IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 3, March 2017

A Study on Classification and Prediction Techniques in Data Mining for Financial Applications

B. Sharmila[1], Dr. R. Khanchana[2]

Research Scholar, Department of Computer Science, Sri Ramakrishna Arts and Science College for Women, Coimbatore, India[1]

Research Supervisor, Department of Computer Science, Sri Ramakrishna Arts and Science College for Women, Coimbatore, India[2]

Abstract: Financial Forecast is an estimate of future financial outcomes for a company. A financial forecast is an economist's best guess of what will happen to a company in financial terms over a given time period. Data mining has made a significant role in the prediction techniques for financial applications. Technology such as data ware house and data mining has made a significant contribution for the prediction in the entire service sector of the organization .The Data mining techniques are used to extract hidden patterns, from large amount of financial data with the effective usage of sentiment features and opinion mining. The paper describes the effective usage of data mining predictive techniques for the financial applications.

Keywords: Data mining, Forecasting, Sentiment analysis, Feature selection, Opinion mining.

I. INTRODUCTION

The application of DM techniques on financial data can contribute to the solution of classification and prediction problems and facilitate the decision making process. Typical examples of financial classification problems are corporate bankruptcy, credit risk estimation, managing the performance of mutual equity funds. The importance of Data Mining in finance and accounting has been recognized by many organizations. Specifics of data mining in finance are coming from the need to forecast multidimensional time series with high level of noise. Efficiency criteria (e.g., the maximum of trading profit) in addition to prediction accuracy.

Make coordinated multi-resolution forecast (minutes, days, weeks, months, and years). Explain the forecasting model for future investment decisions. DM aims to discover valid, complex and not obvious hidden information from large amounts of data. For this reason, another equivalent term for DM is Knowledge Discovery in Databases (KDD), which is equally often met in the literature. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning and artificial intelligence to analyze current data to make predictions about future. The patterns found in historical and transactional data can be used to identify risks and opportunities for future. Predictive analytics allows organizations to become proactive, forward looking, anticipating outcomes and behaviors based upon the data and not on a hunch or assumptions. Industry experts are focusing on customers and developing customer-centric projects. A financial application is a software program that facilitates the management of business processes that deal with money. Finance is a field that deals with the study of investments.

Data Mining Process Data mining, referred to as knowledge discovery from data (KDD) is the extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses. a. Database, data warehouse, or other information repository. This is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request. Knowledge base search, evaluate the interestingness of resulting patterns. Data mining engine is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association analysis, classification, and evolution and deviation analysis. The pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 3, March 2017

method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to find the search to only the interesting patterns.

II. DATA MINING TECHNIQUES FOR CLASSIFICATION AND PREDICTION FOR FINANCIAL DATA

Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. Predictive analytics, pattern recognition, and classification problems are not new. Long used in the financial services and insurance industries, predictive analytics is about using statistics, data mining, and game theory to analyze current and historical facts in order to make predictions about future events.

A. Regression analysis

Regression analysis can be used to model the relationship between one or more independent or predictor variable and a dependent or response variable. Regression analysis is a good choice when all of the predictor variables are continuous .Many problems are solved by linear regression .straight line regression is the simplest form of regression. Polynomial regression is the basic linear model. Generalized linear models represent the theoretical foundation of the methods. Logistic regression and Poisson regression are the common types of linear models for the prediction.

B. Association rule

Association rules are about discovering interesting relationships between variables in large databases. It is a technique applied in data mining and uses rules to discover regularities between products. For example, if someone buys peanut butter and jelly, he or she is likely to buy bread. The idea behind association rules is to understand when a customer does X, he or she will most likely do Y. Understanding those kinds of relationships can help with forecasting sales, promotional pricing, or product placements.

C. Neural network

Neural network has the capacity of high tolerance of noisy data and has the ability to classify patterns. They are suited for continuous valued inputs and outputs. Parallelization techniques can be used to speed up the computation process. Self organizing map (SOM) is one of the most popular neural network methods for cluster analysis. The SOM approach is used for web document clustering. Neural networks are ideal for deriving meaning from complicated data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques. Research result shows that neural networks technique is useful for predicting customer demand and customer segmentation. An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

D. Decision trees

Decision trees use real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process. Decision trees enable to explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action. Such an analysis is useful for investment opportunities, and especially with limited resources of financial data.

E. Bayesian classification

Bayesian approaches are fundamentally important DM technique. Given the probability distribution, Bayes classifier can provably achieve the optimal result. Bayesian method is based on the probability theory. Bayes Rule is applied here to calculate the posterior from the prior and the likelihood, because the later two is generally easier to be calculated from a probability model.

F.GINI index: The GINI index is used in CART (Classification and regression tress) which generates the decision technique. CART handles missing data and trains for the training data. It has multiple information based on the information to classify the dataset.

G. Support vector machine (SVM)

Support vector machine is a method for the classification of both linear and nonlinear data. The SVM finds the hyper plane using support vectors Support vector machine (SVM) is a useful technique for data classification, regression and prediction. However, in financial market, the data often has enormous noises and complex dimensionality, and the problem is overcome by SVM which has been successfully used in the field of prediction. SVM can treat higher dimensional data better even with a relative low amount of training set. Furthermore, it can present a good ability of generalization for complex model.

H. NLP based prediction

The prediction of stock price is very complicated task. This prediction based on news articles which uses sentiment analysis one of the techniques in Text Mining. The steps to be followed in NLP based prediction such as Collection of reviews based on stock market from news articles. R.S.S. feed is the main source of news articles. The news articles mainly consist of business and market

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 3, March 2017

related news. The top news of stock market is retrieved by using R.S.S feed. Sentiment Analysis is applied to the news articles. Natural Language Processing referred to as sentiment analysis is used to extract from computational linguistics and text documents to extract and identify subjective information. The sentiment analysis tasks determine the polarity of the text at document level, sentence level and feature level as positive, negative and neutral.

III. METHODOLOGY FOR PREDICTION

Sentiment features: Sentiment scores are generated from textual data alone. They are used as the input of the prediction model. Technical indicator features from numeric data are introduced into the prediction model. Clustering is conducted based on technical indicator features for prediction. There is an SVR (Support Vector Regression) model for each cluster. The SVR models are modeled based on bag-of-words (BOW) models and topic features from textual data.

Feature combination: Feature combination is conducted on two groups of experiments. One group is based on the same features used in context analysis, namely, technical indicator features, BOW and topic features. The purpose of this group of experiments is to verify if the clustering in context analysis leads to better performance. The other group of experiments is based on technical indicator features and the sentiment features, which shows the best performance as a single type of features. The purpose of this is to identify a better model from feature combination. The method is not widely diffused among statisticians. It combines the qualities based on rank it favorably compared with existing techniques. It has a well behavior even if the ratio between the number of variables and the number of observations becomes very unfavorable, with highly correlated predictors. Another advantage is the principle of kernel (the famous "kernel trick"). It is possible to construct a non-linear model without explicitly having to produce new descriptors. Opinion mining, as a sub-discipline with data mining and computational linguistics, is referred to as the computational techniques used to extract, classify, understand, and assess the opinions expressed in various online news sources, social media comments, and other user-generated content . 'Sentiment' analysis is often used in opinion mining to identify sentiments, affect, subjectivity, and other emotional states in the online text.

Feature selection methods

Feature Selection methods involve into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used. Lexicon-based approaches usually begin with a small set of 'seed' words. Then they bootstrap this set through synonym detection or on-line resources to obtain a larger lexicon. Statistical approaches, on the other hand, are fully automatic. The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string.

Sentiment classification techniques

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called opinion lexicon. Manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

Time series analysis

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

Goals of time series analysis:

- Descriptive: Identify patterns in correlated data—trends and seasonal variation
- Explanation: understanding and modeling the data
- Forecasting: prediction of short-term trends from previous patterns

A time series in finance is a sequence of numerical data points in successive order. In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. The data is accessed from multiple sources including files, spreadsheets, databases, data providers, and the Web. Store data in financial time-series objects to simplify data management, data transformation, missing-data handling, and date-math calculations. Different kinds of financial time-series have been recorded and studied for decades.

Nowadays, all transactions on a financial market are recorded, leading to a huge amount of data available, either for free in the Internet or commercially. Financial time-series analysis is of great interest to practitioners as well as to theoreticians, for making inferences and predictions. Time series forecasting uses information regarding historical values and associated patterns to predict future activity.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 3, March 2017

Deep Learning Approach For Finance Using Big Data Analytics

A key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and un-categorized. In the present study, we explore how Deep Learning can be utilized for addressing some important problems in Big Data Analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We also investigate some aspects of Deep Learning research that need further exploration financial investigation for specific challenges introduced by Big Data Analytics, including streaming data, high-dimensional data, scalability of models, and distributed computing. Deep learning helps in the effective prediction of data analysis which would retrieve the information about the entire portfolio.

IV. CONCLUSION

Many techniques of data mining in prediction for financial data were analyzed and the study shows the effective usage of predictive methods and classifications. In the field of data mining in finance we expect an extensive growth of hybrid methods that combine different models and provide a better performance than can be achieved by individuals. Predictive analytics is ideal for classification and particularly for the investment patterns. The process starts with feature selection then proceeds with representation, data collection and management, pre-processing, data mining, post processing, and in the end performance evaluation. The usage of data mining in the field of finance is endless and provides a unique environment where efficiency of the methods can be tested instantly.

ACKNOWLEDGMENT

I would like to express my deep gratitude to **Dr. R. Khanchana** for the encouragement, comments and help. She helped me to write the research paper and given me the necessary requirements to build my own identity in the research area.

REFERENCES

[1] Azoff, E., Neural networks time series forecasting of financial markets, Wiley, 1994.
[2] Wang J., Data Mining; opportunities and challenges, Idea Group, London, 2003
[3] K.S. Shin and Y.J. Lee: "A Genetic Algorithm Application in Bankruptcy Prediction Modeling", Expert Systems with Applications, Volume 23, Issue 3, October, 2002, pp.321-328.
[4] C. Spathis: "Detecting False Financial Statements Using Published Data: some Evidence from Greece", Managerial Auditing Journal, Volume 17, No 4, 2002, pp.179-191.
[5] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Technique", 3rd edition
[6] Michal Meltzer, Using Data Mining on the road to be successful part III, published in October 2004, retrieved 2nd January
[7] Kang Liu, Liheng Xu, and Jun Zhao,"Co-Extracting Opinion Targets And Opinion Words From Online Reviews Based On The Word alignment Model", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No.3,Mach 2015.
[8] Kovalerchuk B, Vityaev E (2000). Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer.
[9] Lee LW, Wang LW, Chen SM, Leu YH (2006). Handling forecasting problems based on twofactor high-order time series, IEEE Transactions on Fuzzy Systems.
[10] V. Sehgal and C. Song, "SOPS: Stock Prediction using Web Sentiment," Seventh IEEE International Conference on Data Mining – Workshops, 2009, pp.21-26..
[11] Fung, G.P.C.; Yu, J.X.; Lam, W.: Stock Prediction: Integrating Text Mining Approach Using Real-time News. In: Proceedings IEEE Int. Conference on Computational Intelligence for Financial Engineering. Hong Kong 2003, pp. 395-402
[12] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, Opinion Mining and Analysis: A Survey, International Journal on Natural Language Computing (IJNLC) Vol. 2, No. 3, June 2013.
[13] Yune, H.,H. Kim, J.Y. Chang, "An Efficient Search Method of Product Review using Opinion Mining Techniques," Journal of KIISE: Computing Practices and Letters, Vol.16, No.2, 2010.2, pp.222-226.
[14] G. Vinodhini and R M. Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
[15] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Production and hosting by Elsevier B.V. on behalf of Ain Shams University, 27 May 2014.
[16] Data Mining With Predictive Analytics for Financial Applications, ISSN: 2395-3470 International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-1, January 2016.