Applied Mechanics and Materials Vols. 48-49 (2011) pp 1116-1121
Online available since 2011/Feb/02 at www.scientific.net

doi:10.4028/www.scientific.net/AMM.48-49.1116

Stock price's prediction with decision tree

Tun LI[1,a], Gongshen LIU[2,b]

[1]School of information Security Engineering, Shanghai JiaoTong University, Shnaghai, 200240
[2]School of information Security Engineering, Shanghai JiaoTong University, Shnaghai, 200240
[a]tun.lee@gmail.com, [b]lgshen@sjtu.edu.cn

Keywords: J48 decision tree; dynamic-constructed tree; average error distance

Abstract: Establishment of one process and some ameliorations of decision tree's algorithm in order to predict the second day's price change. The experiment builds a J48 tree, which is comfortable with continuous attributes, based on 10 years historical stock prices. After careful selection and preprocessing of financial data, high prediction accuracy is obtained. An introduction of dynamic-constructed tree reduces tree's cost, and increases prediction's quality on accuracy as well as average error distance.

Introduction

With the rapid development of computer science and technology, some data mining methods such as decision trees, neural networks have been widely studied and applied. Consequently, financial market which contains large amount of data flows and huge uncertainties has become an ideal place to practice these methods.

Effectively, stock's price is commonly modeled by scientists as a total stochastic process, which means it is unpredictable. However, in the real market, stock's prices could be impact by lots of important factors, such as investor's behavior, which is impacted by present price again. Affected by this influence circle, some hidden relationship could be enhanced, and the price becomes a fake stochastic process which gives us an opportunity to predict it. Data mining method is to detect these relationships.

There are several data mining methods as mentioned before. However, decision tree could be the best of them, as it could predict the results as well as provide an understandable logic
This paper is aimed to establish one process with some ameliorations of decision tree's algorithm to predict the second day's price change of stocks on financial market. The experiment environment is China's stock market. This paper will show experiment results based on this a process and some amelioration with one real stock. All raw data are from Yahoo web site, which are public free data.

1 Algorithm

Due to some particularities of financial market data, decision trees' algorithm should be adjusted in order to enhance its performance.

1.1 Preparation of data

As there are huge uncertainties and noises in financial data, pre-processing of data will be a crucial point for high quality results. During our experiment, there will be three steps to process financial data.

(1) Maximum selection of related attributes: the purpose of decision tree is to track useful

information from huge volume of data, thus more attributes will provide more information. However, more data means also more noise and more costs. In that case, Principle Component Analysis (PCA) could be used for selecting only important attributes of records. Surprisingly, our experiment shows that PCA could effectively reduce tree's construction cost but as well as prediction accuracy, which is not expected. Thus, on our following experiments, maximum related attributes will be selected for tree's construction. Concretely, all listed parameter on the market of target stock will be used as attributes. In addition, Market index parameter will be also taken into consideration.

(2) Segmentation of target variable's value: The aim of this paper is to predict second day's stock price change, which is a continuous variable. However, till now, decision tree can not efficiently predict continuous variable. In that case, segmentation of target variable's value is necessary. The simplest method of segmentation would be divided second day's price into two parts according to whether it is higher or not than previous price. There would be only two intervals: {Up} and {Down}. It seems that this structure will be easier to be predicted, but following experiments' results show a contrary conclusion: the accuracy is very low. This phenomenon may be explained as: theoretically, {Up} or {Down} could occupy each half chance on statistic of historical data, so it will increase the complexity of prediction by decision tree. Better approach will be dividing target variable's value into smaller intervals.

(3) Quantification of hidden information: Inside financial data, there is huge hidden information which is not directly listed in market data but could strongly influence market movement by impacting investor's attitude. Volatility, for example, is a very important risk parameter in the market. Thus, in this paper, we calculate daily volatility which is defined as $\mathrm{Vol_d} = \frac{\mathrm{P_{high}} - \mathrm{P_{low}}}{\mathrm{P_{open}}}$, weekly volatility which is defined as $\mathrm{Vol_w} = \frac{\sqrt{\mathrm{Var}\{\mathrm{P_t}, \mathrm{P_{t\text{-}1}} \ldots \mathrm{P_{t-4}}\}}}{\frac{1}{5}(\mathrm{P_t} + \mathrm{P_{t\text{-}1}} + \ldots + \mathrm{P_{t-4}})}$, and monthly volatility which is defined the same way as weekly volatility, but using 21 historical values instead of 5. Some other hidden parameters are also calculated, such as inter day change, which is defined as close price minus open price.

1.2 Dynamic constructed Tree

Another significant improvement of decision tree's algorithm on this paper is to propose a dynamic constructed tree. Based on previous study on this paper, more selected attributes could increase decision tree's prediction quality. But whether this principle could be used on sample records' selection? The experiment shows that they are not the same. As financial historical data could cover a very long period on time scope, which means it could contain several different market environments under significantly different economic conditions. Thus, more data may bring more noises into whole sample space and consequently confuse tree's construction logic.

One possible solution will be that, for each prediction, one additional algorithm firstly clean sample space, which means dynamically select only more important sample data for tree's construction. That's why it is called dynamic constructed tree. Concretely, if we will predict tomorrow's price change, the algorithm will select all historical sample records which have similar market environment as today. But, how could we define the similarity of two sample record? There

are many methods, but on this paper, we will use long term volatility as an indicator. Because long term volatility is a good reflection of risk level and is also easy to be calculated.

Three steps will be followed to define two sample's similarity:

1) Normalization: As different attribute generally have values in different scales, some attributes with commonly large numbers could have more impact on tree's construction than others. For example, daily traded amount could be several millions, while stock's close price is only about thirty. In that case, we could transpose each attribute's data into a probability space where, the expected value is 0 while the variance equals to 1. If there is a set of value X_i, we will firstly calculate their average value: $\mu = \frac{1}{N}\sum_i X_i$, and then the standard division: $\sigma = \sqrt{\frac{1}{N}\sum_i (X_i - \mu)^2}$. For each X_i in this set, we calculate $\tilde{X}_i = \frac{X_i - \mu}{\sigma}$, so that $\text{average}(\tilde{X}_i) = 0, \text{var}(\tilde{X}_i) = 1$.

2) Definition of distance between sample records: In order to express similarity in numbers, here we introduce a definition of distance, which is the same idea of Cartesian coordinate system. For example, if we will use four attributes for distance calculation between two records which are separately A (a_x, a_y, a_z, a_h) and B (b_x, b_y, b_z, b_h). Each one could be considered as a position of a point on a four dimension coordinate system. Thus, the distance between these two records could be easily defined as $l = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2 + (a_h - b_h)^2}$. Because we have already normalized data, every attribute should take the same weight during tree's construction.

3) Selection of useful region: Finally, x percentile method will be used for the selection. That means x% of total records according to a sort of distance from the smallest to largest will be selected for tree's construction.

While dynamic constructed tree could increase prediction accuracy, it augments also largely the construction cost, as it will dynamically choose important sample records before each prediction. Fortunately, the application of dynamic constructed tree on this paper is mainly on prediction of next day's price change, which means there will be only one time per day. Additionally, with rapid growth of information technology, computing cost will be less important than before.

2. Experiment results and analysis

The target stock on this paper is Shanghai Pudong Development Bank (SPDB), with the code 600000. As described before, we have included additional information by SSE Composite Index (SSECI) with code 000001. For each stock, those 20 attributes such as open price, highest price, lowest price, close price, daily volatility (defined before), inter day change (defined before), weekly volatility, monthly volatility, traded volume and traded amount, are used for tree's construction. There are totally 2355 effective records in the sample space dated from the 4th January, 2000 till the 12th July, 2010, which are already sufficient to build a decision tree.

On the first experiment, target variable has been divided into 11 intervals segments of 3% movement of the price, and the indicators for each interval are {-5， -4， -3， -2， -1， 0， 1， 2， 3， 4， 5}, which are shown as figure 1.

Price Change	indicator
(-INF, -12%)	-5
[-12%, -9%)	-4
[-9%, -6%)	-3
[-6%, -3%)	-2
[-3%, 0)	-1
0	0
(0, 3%]	1
(3%, 6%]	2
(6%, 9%]	3
(9%, 12%]	4
(12%, +INF]	5

Fig.1 intervals

A self testing method has been used on this paper for evaluation of prediction quality, which means we will use tree's construction sample as also the testing sample. This method is very popular in academic filed.

The result of first experiment is shown as figure 2.

Correctly Classified Instances	1948	**82.72%**
Incorrectly Classified Instances	407	**17.28%**
Kappa statistic	0.7368	
Mean absolute error	0.041	
Root mean squared error	0.1431	
Relative absolute error	33.82%	
Root relative squared error	58.21%	
Coverage of cases (0.95 level)	99.75%	
Mean rel. region size (0.95 level)	15.11%	
Total Number of Instances	2355	

Fig.2 first experiment result

Accuracy of prediction is up to 82.72%.

2.1 Analysis on segmentation's complexity

In this experiment, the target variable's value has been divided into only two intervals, which is expected to show a good result. The intervals are shown as figure 3.

Price Change	indicator
(-INF, 0]	Down
(0, +INF)	Up

Fig.3 simplest intervals

However, the result indicates that tree's prediction with this a simple segmentation is worse than a relative complex segmentation used in first experiment.

It seems that there is a relationship between the complexity of segmentation and the accuracy of prediction. To detect this important relationship, we have tested trees constructed with different complexity of segmentation. The target variable's value is divided by following 6 sets of separate points.

{0}

{-6%，0，6%}

{-8%，-4%，0，4%，8%}

{-9%，-6%，-3%，0，3%，6%，9%}

{-8%，-6%，-4%，-2%，0，2%，4%，6%，8%}

{-10%，-8%，-6%，-4%，-2%，0，2%，4%，6%，8%，10%}

And the respective results are shown on a curve as figure 4

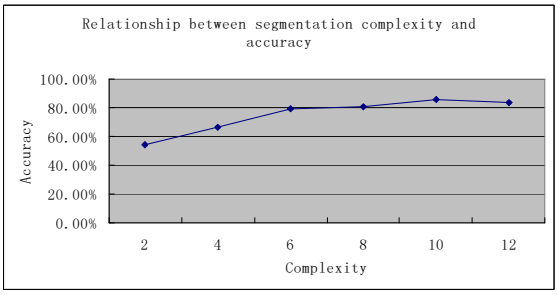

Fig.4 results with different intervals

It could be concluded that more complexity will bring more accuracy of prediction. However, this curve tends to be stable after the complexity of 10 intervals. So 10-intervals will be a good choice for following experiments.

2.2 Dynamic constructed tree

As long term volatility is a good indicator of market environment, on our experiment, we will use monthly volatility of SPDB as well as SSECI to calculate two sample records' distance according to previous description of method. And then we selected 60 percentile data, which is equivalent to 1413 sample records for tree's construction.

The result is shown as figure 5.

Correctly Classified Instances	1184	**83.79%**
Incorrectly Classified Instances	229	**16.21%**
Kappa statistic	0.7388	
Mean absolute error	0.0394	
Root mean squared error	0.1404	
Relative absolute error	34.46%	
Root relative squared error	58.80%	
Coverage of cases (0.95 level)	99.79%	
Mean rel. region size (0.95 level)	14.71%	
Total Number of Instances	1413	

Fig.5 result of dynamic constructed tree

Accuracy increases from 82.72% to 83.79%. Even this improvement is not so significant, dynamic constructed tree could save 40% of construction cost as it has used less data to build a tree. In that case, it is much better than a normal decision tree.

Further more, the prediction quality should be evaluated in an additional aspect, the average error distance. In fact, our target variable's value contains not only two possible values like Y or N, but lots of different intervals, so prediction accuracy could not completely reflect prediction's quality. For example, if the real change of price for tomorrow will be -1%, whether the predicted value is 1% or 10% will not impact prediction accuracy as they are all wrong. However, 1% or 10% prediction could deeply impact investor's P&L who has made decision according to predicted value. In that case, the Average error distance is defined as:

$$E = \sqrt{\frac{1}{N}\sum_{i}(\hat{C}_i - C_i)^2}$$

Where C_i is the real second day's price change, and $\hat{C}_i$ is the predicted value.

The comparison of average error distance (AED) between normal decision tree and dynamic constructed tree is shown as figure 6.

Tree type	Records number	Accuracy	AED
Normal	2355	82.7176 %	1.0602
Dynamic	1413	83.7933 %	0.9933

Fig.6 comparison of prediction quality

So we could conclude that dynamic constructed tree has better prediction quality and lower cost

than normal decision tree.

3. Conclusion

This paper has established one process which could predict the second day's price change on stock market. At the same time, several improvement of decision tree's algorithm, especially dynamic constructed tree have been proposed and tested. The experiments results show that such a process could predict stock price's trend with high accuracy and relatively low costs. There are still lots of potential applications of data mining on financial market to be explored.

Reference:

[1] Boris Kovalerchuk and Evgenii Vityaev, in: *Data Mining For Financial Applications* [M]. Springer, 2005

[2] J. R. Quinlan, in: *Improved Use of Continuous Attributes in C4.5* [D]. Sydney: University of Sydney, 2006

[3] J. Hull, in: *Options, Futures and Other Derivatives*. *6th edition* [M]. New Jersey: Prentice Hall, 2005

[4] Damiano Brigo and Fabio Mercurio, in: *Interest Rate Models - Theory and Practice: With Smile, Inflation and Credit* [M]. Springer, 2007

[5] Kohavi, R. and John, G. H, in: *Automatic parameter selection by minimizing estimated error* [D]. San Francisco: Morgan Kaufmann, 1995

[6] Witten, Ian H. and Frank E, in: *Data Mining. Practical Machine Learning Tools and Techniques. Second Edition* [M]. Morgan Kaufmann Publishers, 2005

Measuring Technology and Mechatronics Automation
10.4028/www.scientific.net/AMM.48-49

Stock Price's Prediction with Decision Tree
10.4028/www.scientific.net/AMM.48-49.1116