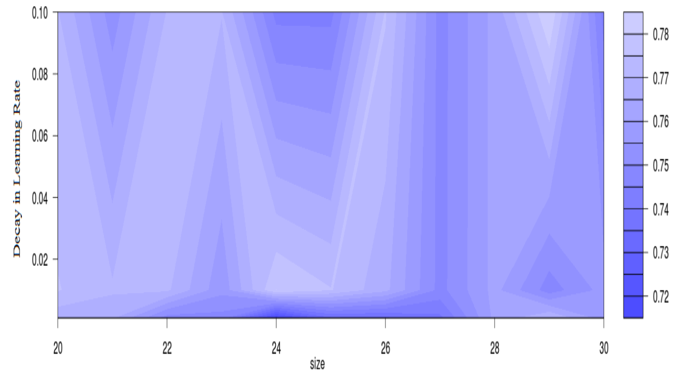

Fig. 6 Performance of Neural Network against decay and number of hidden units

IV. RESULTS AND INFERENCE

The three predictive models and classification are based on supervised learning techniques where each classification approach decides the order of importance of parameters. This generated model in turns classifies the new case and the result is a recommended university for the prospective student.

V. FUTURE SCOPE

The Job Experience is often evaluated in terms of:

- The duration of the Job
- The position held
- Domain of the Job
- Company of Employment

Apart from these, similar factors while evaluating the internships as well as other extracurricular activities must be considered. Also, factors pertaining the evaluation of research papers and their corresponding impact index shall improve the overall recommendation

REFERENCES

[1] Geisser, Seymour (1993). Predictive Inference: An Introduction. New York: Chapman & Hall. p. [page needed]. ISBN 0-412-03471-9

[2] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.

[3] Chavan, T., Barretto, A., & Chavan, T. Data Evaluation and Knowledge Extraction for Students Using Clustering and Classification.

[4] (2013) INTERNATIONAL STUDENTS FLOCKING TO U.S. SCHOOLS[Online].Available:http://www.usatoday.com/story/news/nation/2013/11/11/international-students-and-study-abroad/3442733/

[5] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.