IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5253

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour. Copyright 2005, Idea Group Inc.

Dynamic Data Driven Forecasting Between Foreign Stock Markets

Chiu-Che Tseng, Ching-Tsai Kang and Jun Wei
Dept of Computer Science & Info. Systems, Texas A&M University - Commerce, TX 75429, USA, &
Dept of Mgt, & Mgt. Info. Systems, The University of West Florida, Pensacola, FL 32514, USA

ABSTRACT

This study examines the stock price effects of cross-listings ADRs by 8 Taiwanese companies during the period 1996 to 2003. After analyzing the numerical information, the result is going to be compared with those in 2004 to estimate the accuracy of prediction and sees if there is any positive co-relation between the stock prices in those two countries. In the study, we use decision tree and rule base system which is different from the traditional statistical methodology, which has been used in a fairly extensive empirical researches, to examine stock price information.

INTRODUCTION

There has been a dramatic increase in the trading of foreign stocks as investors recognize the need for international diversification and as foreign companies seek to broaden their shareholder base and raise capital. As a result, the number of American depositary receipts (ADR) listings on U.S. exchanges has also risen sharply. Though corporations view cross-listings as value enhancing, the changes in liquidity and volatility, and the cost of training associated with order flow migration following cross-listing may affect the quality of the domestic equity market.

The decision tree approach in this study is based on the C5.0 implementation of SPSS' Clementine[8]. The C5.0 decision tree learning algorithm is a commercial decision tree and rule induction engine developed by Ross Quinlan[17,20]. It is the state-of-the-art successor of the widely used C4.5 decision tree algorithm[20]. In contrast to other decision tree algorithms such as CART[3], C5.0 is able to generate trees with a varying number of branches per node. Decision trees based on C5.0 algorithm provide a clear indication of which attributes are important for the classification task at hand.

Since the trading hours of US markets do not coincide with Taiwanese markets, in this study, we apply decision tree and rule-based to analyze the stock price variances of ADRs in the US and those in Taiwanese market and see if the ADR listed in the US market really reflect the real-time information that became available while the US market was open right after the Taiwanese market was closed.

RELATIVE WORKS

Among numerous empirical researches, the co-relationship between international stock prices has always been discussed. Jayaraman et al. [10] show ADR listing to be associated with both positive abnormal returns on the listing day and an increase in the volatility of returns to the underlying stock. Foerster and Karolyi find that their sample of non-US firms cross-listing on US exchanges, over the period 1976 to 1992, experienced average excess returns of 19% during the year before listing, 1.2% the listing week, and — 14% the year following listing.

Moreover, Jiang [11] uses weekly data, over the sample period January 1980 to September 1994, on ADRs and market indices to conduct co-integration tests and to estimate EC and multifactor models. The study's findings shows that, most of the time, ADRs and the home markets are interrelated and do influence each other. As a result, the inter-relationship among international markets does exist.

Nevertheless, despite of the existing interrelationship between ADRs and stocks in home country, there are many other science and technical literatures which discuss the factor that really affect the price of ADRS and its returns. For instance, Park uses the data from July 1997 to June 1987 of the ADRs cross-listed by Japanese and English companies. He found that the prices of ADRs are mainly affected by those issued in home country but lightly affected by US market instead. Karolyi and Stulz [12] uses the daily ADRs data of eight Japanese companies during May 31st, 1988 to May 31st, 1992 as sample. He also found that the ADRs return is barely related to the daily exchange and bond return's impact in the USA. What's really matter to the ADRs return are Nikkei index and S&P 500 index in Japan. Besides; they also have positive movement.

Even so, there are not many empirical researches which discuss the issue of whether the cross-listed ADR has any influence on the stock issued in the home country. Whether there is any positive movement between the return of ADRs and that issued in the home country. In the following, we are going to apply decision tree and do adverse analysis to make contrasts with above researches.

TRAINING METHOD

Decision Tree Algorithm

We chose to use decision trees because they provide a comprehensible representation of their classification decisions. Although techniques such as boosting [5, 19] or support vector machines might obtain slightly higher classification accuracy, they require more computation during classification and they further obscure the decision making process.

A decision tree is a tree structure where each internal node denotes a test on a feature, each branch indicates an outcome of the test, and the leaf nodes represent class labels. An example of a decision tree is shown in Figure 1. To classify an observation, the *root* node tests the the value of feature *A*. If the outcome is greater than some value *x*, the observation is given a label of *Class 1*. If not, we descend the right subtree and test the value for feature *B*. Tests continue until a leaf node is reached. The label at the leaf node provides the class label for that observation.

We chose to use the C5.0 decision tree algorithm[17] a widely used and tested implementation. For details regarding the specifics of C5.0 the reader is referred to[17, 18]. Here we provide only the key aspects of the algorithm related to decision tree estimation, particularly as it pertains to feature selection. The most important element of the decision tree estimation algorithm is the method used to estimate splits at each internal node of the tree. To do this C5.0 uses a metric called the information gain ratio that measures the reduction in entropy in the data produced by a split. In this framework, the test at each node within a tree is selected based on splits of the training data that maximize the reduction in entropy of the descendant nodes. Using these criteria, the training data is recursively split such that the gain ratio is maximized at each node of the tree. This procedure continues until each leaf node

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

contains only examples of a single class or no gain in information is given by further testing. The result is often a very large, complex tree that overfits the training data. If the training data contains errors, then overfitting the tree to the data in this manner can lead to poor performance on unseen data. Therefore, the tree must be pruned back to reduce classification errors when data outside of the training set are to be classified. To address this problem C5.0 uses confidence-based pruning[17].

When using the decision tree to classify unseen examples, C5.0 supplies both a class label and a confidence value for its prediction. The confidence value is a decimal number ranging from zero to one – one meaning the highest confidence – and it is given for each instance.

Rule-Based System

Rule-based systems are a relatively simple model that can be adapted to any number of problems. As with any AI, a rule-based system has its strengths as well as limitations that must be considered before deciding if it's the right technique to use for a given problem. Overall, rule-based systems are really only feasible for problems for which any and all knowledge in the problem area can be written in the form of if-then rules and for which this problem area is not large. If there are too many rules, the system can become difficult to maintain and can suffer a performance hit.

The rule-based system itself uses a simple technique: It starts with a rule-base, which contains all of the appropriate knowledge encoded into If-Then rules(Figure 3), and a working memory, which may or may not initially contain any data, assertions or initially known information. The system examines all the rule conditions (IF) and determines a subset, the conflict set, of the rules whose conditions are satisfied based on the working memory. Of this conflict set, one of those rules is triggered (fired). Which one is chosen is based on a conflict resolution strategy. When the rule is fired, any actions specified in its THEN clause are carried out. These actions can modify the working memory, the rule-base itself, or do just about anything else the system programmer decides to include. This loop of firing rules and performing actions continues until one of two conditions are met: there are no more rules whose conditions are satisfied or a rule is fired whose action specifies the program should terminate.

TRAINING MODEL

Our study is designed to estimate the accuracy of the prediction. We first normalize the data used in the study and describe the supervised learning algorithm we chose. Then we train the data by C 5.0 decision tree classifier and rulesets classifier. After that, we use the data of 2004 to test and estimate the accuracy of prediction.

Figure 1. Decision Tree Abstraction (This shows how the values associated with certain features determine the class label. In this example, observations whose value for feature A is greater than x are assigned a class label of Class 1. Other classifications are based on the values of features B and C.)

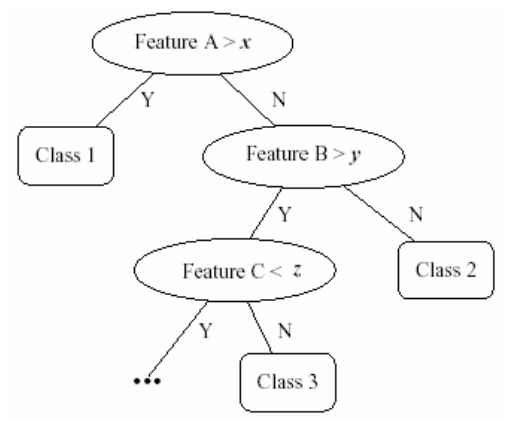

Figure 2. Portion of a Decision Tree Generated by C5.0

```
d5 > 0.0719:
:...d2 > 0.0347:
:  :...d2 <= 0.0415:
:  :  :...d5 <= 0.0833: -1 (4)
:  :  :  d5 > 0.0833: 1 (4/1)
:  :  d2 > 0.0415:
:  :  :...d3 <= 0.0188: -1 (5/2)
:  :      d3 > 0.0188: 5 (4)
:  d2 <= 0.0347:
:  :...d1 > 0.0583: 5 (5/1)
:      d1 <= 0.0583:
:      :...d2 <= -0.031:
:          :...d4 <= -0.0197: 0 (4/1)
:          :  d4 > -0.0197:
:          :  :...d2 <= -0.0483: -5 (3)
:          :      d2 > -0.0483:
:          :      :...d3 <= 0.0104: -3 (3/1)
:          :          d3 > 0.0104: 0.5 (3)
```

Data Source

We used daily data during June of 1996 to 2004 8 Taiwanese companies which have cross-listings of their stocks in the US stock market as American Deposit Receipt (ADR) is being applied and these Taiwanese companies daily data of Taiwanese stock market as sample. The data during 1996 to 2003 will be trained and constructed by C 5.0 decision tree classifier and rulesets classifier. Then they will use the decision tree and rulesets to run the data of 2004. The result would be used to compare with existed data of Taiwanese stock variation or ADR variation in 2004.

Before training, the data will be normalized and change into the format which could be recognized by C 5.0. The following are the data used in the training set:

*ADR or Stock Variation: (Close price – Open price) / Open price * 100%*

Figure 3. Portion of a Rulesets Generated by C5.0

```
Rule 1: (2, lift 22.1)
      d3 <= -0.0017
      d2 <= -0.0557
      d1 > 0.1148
      ->  class -5  [0.750]

Rule 2: (8/3, lift 17.7)
      d5 > -0.0348
      d5 <= 0.0112
      d4 > -0.0339
      d3 > -0.0478
      d3 <= -0.042
      d2 > -0.0211
      d2 <= 0.0098
      d1 > -0.0029
      ->  class -5  [0.600]

Rule 3: (3/1, lift 17.7)
      d1 > 0.0812
      d1 <= 0.0814
      ->  class -5  [0.600]
```

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Specifying the Classes

C5's job is to find how to predict a case's class from the values of the other attributes. C5 does this by constructing a *classifier* that makes this prediction. As we will see, C5 can construct classifiers expressed as *decision trees* or as sets of *rules*.

Before constructing decision tree and rulesets, we normalize two data sets. First is Taiwanese is stock previous 5 days variation and the class that was specified by today ADRs variation. Second is ADR is previous 5 days variation and the class that was specified by today Taiwanese stock.

Training Process

Decision Tree

Decision tree learning follows a kind of top-down, divide-and-conquer learning process. The basic algorithm for decision tree learning can be described as follows:

1. Based on an information gain measure, select an attribute to place at the root of the tree and branch for each possible value of the tree. Thereby, the underlying case set is split up into subsets, one for each value of the considered attribute.
2. Recursively repeat this process for each branch, using only those cases that actually reach that branch.
3. If at any time all instances at a node have the same classification, stop developing that part of the tree.

Rulesets

The Rulesets option causes classifiers to be expressed as rulesets rather than decision trees, here giving the following rules:

Rule 1: (31, lift 42.7)
thyroid surgery = f
TSH > 6
TT4 <= 37
-> class primary [0.970]

Rule 2: (63/6, lift 39.3)
TSH > 6
FTI <= 65
-> class primary [0.892]

Rule 3: (270/116, lift 10.3)
TSH > 6
➔ class compensated [0.570]

Each rule consists of:

1. A rule number — this is quite arbitrary and serves only to identify the rule.
2. Statistics (n, lift x) or (n/m, lift x) that summarize the performance of the rule. Similar to a leaf, n is the number of training cases covered by the rule and m, if it appears, shows how many of them do not belong to the class predicted by the rule. The rule's accuracy is estimated by the Laplace ratio (n-m+1)/(n+2). The lift x is the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set.
3. One or more conditions must be satisfied if the rule is to be applicable.
4. A class predicted by the rule.
5. A value between 0 and 1 that indicates the confidence with which this prediction is made. (Note: If boosting is used, this confidence is measured using an artificial weighting of the training cases and so does not reflect the accuracy of the rule.)

Table 1. Result is Acquired by Training the Data Using Rule Base of Each Company

	TW predicts ADR	ADR predicts TW
MXICY	55.26%	48.62%
ASTSF	50.43%	47.71%
TSM	54.39%	44.04%
ASX	62.28%	54.13%
UMC	56.14%	47.22%
SPIL	50.00%	43.12%
AUO	64.35%	47.71%
CHT	61.74%	39.45%

Table 2. Result is Acquired by Training the Data Using Rule Base of Total Eight Companies

	TW predicts ADR	ADR predicts TW
MXICY	56.14%	51.38%
ASTSF	57.39%	47.71%
TSM	45.61%	52.29%
ASX	62.28%	58.72%
UMC	64.04%	45.73%
SPIL	51.75%	55.96%
AUO	53.04%	50.46%
CHT	55.65%	37.61%

Test and Run

Once trained, a decision tree can predict a new data set by starting at the top of the tree and following a path down the branches until a leaf node is encountered. The path is determined by imposing the split rules on the values of the independent variables in the new data set.

The data of eight companies in 2004 are used to test the decision tree and make comparison with the analytical result. Run the network to predict future results. Run the network, and show it new input data and read the results.

SUMMARY AND CONCLUSION

The predicting which ranges from January 2, 2004 to June 16, 2004 is based on the stock price variation of Taiwanese's stocks and their ADRs. Each company has 115 trading days. The accuracy rate shows the moving trend between the predicted data and the real data.

Rule Base

Table 1 is the accuracy rate of prediction acquired by rule base training.

Table 1 shows the predicting accuracy rate by using each company's historical data. The second column shows the rate of the ADRs prediction, and contrary the third column shows the Taiwanese stocks'. From the table, apparently most company has higher accuracy rate of ADRs prediction.

Table 2 shows the predicting accuracy rate by using total eight companies' historical data. In table 2, you can see apparently most company has higher accuracy rate of ADRs prediction as Table 1.

The ADR of AUO (AU Optornics Corp.), for instance, has the result of 64.35% (74/115) same moving trend as the Taiwanese stock price within

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 3. Result is Acquired by Training the Data Using Decision Tree of Each Company

	TW predicts ADR	ADR predicts TW
MXICY	60.53%	44.95%
ASTSF	49.57%	50.46%
TSM	47.37%	45.87%
ASX	61.40%	49.54%
UMC	53.51%	37.96%
SPIL	50.88%	47.71%
AUO	54.78%	46.79%
CHT	60.00%	44.04%

Table 4. Result is Acquired by Training the Data Using Decision Tree of Total Eight Companies

	TW predicts ADR	ADR predicts TW
MXICY	53.51%	47.71%
ASTSF	49.57%	42.20%
TSM	43.86%	58.72%
ASX	54.39%	55.05%
UMC	51.75%	38.89%
SPIL	46.49%	58.72%
AUO	50.43%	48.62%
CHT	52.17%	50.46%

the total 115 trading days. Based on both two tables above, it can approve that Taiwanese stock price plays the main role to affect those price of the stock in America.

Decision Tree

Table 3 is the accuracy rate of prediction acquired by decision tree training.

Table 3 shows the predicting accuracy rate by using each company's historical data. From the table, apparently most companies have a higher accuracy rate of ADRs prediction.

Table 4 shows the predicting accuracy rate by using total eight companies' historical data. In table 4, you can see apparently most company has higher accuracy rate of ADRs prediction as table 3.

The ADR of ASX (ADV SEMICON ADR), for instance, has the result of 61.40% (70/115) same moving trend as the Taiwanese stock price within the total 115 trading days. It can prove that Taiwanese stock price plays the main role to in affecting the prices of the stocks in America.

CONCLUSION

As the result above, not all the companies have the same moving trend. But for most of the cases, the results show that it has higher accuracy rate of ADRs prediction by using both rule base and the decision tree. Some companies even have more than 60% accuracy rate. Nevertheless, it might be some other factors such as politics, economics that are possible to affect the stock price, so not all of the companies' stock price can be predicted by using rule base or decision tree. Overall, we believe that the ADRs co-related with stock prices, especially using Taiwanese stock price to predict ADR's.

Future Work

In the future, we will continue our job to compare the result obtain by other AI techniques such as neural network.

REFERENCES

[1] Anant K., S., Dennis E., L. (1st Qtr., 1996). *Valuation Effects of Foreign Company Listings on U.S. Exchanges*. Journal of International Business Studies, 27(1), p.67-88.

[2] Bennett K. and Campbell C. (2000). *Support Vector Machines:Hype or Hallelujah?* SIGKDD Explorations, 2:1–13, 2000.

[3] Berry, M.J. and Linoff, G(1997*). Data Mining Techniques For Marketing, Sales and Customer Support*, John Wiley & Sons, Inc., New York, 1997.

[4] David E., Mehdi, S. (2001). *American depositary receipts: An analysis of international stock price movements*. International Review of Financial Analysis, 10, p.323-363.

[5] Freund Y. (1995). *Boosting a Weak Learning Algorithm by Majority*. Information and Computation, 121(2):256–285, 1995.

[6] Geng, C., Man, L., W. *Implementing neural networks for decision support in direct marketing.*International Journal of Market Research. 46(2).

[7] Granzow M.., Berrar D.., Dubitzky W.., Schuster A.., Azuaje F.J.., Eils R. (2001) *Tumor Classification by Gene Expression Profiling:Comparison and Validation of Five Clustering Methods,* ACM SIGBIO Newsletter,Volume 21 , Issue 1 (April 2001), 16 - 22

[8] Huntsberger, T.L. and Aijimarangsee P(1992). *Parallel selforganising feature maps for unsupervised pattern recognition.* Bezdek J.C. and Pal N.R, Fuzzy models for pattern recognition, 483-495. IEEE Press, New York. 1992

[9] James P. Early, Carla E. Brodley, Catherine Rosenberg(2003), *Behavioral Authentication of Server Flows*, 19th Annual Computer Security Applications Conference, Las Vegas, Nevada, December 8-12, 2003

[10] Jayaraman, Shastri and Tandon (1993*). The impact of international cross-listings on risk and return: the evidence from ADRs.* Journal of Banking and Finance. 91-103.

[11] Jiang C. (1998)"*Diversification with ADRs: The Dynamics and the Pricing Factors."* Journal of Business Finance and Accounting, 25, (1998), pp. 683–700.

[12] Karolyi (1996). *What happens to stocks that list shares abroad?* A Survey of the evidence and its managerial implications. University of Western Ontario Working paper.

[13] Karolyi, G. Andrew, and René M. Stulz, (1996), "*Why do Markets Move Together?, An Investigation of U.S.-Japan Stock Return Comovements,"* Journal of Finance 51, 951-986.

[14] Lan D., Jack G., Ananth M(1998*)., International Cross-Listing and Order Flow Migration : Evidence From An Emerging Market*, The Journal of Finance.vol LIII,No. 6, Dec 1998

[15] Mark E., W., William J., Crowder. (1998). *COINTEGRATION, FORECASTING AND INTERNATIONAL STOCK PRICES*. Global Finance Journal. 9(2).

[16] Niklas, A., Jan, A. (2002). *Testing for cointegration between international stock prices*. Applied Financial Economics. 12, p.851-861.

[17] Quinlan. J. R. *C4.5: Programs for Machine Learning*.Morgan Kaufmann, San Mateo, CA, 1993.

[18] Ross Quinlan. Data Mining Tools See5 and C5.0. URL http://www.rulequest.com/see5-info.html.

[19] Robert E. Schapire. *A Brief Introduction to Boosting*. In IJCAI, pages 1401–1406, 1999. URL citeseer. nj.nec.com/schapire99brief.html.

[20] Witten, I.H., and Frank, E(1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Pub., San Francisco, 1999.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

0 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/proceeding-paper/dynamic-data-driven-forecasting-between/32664

Related Content

A Rough Set Theory Approach for Rule Generation and Validation Using RSES
Hemant Ranaand Manohar Lal (2016). *International Journal of Rough Sets and Data Analysis (pp. 55-70).*
www.irma-international.org/article/a-rough-set-theory-approach-for-rule-generation-and-validation-using-rses/144706

Health Wearables Turn to Fashion
Lambert Spaanenburg (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 6114-6123).*
www.irma-international.org/chapter/health-wearables-turn-to-fashion/184310

Medical Software Engineering Cost Estimation Model Based on Multimodal Improved Genetic Algorithm
Liangyu Li, Zulkefli Bin Mansor, Siyi Li, Xiaoyan Zhao, QingJie Zhongand Xuwei Guo (2024). *International Journal of Information Technologies and Systems Approach (pp. 1-22).*
www.irma-international.org/article/medical-software-engineering-cost-estimation-model-based-on-multimodal-improved-genetic-algorithm/364841

Analysis of Gait Flow Image and Gait Gaussian Image Using Extension Neural Network for Gait Recognition
Parul Arora, Smriti Srivastavaand Shivank Singhal (2016). *International Journal of Rough Sets and Data Analysis (pp. 45-64).*
www.irma-international.org/article/analysis-of-gait-flow-image-and-gait-gaussian-image-using-extension-neural-network-for-gait-recognition/150464

Designing Personalised Learning Resources for Disabled Students Using an Ontology-Driven Community of Agents
Julius T. Nganjiand Mike Brayshaw (2013). *Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies (pp. 81-102).*
www.irma-international.org/chapter/designing-personalised-learning-resources-disabled/70711