Application of Decision Trees for Portfolio Diversification in Indian Share Market

Shehroz S Khan
Department of Information Technology,
National University of Ireland Galway,
Galway, Republic of Ireland
s.khan1@nuigalway.ie

Amir Ahmad
Department of Physics,
Indian Institute of Technology,
Delhi, India-110016
amirahmad01@gmail.com

Abstract

Portfolio Diversification (i.e. possessing shares of not one but many companies) is considered as an important task in the investors' community to minimize investment risks. Classification of companies (belonging to various Industrial Sectors) into different categories and assign ratings on the basis of their performance is a critical step for Portfolio Diversification. In this paper we present a machine learning approach to identify different categories of companies on the basis of their annual balance sheets. We employed C4.5, a decision tree based machine learning algorithm to first learn and understand the classification rules generated for companies of each industrial sector and then predicting the category of uncategorized companies (companies without much research) within their respective sector. The results were impressive and shows that using this approach portfolios can be discretely diversified effectively with less time and effort involved into profit making and non-vital assets, paving a way for less risks higher returns on investments. The methodology works well for non-experts of finance too. We compared our results with the categories of these companies as suggested by ICICIDIRECT service (a renowned financial service company). The comparison shows the applicability and usability of decision tree approach as an important tool for taking investment decisions in respect to Indian share market.

Keywords

Portfolio Diversification, Decision Trees, C4.5, Industrial Sectors

1. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a new powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is an ideal tool to model different business applications, such as investment return prediction, market fluctuation simulation, stock / mutual fund analysis, consumption categorization etc. Data mining tools can predict future market trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [1] [2] [3]. The finance industry is a real veteran in such technology. Banks, brokerages, insurance and pharmaceutical companies [4] have been relying on various data mining tools for over two decades.

Lately various machine learning algorithms have been studied to understand market severity and predict the future trends. Neural networks are extensively used for stock predictions and have become the standard for detecting credit-card fraud. Since 1992, neural networks usage have slashed such incidents by 70% or more in U.S.A. [5]. Hochreiter, S. et al [6] presented an efficient algorithm called "flat minimum search" that outperforms other widely used methods on stock market prediction tasks. E. W. Saad et al [7] compared three neural network models for low false alarm stock trend predictions. Skabar et al [8] describe a methodology by which neural networks

can be trained indirectly, using a genetic algorithm based weight optimization procedure, to determine buy and sell points for financial commodities traded on a stock exchange. Other machine learning paradigms have also been studied for stock market analysis like genetic programming [9], decision trees [10] [11] etc. M. Harries et al [10] investigated the use of strategies to enhance an existing machine learning tool, C4.5, to deal with concept drift and non-determinism in a financial time series domain. M. Noor et al [11] presented new Fuzzy Decision Tree (FDT) approach that uses weighted fuzzy production rules. They calculated several certainty factors using some important variables (e.g. effect of other companies, effect of other stock exchanges, effect of overall world situation, effect of political situation etc) in dynamic stock market. They predicted stock share indices and showed improve computational efficiency of data mining approaches. Ahmad et al [12] presented an unsupervised learning method to group Indian companies of same industry sector into different categories on the basis of their annual balance sheets to help investors identifying companies with maximum return.

The rest of the paper is as follows. Section 2 describes the problem of Portfolio Diversification in detail. In Section 3, we briefly introduce C4.5 decision tree algorithm. In section 4 we present experimental results on real time data taken from ICICIDIRECT, a famous Indian Online Trading Web Portal. Section 5 concludes our presentation.

2. Portfolio Diversification

Sharpe et al [13] pointed out that there exist two types of risk in holding shares

- Market Risks or Systematic Risks
- Firm Specific Risks or Unsystematic Risks

All shares are exposed to market risk due to changes in the level of interest rate, changes in tax laws, government trade policies, general economic activities and possibility of war and peace. The impact of these factors on all stocks is not identical. These kinds of risks are known as Market Risks. Firm Specific Risks are only limited to a particular firm or industry sector. For example a particular economic event can affect the price of a specific stock (of a given industrial sector) but still has no impact on other stocks in the market.

Modern portfolio theory [13] has shown that holding a diversified portfolio of many shares can eliminate Unsystematic Risks. Over time some firms will perform better than others in properly diversified portfolio; then differences will balance. The gain in one stock is offset by loss in other, stabilizing the overall earnings of the investment. Portfolio Diversification [13] (holding shares of many companies) is an important strategy in shares business to reduce financial risks. Investor should diversify their investment, in other words they should include shares of more than one company in their investment basket.

There are two steps in selection of shares of good companies that generate good return on investment

- Selection of Good Industrial Sectors.
- Selection of profit making companies in those Good Industrial Sectors.

Performance of companies depends upon different factors, *Internal Factors* (companies' management, R&D, good marketing etc.) and *External Factors* (cost of raw material, government policies, world scenario etc.). Every industrial sector has few External Factors associated with them that affect the profitability of companies within these industrial sectors. It is difficult to use computational methods to compare companies of different Industrial Sectors because sector related factors cannot be quantified. But in a given sector where all Internal Factors affect the companies in similar manner, the attributes based on company's balance sheet could be used to compare them. Financial analysts wish to categorize companies into three groups viz **Good**, **Average** and **Bad** on the basis of expected return on the investment [13]. Though it may be easy to draw a clear-cut boundary between Good companies and Bad companies, it is potentially arduous to differentiate Good companies' category and Average companies'

category. While deciding their portfolio, investors never want Bad companies in their portfolios because that will have devastating effect on their subsequent return on investment. They also want to reduce the possibility of having Average companies in their portfolio because that will reduce return on their investment. They also don't want to categorize Good company as Average company because by doing this they will miss the opportunity of getting good return on their investment.

3. Brief Introduction to C4.5

Most of Machine Learning algorithms use Neural Networks, Statistical Learning and Decision Trees to accomplish various classification tasks. There exists a genre of supervised machine learning algorithms that build decision trees, association / classification rules as a result of their classification process. There are variety of methods, tools and software that work in this way, and they differ in the manner they construct these rules. Many inductive algorithms have been reported, like CART [14], ID3 [15], C4.5 [16], that works around the similar concept. The C4.5 algorithm induces classification rules from training sets to form decision trees. C4.5 uses the concept of gain ratio to make a tree of classificatory decisions with respect to a previously chosen target classification. By considering which of the attributes is best for discriminating among cases at a particular node in the tree, it build up a tree of decisions that allows us to navigate from the root of the tree to a leaf node by continually examining attributes.

4. Experiments

Financial data of different companies of various industrial sectors was taken from ICICIDIRECT web portal [17]. Each company's balance sheet was represented by 12 attributes that signify the financial condition of the company. These attributes are presented in Table 1.

Table 1. List of financial attributes (FY=Financial year)

1	**2**	**3**	**4**
FY2001 Net Profit	FY2002 Net Profit	Latest FY Net Profit	Earning per Share
5	**6**	**7**	**8**
Price of Share / Earning per Share	Price of Share /Book Value	Yield (%)	Return on Capital Employed (%)
9	**10**	**11**	**12**
Market Capitalization	365 days Average Traded Quantity	Last Dividend (%)	Last traded Price

ICICIDIRECT has a three level rating system based on the expected return on investment (on the basis of these attributes) – OverWeight (Good), EqualWeight (Average) and Underweight (Bad). These rating can be broadly defined as:-

OverWeight - Stocks with this rating may be the best candidates for investing.

EqualWeight -An investor who has stocks with an EqualWeight rating can consider continuing to hold the stock. But he should monitor the stock's ongoing performance and compare the potential benefits of owning a stock with higher ratings.

UnderWeight - An investor having stocks with an Underweight rating should consider the benefits of owning a stock with higher ratings.

In our study we used decision trees to first learn (on above mentioned) predefined categories from the dataset of various industrial sectors and then predict the category of uncategorized companies'. We used C4.5 algorithm for classification of company's of each sector with their ratings considered as their respective classes. We performed ten 10-folds Cross Validation (C.V.) to estimate the prediction accuracy by randomly choosing 90% of the companies for classification (learning) and remaining 10% (uncategorized) were used for predicting their respective ratings. We also carried out a simultaneous

study to understand the classification rules induced from training sets. The study reported in this paper is carried out for companies belonging to four different industrial sectors viz Textile, Information Technology, Automobiles and Fast Moving Consumable Goods.

4.1 Textile Sector

We have data and categories of 33 companies (8 OverWeight, 12 EqualWeight, and 13 Underweight). Using C4.5 algorithm, we observed an average learning error of 3.44% and average prediction error of 22.5%. When all the 10 C.V. runs are clubbed together, we observed that 10 out of 12 attributes played role in classification (which are attribute number 1, 2, 3, 4, 6, 8, 9, 10, 11, 12 (see table1)).

4.2 Information Technology Sector

We have data and categories of 53 companies (14 OverWeight, 14 EqualWeight, and 25 Underweight). Applying C4.5 algorithm gave an average learning error of 6.72% and average prediction error of 28.34%. In this sector, 8 out of 12 attributes contributed (combining all the 10 C.V. runs) in making classification decision; these are number 1, 3, 4, 6, 7, 8, 9, 11 (see table1).

4.3 Automobile Sector

We have data and categories of 34 companies (14 OverWeight, 16 EqualWeight, and 4 Underweight). Application of C4.5 algorithm gave us an average learning error of 5.01% and average prediction error of 30.0%. In this case, when the classification rules of all the 10 C.V. runs were combined, we observed that only 6 attributes were pivotal for learning classification rules from the data. These attributes are number 1, 3, 8, 9, 11, 12 (see table1).

4.4 Fast Moving Consumable Good (FMCG) Sector

We have data and categories of 25 companies (8 OverWeight, 13 EqualWeight, and 4 Underweight). C4.5 algorithm yields an average learning error of 4.05% and average prediction error of 17.76%. In this case, when the classification rules of all the 10 C.V. runs were combined, only 3 attributes proved to be important for learning classification rules from the data. These attributes are number 5, 11, 12 (see table1).

We made several inferences from this study. The first one is the steady rate of self-classification (learning accuracy) and less variant value of prediction error. There were some overlapping between OverWeight and EqualWeight rated companies and between EqualWeight and Underweight rated companies, that is why sometimes the prediction rate looks a noticeable number. But there was negligible overlapping between OverWeight and Underweight rated companies which conform to the basic motive of this study - to identify OverWeight and Underweight companies with higher accuracy so as to prevent investor to make any adverse investment decision (Figure 1 and 2). EqualWeight companies should be dealt with uttermost care, as our analysis also shows; these are the categories of companies which are more susceptible to plunge to OverWeight and Underweight ratings. For making more accurate prediction to increase returns on investment, more informative financial attributes may be needed.

Figure 1: Ideal case classification scenario for the distinction of OverWeight, EqualWeight and Underweight rated Companies

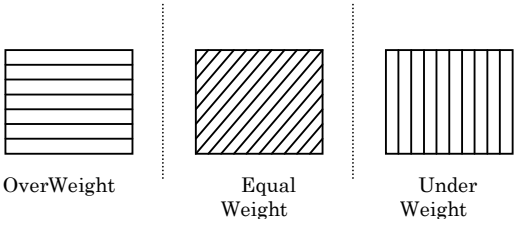
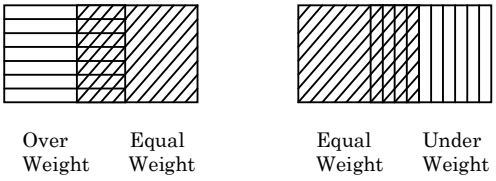


Figure 2: Overlapping between OverWeight and EqualWeight rated Companies and EqualWeight and UnderWeight rated companies may cause sub-optimal returns on investments



Analysis of Classification Rules

When the classification rules for all the companies of all given sectors were studied some interesting results came upfront. From the experiments we found out that we get different set of classification rules for companies of different sectors, which is in conformity that External Factors vary for different industrial sectors (refer section 2). We also observed a similar pattern of classification rules for the companies within any given sector, which shows that Internal Factors for a given industrial sector affects the companies in a similar manner (refer section 2). We also observed that '*not all attributes*' contributed in taking classification decision (from the available data and attribute set). We found that attribute number 1, 3, 8, 9 and 11 (see table 1) played prominent role in predicting portfolio ratings in all the sectors studied and therefore these attributes are unaffected by the prevailing External / Internal Factors. Attribute number 4, 6 and 12 (see table 1) were present in at least three of the four sectors studied. We infer that these attributes shows the mounting influence of Internal Factors governing companies within a particular industrial sector. Interestingly attribute number 5 i.e. Price of Share/Earning per Share, has not contributed in determining rating of any company in any sector except the FMCG sector.

5. Conclusion

While analyzing financial credibility of companies', individual preferences and projections play important role and can have adverse effects on return on investments. By using machine-learning algorithms we can reduce such kind of biased decision possibility. In this paper we have shown the use and applicability of C4.5 decision tree algorithm for Portfolio Diversification problem for Indian Share Market. From the analysis we found that, that this technique ensures negligible overlap between OverWeight and Underweight rated companies but there is some overlapping between EqualWeight and OverWeight rated companies and EqualWeight and Underweight rated companies. This analysis can help an investor to identify companies for achieving optimum returns on investments by chalking out OverWeight/Underweight companies in a given industrial sector. An investor can acquire portfolios of just not one company but a congregation of profit making OverWeight rated companies from different Industrial Sectors and consider abandoning Underweight rated companies from his basket. After studying the classification rules we observed that there is a need for more informative attributes to describe the data to increase the prediction rate. We also observed that only few of the attributes played a pivotal role in learning classification rules from the data. A few of the attributes did not contribute much in classification and hence may be considered less important while taking investment decisions. A limitation of this study is the lack of sufficient data, but we plan to acquire more data to perform a more comprehensive study. This technique can be used with other conventional investment analysis tools for better financial analysis of the companies.

References

[1]. Wang, H. and Weigend, A.S. (2002): "Data mining for financial decision making", EDITORIAL, Journal of Decision Support Systems, 457-460.

[2]. http://www.thearling.com/text/dmwhite/dmwhite.htm

[3]. Bonchi, F., Giannotti, F., Mainetto, G., and Pedreschi, D. (1999): "Using Data mining techniques in Fiscal Fraud Detection", In Proc. DaWak'99, Firtst Int'l Conf on Data Warehousing and Knowledge Discovery

[4]. Cohen, J.J. and Olivia Parr Rud, C. (2000): "Data Mining of Market Knowledge in The Pharmaceutical Industry", NESUG, Philadelphia, PA,

[5]. http://www.businessweek.com/bw50/content/mar2003/a3826072.htm

[6]. Hochreiter, S. and Schmidhuber, J. (1997): "Flat Minima", Neural Computation, 9(1), 1-43,

[7]. Saad, E. W. , Prokhorov, D. V. and Wunsch, D. C. (1998): "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks", IEEE Transactions on Neural Networks, 9(6), 1456-1470,

[8]. Skabar, I. and Cloete, I. (2002): "Neural Networks, Financial Trading and the Efficient Markets Hypothesis", Australian Computer Science Communications, Volume 24 , Issue 1, 241 - 249,

[9]. Markose, S. Tsang, E., Er, H. and Salhi, A. (2001): "Evolutionary Arbitrage for FTSE-100 Index Options and Futures", In Proceedings of the 2001 Congress of Evolutionary Computation, 275-282,

[10]. Harries, M. and Horn, K. (1995): "Detecting Concept Drift in Financial Time Series Prediction using Symbolic Machine Learning", In Xin Yao, editor, Eighth Australian Joint Conference on Artificial Intelligence, World Scientific Publishing, Singapore, 91- 98,

[11]. Noor, M. and Khokhar, R. H. (2004): "Fuzzy Decision Tree for Data Mining of Time Series Stock Market Databases", 5th International conference for the Critical Assessment of Microarray Data Analysis, CAMDA,

[12]. Ahmad, K., Warsi, S. and Ahmad, A. (2004): "Application of K-means Clustering for Portfolio Diversification in Indian Share Market", KBCS-2004, Fifth International Conference On Knowledge Based Computer Systems, India,

[13]. Sharpe, W. F., Alexander, G. J. and Bailey, J. V. (1999), "*Investment*", 6th Edition, *Prentice-Hall*

[14]. Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and Regression Trees*

[15]. Quinlan, J, R. (1986): "Induction of Decision Trees", Machine Learning Journal,

[16]. Quinlan, J, R. (1993), "*C4.5: Programs for Machine Learning", Morgan Kaufmann*, San Francisco, CA.

[17]. http://secure.icicidirect.com/

Photograph:

Name of Author: Shehroz S. Khan

Areas of Research Interest: Clustering, One-sided Classification, Bayesian networks, Decision Trees

Organization: National University of Ireland Galway, Galway, Ireland

E-mil ID: s.khan1@nuigalway.ie

Photograph:

Name of Author: Amir Ahmad

Areas of Research Interest: Classifiers Ensemble, Clustering, Application of Data Mining Algorithms

Organization: Department of Physics, Indian Institute of Technology, Delhi, India

E-mil ID: amirahmad01@gmail.com