Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 675806, 7 pages
http://dx.doi.org/10.1155/2014/675806

Research Article

Decision Tree Classification Model for Popularity Forecast of Chinese Colleges

Xiangxiang Zeng, Sisi Yuan, You Li, and Quan Zou

Department of Computer Science, Xiamen University, Xiamen, Fujian 361005, China

Correspondence should be addressed to Quan Zou; zouquan@xmu.edu.cn

Received 25 December 2013; Accepted 5 April 2014; Published 24 April 2014

Academic Editor: Jose L. Gracia

Prospective students generally select their preferred college on the basis of popularity. Thus, this study uses survey data to build decision tree models for forecasting the popularity of a number of Chinese colleges in each district. We first extract a feature called "popularity change ratio" from existing data and then use a simplified but efficient algorithm based on "gain ratio" for decision tree construction. The final model is evaluated using common evaluation methods. This research is the first of its type in the educational field and represents a novel use of decision tree models with time series attributes for forecasting the popularity of Chinese colleges. Experimental analyses demonstrated encouraging results, proving the practical viability of the approach.

1. Introduction

College selection is a complicated decision-making activity for prospective college students and their parents. Such decision is typically made on the basis of subjective judgment or experience of the decision makers involved. The diverse information about colleges vis-a-vis the narrow expertise of students presents a challenging situation in which college selection cannot be fully justified. Moreover, college rankings vary every year and are beyond personal analyses.

With the vast amount of previous data, data mining appears to be well-suited technique that could provide an objective approach. Data mining, which is the process of exploring data to discover unknown patterns, is an essential part of the overall knowledge discovery in databases [1, 2]. This process can determine underlying patterns among historical cases and deliver knowledge to support decision making.

College popularity is the state of being applied by a number of students. The more the students applying for the college are, the higher the popularity of the college will be. Apparently, the number of students who choose the colleges to apply for varies every year, causing the increase or decrease in college popularity.

In this work, college popularity prediction is considered a time series forecast problem because the information of students accepted for enrollment into colleges is cumulated through consecutive years (from 2005 to 2012). A time series is a sequence of regularly sampled quantities from an observed system. A time series is useful in discovering and studying a system's behaviors, such as periodicity and regularity. A reliable time series prediction method would enable researchers to accurately model a system and forecast its behaviors. A great number of prediction methods in time or frequency domain have been proposed since the 1970s. Auto regressive (AR) model [3], AR moving average model [4], and AR conditional heteroskedasticity model [5] are very popular algorithms. Recent prediction approaches include wavelet networks [6] and hierarchical Bayesian approach.

A decision tree represents a tree-structured classifier that performs a split test in its internal node and predicts a target class of an example in its leaf node. With their simplicity and transparency, decision trees are widely used in data mining [7, 8]. In this work, we employ a decision tree algorithm in the prediction problem with a large number of colleges and corresponding average passing score, which is simply referred to as score in this study. We propose a simplified but efficient decision tree data-mining algorithm based on entropy

splitting criterion combined with prepruning to limit the tree growth. The scores are collected during the period from 2005 to 2012 from six representative provinces, namely, Anhui (eastern China), Heilongjiang (northern China), Xinjiang (western China), Yunnan (southern China), and Hebei and Henan (mid-China). For each province, the actual decision tree model is built by applying our algorithm to the scores from 2005 to 2011. Then, the data from 2006 to 2011 are employed in the decision tree to forecast the college popularity in 2012. Finally, a confusion matrix is used to evaluate the classifier. The experiments performed using different real datasets reveal satisfactory results in comparison with previous classification approaches.

The rest of the paper is organized as follows. Section 2 presents the proposed decision tree algorithm, including splitting criterion and decision tree pruning. Section 3 evaluates our algorithm using confusion matrix and receiver operating characteristic (ROC) curve. Section 4 presents and analyzes the experimental details. Finally, Section 5 presents the conclusion.

2. Proposed Algorithm

2.1. Data Preprocessing. Before we present the data used in this work, we briefly introduce the admission process of Chinese colleges as follows.

Step 1. The candidate students are ranked in a queue descendingly by their scores.

Step 2. The queue header is picked to fulfill his or her application if the colleges are not already in full recruitment.

Step 3. Delete the current queue header and repeat Step 2.

The original data are collected from Sina Education Channel (http://edu.sina.com.cn/); the data include the following elements:

(i) province refers to the location of students (colleges implement different enrollment policy among provinces);

(ii) type is the kind of colleges (arts or science);

(iii) year is the year of enrollment;

(iv) college name refers to the college that recruits students;

(v) score is the passing score of the college; if a student gets a score higher than this passing score and the student chooses the college to be his/her desired college, then the student will be enrolled by the college.

For example, "Hebei, science, 2012, Xiamen University, 692" means that, in 2012, the score of Xiamen University given by students from Hebei Province who majored in science was 692.

Our objective is difficult to predict directly because the complexity of college entrance examinations varies every year. To eliminate such disparity, college score ranking is transformed to amend the original data. For example, the score and ranking of "Hebei, science, 2005 to 2011, Xiamen University" for each year are listed in Table 1.

To achieve further normalization, popularity change ratio (PcR) is used to reduce inherent distinction. Consider

$$p_t = \left(\ln R_t - \ln \overline{R}\right) \times 100, \tag{1}$$

where the notation R_t denotes the score ranking of a college in a province at year t and $\overline{R}$ denotes the previous average ranking. In building decision trees, scores and PcRs were used as attributes and popularity was used as target class. For the target class, the value "1" indicates an increase in popularity, whereas "0" indicates a decline in popularity.

2.2. Splitting Criterion. To evaluate the classification capability of attributes, we utilize the information gain ratio of attributes, as proposed by Quinlan [11].

To define this metric, we first define the information entropy that measures the degree of impurity of a certain labeled dataset. For a given dataset S, with n target classes $c_1, c_2, \ldots, c_n$, we define information entropy info(S) as

$$\text{info}\,(S) = -\prod_{i=1}^{n} \frac{C_i}{|S|} \times \log_2\left(\frac{C_i}{|S|}\right), \tag{2}$$

where C_i is the subdataset whose samples have the same target class c_i.

2.2.1. Information Gain. Assume that S is a training sample set. S can be partitioned into $\{S_1, S_2, \ldots, S_n\}$ according to the $\{n\}$ different values of attribute X, that is, in each subset the samples have the same value of X; the expected information requirement can be defined as the weighted sum over the subsets, as expressed in (3). Consider

$$\text{info}_X\,(S) = \prod_{i=1}^{n} \frac{S_i}{|S|} \times \text{info}\,(S_i)\,. \tag{3}$$

The quantity

$$\text{gain}\,(X) = \text{info}\,(S) - \text{info}_X\,(S) \tag{4}$$

measures the information, which is gained by partitioning S in accordance with the test X.

2.2.2. Information Gain Ratio. According to the definition of info(S),

$$\text{split info}\,(X) = -\prod_{i=1}^{n} \frac{S_i}{|S|} \times \log_2\left(\frac{S_i}{|S|}\right) \tag{5}$$

represents the potential information generated by dividing S into n subsets, and the information gain measures the information relevant to classification that arises from the same division. Meanwhile,

$$\text{gain ratio}\,(X) = \frac{\text{gain}\,(X)}{\text{split info}\,(X)} \tag{6}$$

expresses the proportion of information generated by the split, which is useful for classification [11].

Table 1: Scores, ranking, and PcR of the instance.

Year	2005	2006	2007	2008	2009	2010	2011
Score	604	618	636	603	628	622	643
Ranking	23	26	26	32	29	26	27
PcR	−16	−3	−3	16	7	−3	0

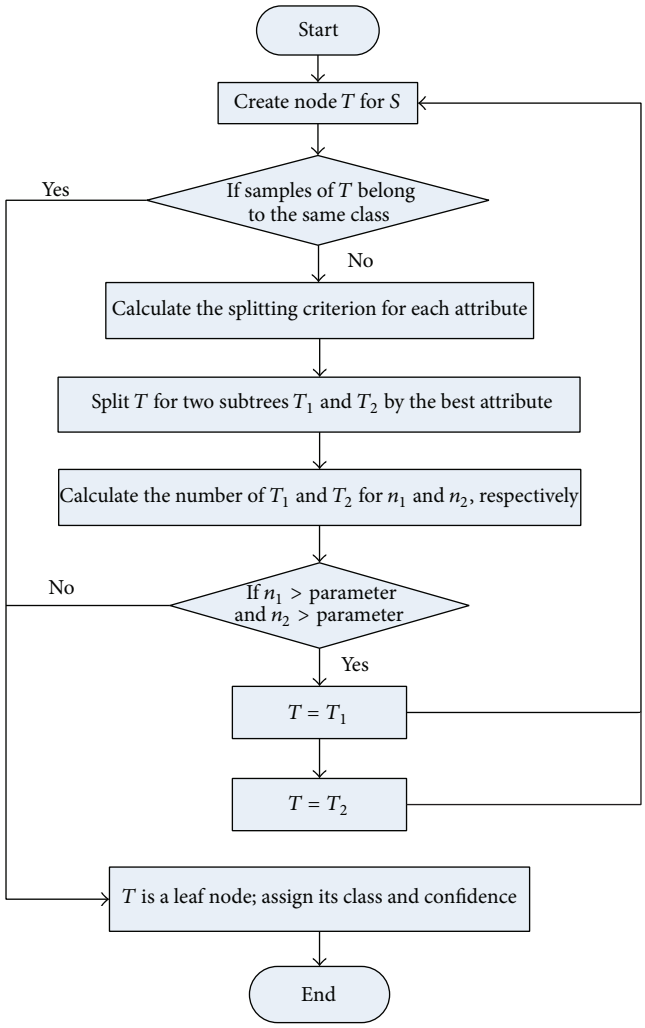

Figure 1: Decision tree construction flow path.

2.3. Decision Tree Construction. Let N denote the root node, which represents the entire dataset. For every value x of an attribute A, N is partitioned into two parts: one contains the samples whose value of A is smaller or equal to x and the other consists of the rest. By using (6), gain ratio (x) is obtained, where n is 2. Among all the gain ratios, the maximum is labeled as the gain ratio of attribute A, and the attribute with the maximum gain ratio is regarded as the best attribute. N, which is split by the best attribute, is divided into two subnodes, which continue splitting as N until they meet the requirements of a leaf node. The generated decision tree is a binary tree with two target classes.

If S is the current sample dataset, the decision tree construction flow path is as shown in Figure 1.

2.4. Decision Tree Pruning. When a system is trained by the training dataset, its efficiency with respect to instances outside the training dataset is an important issue. If a system accurately memorizes the training samples, it may fail miserably when provided with similar but slightly different inputs. In real-life classification tasks, the target class of samples in the training dataset generally cannot be expressed simply by the attribute values. Such case could happen either because the attribute values contain errors or because the attributes cannot collectively provide sufficient information to classify a new instance. In these circumstances, the tree might model the idiosyncrasies of the training dataset rather than a structure, which is useful for classifying unseen instances.

Two methods are used to cope with this problem. One is a heuristic method called stopping criterion [11], which determines whether a multiclass set of training objects should be divided further by evaluating its features, such as size, or by statistical significance tests. The other approach is to allow the tree to grow without constraints, followed by the removal of unimportant or unsubstantiated portions by pruning [9, 10].

The former method, which is also called "prepruning," is adopted in this study. A parameter is used to limit the growth of the decision tree, that is, the minimum object number of the subtree of the current node. The constraint should be satisfied until the tree stops growing.

2.5. Algorithm Description. The algorithm is shown in Algorithm 1, where n denotes the number of provinces experimented.

3. Performance Evaluation Measures

The output of a classification model is generally the counts of correct and incorrect instances or the counts with their confidence (for probabilistic decision tree). Table 2 shows the confusion matrix of a two-class (positive and negative) classifier.

Numerous evaluation measures are used for evaluating classifier performance. In our experiments, we elucidate two commonly used measures by using the elements of the confusion matrix.

3.1. Classification Accuracy. Classification accuracy (Acc) is the most frequently used measure for evaluating classifier performance. This measure correctly predicts instances against the total number as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\%. \quad (7)$$

3.2. Area under ROC Curve (AUC). However, most classifiers (including decision trees [11, 13]) could produce the probability estimations or the "confidence" of the target class prediction. Unfortunately, Acc completely ignores this information. Thus, Acc cannot sufficiently evaluate probabilistic classifiers. Another common evaluation measure is ROC curve [14], which is a simple graph that plots the relationship between

```
procedure PREDICT POPULARITY (int n)
    i ← 0
    while i < n do
        number ← college number of the province
        for (k = 0; k < number; k++) do
            data[k][9]← scores of 2005 to 2011, PcRs of 2005 to 2011
            para ← experimental parameter
            tree = Create_tree(data, para)
        end for
        for (k = 0; k < number; k++) do
            test[k][9]← scores of 2006 to 2012, PcRs of 2006 to 2012
            predict = Predict(tree, test)
        end for
        return predict
    end while
end procedure
```

Algorithm 1: Using decision tree to predict Chinese colleges' popularity.

Table 2: Confusion matrix representation.

	Predicted positives	Predicted negatives
Real positives	TP	FN
Real negatives	FP	TN

the false positive rate (x-axis) and true positive rate (y-axis) for different available cut-points. The two metrics can be defined as follows:

$$\text{false positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$
$$\text{true positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (8)$$

In this study, ROC curve is generated over real target class and its probability of being positive is based on testing records through IBM SPSS Statistics 21. We can explicitly obtain the AUC for evaluating decision trees. An area of 1 represents a perfect test, whereas an area of 0.5 represents a worthless test. Therefore, a desirable algorithm with a high true positive rate and a low false positive rate should have an AUC value closer to 1.

4. Experiments

4.1. Parameter Value. Numerous experiments are conducted with different parameter values. Comparative analyses reveal that different parameters significantly influence the accuracy of our decision tree. For example, for Hebei Province, where Beijing is located, the accuracy of the decision tree changes when the parameter value changes. Figure 2 shows the relationship between parameter value and accuracy.

To achieve an accurate prediction result, the parameter value is set to 10. Experiments reveal that the parameter value also produces satisfying results for other provinces.

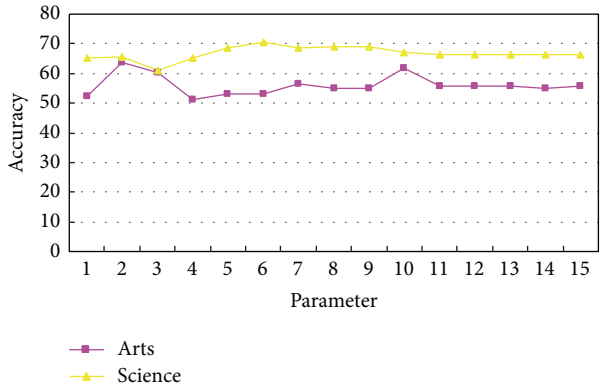

Figure 2: Relationship between parameter value and accuracy of Hebei Province.

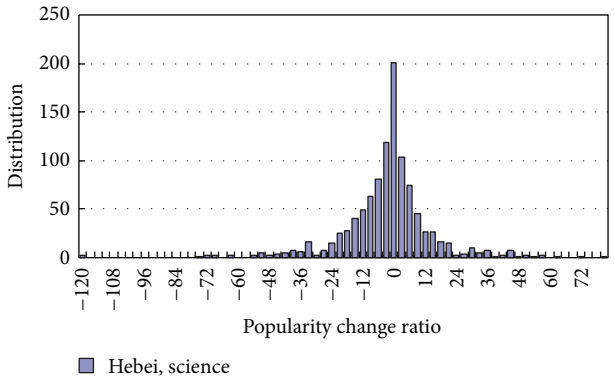

Figure 3: Distribution of PcR.

4.2. Analysis of One Province for Experimental Details. In this section, we consider "science-" type colleges in Hebei Province to gain experimental details.

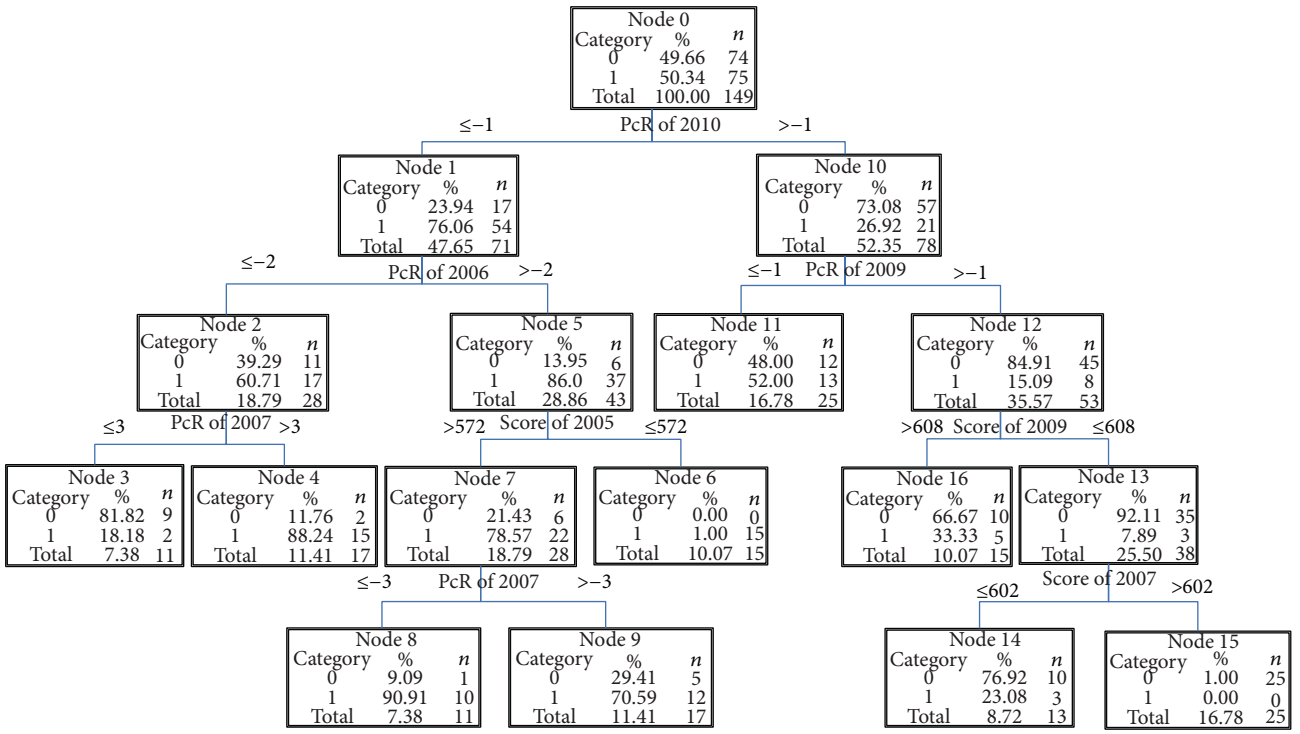

Figure 4: Decision tree of Hebei Province typed in "science."

4.2.1. PcR Distribution. In time series forecast algorithms, PcR fluctuation allows a normal distribution. Figure 3 is plotted with PcR as the horizontal axis and frequency distribution as the vertical axis. The frequency distribution of the PcR approximately accords with normal distribution.

4.2.2. Decision Tree. We use the dataset from 2005 to 2011, including original scores and PcRs, to build a decision tree with the use of the aforementioned algorithm. A leaf node may contain at least one object because prepruning stipulates the minimum objects of subnodes. We set the class of a leaf node in terms of its major component and its confidence. The actual decision tree is shown in Figure 4.

4.2.3. Decision Tree Evaluation. In this section, we apply the generated tree on the dataset of "science-" type colleges in Hebei from 2006 to 2011 to predict the popularity for 2012. The confusion matrix values are shown in Table 3, where positive and negative mean "1" (popularity rises) and "0" (popularity declines), respectively.

According to Table 3, we can obtain the Acc of the decision tree by using (7). Consider

$$\text{Acc} = \frac{53 + 47}{53 + 25 + 24 + 47} \times 100\% = 67.11\%. \quad (9)$$

The evaluation measure shows that the proposed classifier achieves a satisfying prediction result.

The decision tree is a probabilistic classifier; thus, a leaf node has its class and corresponding confidence, which are considered as its real target class and probability of being positive for ROC experiment, respectively. The ROC curve is shown in Figure 5.

Table 3: Experimental confusion matrix.

	Predicted positives	Predicted negatives
Real positives	53	24
Real negatives	25	47

AUC can be directly obtained. In this experiment, the value of AUC is 0.693, suggesting that the decision tree is considerably effective.

4.2.4. Experiments on Previous Classification Approaches. Two previous classification approaches, namely, Naive Bayes and SVM, are used to model the classifier generated over colleges ranked by Weka. The experimental results (Figure 6) show that our algorithm is a more effective approach in comparison with previous methods and has practical viability for forecasting the popularity of Chinese colleges.

4.3. Overall Result. To show that the proposed algorithm is not specially designed to predict a particular pattern, we use the data from 2005 to 2010 to build decision trees and then predict the popularity for 2011. Experimental results show that the algorithm works well on other datasets. Table 4 shows the overall results of the experiments.

According to Table 4, almost all the Accs and AUCs of "science-" type colleges are greater than those of "arts-" type

Table 4: Overall experimental results.

Province, type	Acc of 2012 (%)	AUG of 2012 (%)	Acc of 2011	AUG of 2011
Anhui, science	70.51	0.767	75.64	0.764
Anhui, arts	58.41	0.633	72.57	0.734
Hebei, science	67.11	0.693	73.83	0.787
Hebei, arts	61.95	0.635	66.37	0.664
Henan, science	73.75	0.790	75.63	0.777
Henan, arts	73.45	0.802	63.72	0.628
Heilongjiang, science	60.16	0.637	59.38	0.610
Heilongjiang, arts	67.01	0.597	56.70	0.637
Xinjiang, science	60.66	0.657	63.93	0.642
Xinjiang, arts	53.52	0.575	60.56	0.662
Yunnan, science	69.44	0.730	64.58	0.760
Yunnan, arts	57.89	0.612	63.16	0.654
Average	64.49	0.677	66.34	0.693

colleges. In the original data, "science" colleges outnumber "arts" colleges. Therefore, we assume that a greater number of training samples correspond to better decision trees for test instances.

In the case of "Xinjiang, science, 2012," the measured values are only 53.52% and 0.575. This result is attributed to the following reasons. First, Xinjiang Province only has 72 colleges available for modeling the decision tree, and this number is not sufficient to predict new instances. Second, Xinjiang Province is a minority municipality, such that its enrollment policy differs from other provinces.

The overall results are satisfactory, with an average Acc of 65.42% and an AUC of 0.685. Hence, the prediction tool improves the efficiency and effectiveness of the application process. In China, every prospective undergraduate can apply for at most five colleges. Therefore, our prediction is useful for the students to make decisions. The classifier aims to filter out the popularity-risen candidate colleges and forecast such colleges whose popularity may decrease in current year, so that the prospective students can focus on the most promising colleges, thereby allowing them to make a better selection job, such as choosing a low-popularity college that has a relatively better ranking.

5. Conclusion

In this paper, we present an efficient classification model that uses decision tree for forecasting the popularity of Chinese colleges. Experimental results show that the classifier is applicable to different patterns. Although our work performs a broad search to build decision trees and our experimental results are encouraging, analyzing other relational datasets or studying other classification methods is recommended to achieve better experimental results in future works.

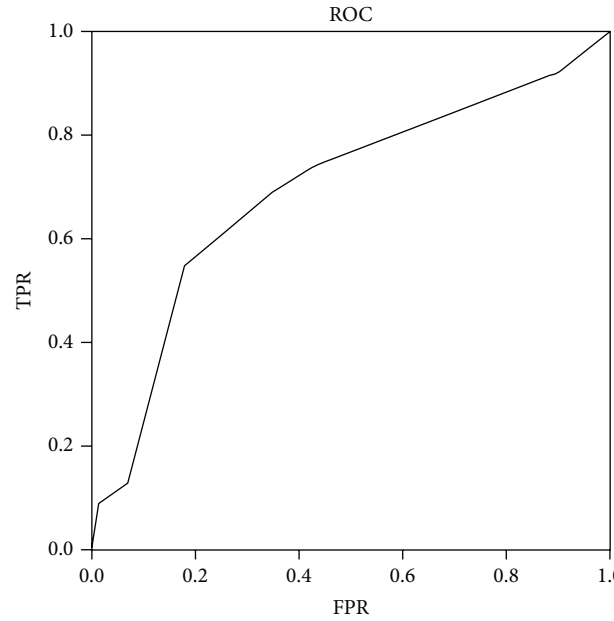

Figure 5: Experimental ROC curve.

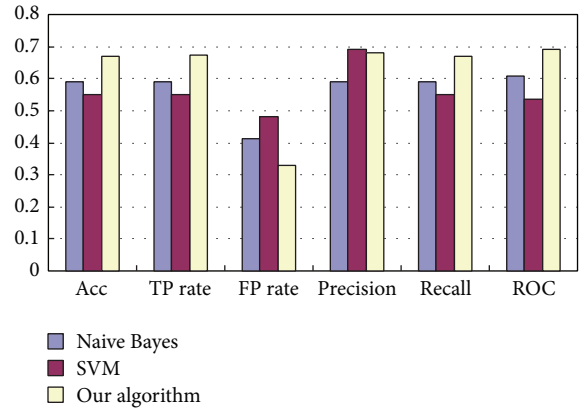

Figure 6: Evaluation measures of different classification approaches.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (60971085 and 61272071), the Base Research Project of Shenzhen Bureau of Science, Technology, and Information (JC201006030858A), and the Major Program of the National Social Science Foundation of China (Grant no. 13&ZD148).

References

[1] A. F. Mashat, M. Fouad, P. Yu, and T. Gharib, "A decision tree classification model for university admission system," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 10, pp. 17–21, 2012.

[2] J. Sun and H. Li, "Data mining method for listed companies' financial distress prediction," *Knowledge-Based Systems*, vol. 21, no. 1, pp. 1–5, 2008.

[3] H. Akaike, "Fitting autoregressive models for prediction," *Annals of the Institute of Statistical Mathematics*, vol. 21, pp. 243–247, 1969.

[4] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American Statistical Association*, vol. 65, pp. 1509–1526, 1970.

[5] R. Engle and V. Ng, "Measuring and testing the impact of news on volatility," *The Journal of Finance*, vol. 48, no. 5, pp. 1749–1778, 1993.

[6] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 889–898, 1992.

[7] L.-M. Wang, X.-L. Li, C.-H. Cao, and S.-M. Yuan, "Combining decision tree and Naive Bayes for classification," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 511–515, 2006.

[8] M. J. Aitkenhead, "A co-evolving decision tree classification method," *Expert Systems with Applications*, vol. 34, no. 1, pp. 18–25, 2008.

[9] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using the minimum description length principle," *Information and Computation*, vol. 80, no. 3, pp. 227–248, 1989.

[10] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.

[11] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, vol. 1, Morgan kaufmann, 1993.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth amd Brooks, Monterey, Calif, USA, 1984.

[13] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, vol. 3, Wiley, New York, NY, USA, 1973.

[14] J. Hanley, "Characteristic (ROC) curvel," *Radiology*, vol. 743, pp. 29–36, 1982.