ISSN 2320 - 2602

Volume 2, No.11, November 2013

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2013/ijacst012112013.pdf

KENERL BASED SVM CLASSIFICATION FOR FINANCIAL NEWS

Ms. D. Preetha [1], Mrs .K.Mythili [2]

[1]Research Scholar, Department of Computer Science
Hindusthan College of Arts and Science, Coimbatore-28.
India, it.preetha@gmail.com

[2] MSc., M.Phil.,(Ph.D)., Associate Professor PG & Research Department of Computer Application
Hindusthan College of Arts and Science, Coimbatore-28. India, Mythiliarul@gmail.com

ABSTRACT

Financial news based forecasting is still becomes major important problem to classify the new under various categories based on modern time series .Further it can be used for stock market prediction analysis and stock market improvement in the business environment for individual organization. From this new content are major important to influence market prediction analysis. In this research the stock market prediction data as considered as input text data with the stock price can be predicted much well by looking at appeared news articles in stock market by sharing numeral of factors of news from companies in local and global politics to news of superpower economy. In this paper proposed a classification based method to classify news data for given stock market prediction system .To execute efficient classification by establishing efficient multiple kernel learning in SVM classification. Kernel function called Gaussian Radial basis Polynomial Function (GRPF) is introduced that could improve the classification accuracy of Support Vector Machines (SVMs) for both linear and non-linear data sets. The aspire is to train Support Vector Machines (SVMs) with dissimilar kernels compare with SVM learning algorithm in classification task .Kernel SVM takes a set of input data and predict possible outcomes fro given input data from news for making the predication result . Experiments Results show that the Kernel SVM classification can provide high accuracy for prediction of the result than the previous SVM methods.

Keywords: Stock market prediction, News classification, Kernel based SVM, Data mining

1. INTRODUCTION

Data mining has been considerably used in the analysis/prediction of the stock market results bases on the news. During the earlier years data mining is also known as the knowledge discovery phase in the databases also known as knowledge discovery in databases has recognized its position as a high-flying and important research area. Data mining has been used in various data domains. Data mining can be regarded as an algorithmic process that takes data as input and yields patterns, Data mining task can be classified into two categories such that Descriptive Mining and Predictive Mining. The expressive Mining technique such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data,. The Predictive Mining techniques like Classification had been used for prediction analysis, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables. In this paper study the data mining under the predicative analysis for future values and find unknown variables to predict data for news based results

But the KDD data are considered as the textual format, numerical values and data warehouses. Though, a group of information these days is obtainable in the appearance of text, counting documents, information, manual, email, and etc. The mounting numeral of textual data has led to information detection in unstructured data known as text mining or text data mining. Text mining is a promising knowledge for analyze large collection of unstructured documents for the purpose of extract attractive and non-trivial pattern or information.

Financial making statements about events are still regard as one of the majority challenging application of up-to-the-minute moment in time series forecasting. Because financial series have very complex behavior, it becomes intrinsically noisy, non-stationary, and deterministically chaotic. The figure of proposed method in financial instance series prediction is extremely great. These methods rely seriously by means of structured and numerical databases.

Specific features of the input data are selected to represent the whole document. In classifying the news two significant factors be supposed to be measured; the news content and the numerical market data such as stock prices. Both of these factors shall be measured for classifier input production on behalf of the news part consists of two main responsibilities: feature selection and feature weighting. Primary a set of features will be chosen to represent a piece of news and next step is to assign weights to theses selected features.

The analysis of the news based result produces the prediction results .The news based market prediction can be performed by using the classification in both training phase and functional phase. In functional phase one of the predefined class is automatically assigned to the news and then classifier should be trained in the training phase based on the class labels in the functional phase .Machine learning based methods help to classify or automate the process to classify the news prediction results .In this paper Kernel based SVM classification methods training phase, a set of preparation data shall be prepared which in our case the train data are the pre-classified news and market information such as market price. Pre-classified news and probably market numerical data will be process to be feed into the classifier for training. The trained classifier would be prepared to get a portion of news and allocate a class to it in prepared phase of the organization.

The main objective of this design is to reply the query of how to predict the response of stock market to news article, which are rich in precious information and are additional superior to numeric data. The pressure of news articles on stock price pressure group, dissimilar data and text mining techniques are implementing to make the prediction model. The entire work can be done in the following manner:

- To perform the prediction goal news labeling is important to classify the news/information identify the class that each news article belongs to and label them accordingly. It can be done either manually /automatically based on the financial market experts will read the news and assign a class based on their opinion.
- Classifier input generation with features of the input data are selected to represent the whole document. Major important factors are considered to classify the news articles in both news content and numerical market data such as stock prices.
- Finally classify the news information from the above input generation result it shows that best classification results.

2. RELATED WORK

K. Kyong-jae and I. Han [2] et.al including artificial neural networks (ANNs) with GA . ANN has preeminent learning capability whilst it is repeatedly confront with incompatible and impulsive performance for boisterous data. In addition, sometimes the amount of data is so large that the learning of patterns may not work well. In exacting, the subsistence of continuous data and large quantity of data may pretense a challenging task to explicit concepts extraction from the raw data due to the huge amount of data space determined by continuous features. Many researchers in the society of data mining are interested in the reduction of dimensionality. Feature discretization is closely related to the dimensionality reduction may improve the generalizability of the learned results. K. Kyong-jae and I. Han [2] et.al study uses GA to search the optimal or near-optimal thresholds for feature discretization. In addition, this study simultaneously searches the connection weights between layers in ANN. The genetically evolved connection weights mitigate the well-known limitations of the gradient descent algorithm.

Quah and B. Srinivasan et.al [3] rising significance in the role of equities to both the worldwide and restricted investors, the assortment of good-looking stock is of greatest meaning to ensure a good come back. Consequently, dependable tools in the assortment process know how to be of great support to this investor. Successful and resourceful tool/system gives the investor the aggressive edge over others as he/she can identify the theater stocks with bare minimum effort. Innovative investors opt to employ information technology to improve the efficiency in the process. This is done through transforming trading strategies into computer known languages so as to exploit the logical processing power of the computer. This greatly reduces the moment in time and attempt in short-listing the list of good-looking stocks.

W. Huang et.al [4] investigates the predictability of financial association way by means of SVM by forecasting the daily association direction of NIKKEI 225 index. To appraise the forecasting facility of SVM it compares its presentation with individuals of Linear Discriminant Analysis(DA), Quadratic Discriminant Analysis(QDA) and Elman Backpropagation Neural Networks(EBNN). Finally show that SVM outperforms the other classification methods.

In automated conveying, occasion embossed arithmetical market data is analyze to decide the correct group for a part of news [5][6][7]. Usually an occasion time approximately the news let go will be chosen and the price is analyzed in with the aim of interval to decide the news contact. For illustration Fung et al. [8] divided the moment in time series data into independent segment and label the segment according to its average slope. Mittermayer and Knolmayer [9] labeled the news based on the percentage change of the price 15 minutes after the news release

In the first move toward the news comfortable is used as enter data source, whilst in the second approach marketplace information such as stock price at the moment of news release [7], closing price and change indicator values [10] are included in the classifier input.

R.P. Schumaker et.al [7] proposed a stock market prediction research encapsulates two elemental trading philosophies; fundamental and technical approaches. In fundamental analysis, stock market price movements are believed to derive from a security's relative data. General list use numeric information such as salary, ratio, and administration usefulness to establish future forecasts. In technical analysis, it is whispered that marketplace timing is key.

Number of final classes can affect formative the criterion for transmission a class to a piece of information. E.g. if the numeral of final classes are 2, then more often than not the price in a time range after or before the news discharge will be compare with the value at the time of news release to decide about the news label [9][11].

3. FINANCIAL NEWS CLASSIFICATION WITH SVM AND KERNEL SVM

In this research first the identify the important researchers result when time the stories are released and identify the strong relationship among them as well as the time when the stock prices fluctuate. It becomes to make enter pat of the new area for predicting the movement of stock trend movement based on the content of news story. At the same time as there are numerous shows potential forecasting methods to predict accumulate marketplace actions based on top of numeric moment in time series data, the number of predict method relating to the submission of text mining techniques using news articles is few. This is because text mining seems toward being more multifaceted than data mining as it involves commerce by means of text data to facilitate are intrinsically unstructured and fuzzy.

Predict the reaction of stock market result to the news article it contains the enrich valuable information and some of them becomes greater to numeric data. To the influence of the new article to stock price movement different data mining methods and text mining based methods are used to predict the stock market results .For this first identify the important news with time series and after that name the information under different categories to classify the data the learning process is managed by Machine learning algorithm, Before that first label the data by identifying the important features in the news articles by measuring the term frequency with the help of TF-IDF frequency .Then the results are feed into the system for classify the news article data for stock prediction. Moreover in this design aims to show that how much valuable information exists in textual databases which with the help of text mining techniques can be extracted and used for various purposes.

Features are the representatives of a document in a classification problem. Based on the classification goal a setoff features from document should be selected which best convey the document content in feature selections two main approaches has been followed. Generating a term dictionary a set of terms are gathered and used as the fixed vector fundamentals. It represents the group of financial expert choose the delegate terms, for every group there exists a set of special vocabulary that if exist in a manuscript the possibility of belong the document to that category would be superior. It can be used extracted for training process, then stop words are removed.

For selection of the best feature by calculation of weight values to each terms /features in the news .Here weight values are randomly assigned from zero to one .The words with higher degree of membership as input features tf- idf is used to calculate the weights,

Term Frequency:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the divider is the sum of number of occurrences of all terms in document d_j, that is, the size of the document | d_j|

Inverse document frequency ,

$$\mathrm{idf}_{i,j} = \log \frac{|D|}{|\{j: t_i \in d_j\}|}$$

IDF = log (total-number-of-documents / number-of-documents-containing-t)

SVM classification

In dispensation the news normally the goal is to categorize the information/news into two classes either good news or bad news as regards the selected stock. From time to time this classification is comprehensive and additional category representative neutral news is further added. If the degree of news authority is significant to be recognized more concluding categories will be defined .In the majority of the methodologies the Support Vector Machine (SVM) selected as their classification algorithm. SVM is dual classifiers which try to classify the input data by important a hyper plane or a set of hyper planes in high-dimensional space. SVM tries to maximize the distance of the hyper plane with the nearest data points of each class.

Kernel based SVM classification

Support vector machine classification is choosing a suitable kernel of SVMs for a particular application, i.e. various applications need different kernels to get reliable classification results. It is well known that the two typical kernel functions often used in SVMs are the radial basis function kernel and polynomial kernel. More recent kernels are presented to handle high dimension data sets and are computationally efficient when handling non-separable data with multi attributes. However, it is difficult to find kernels that are able to achieve high classification accuracy for a diversity of data sets. In order to construct kernel functions from existing ones or by using some other simpler kernel functions as building blocks, the closure properties of kernel functions are essential. Classical kernels, such as Gauss RBF and POLY functions, can be used to transfer non-separable data to separable, but their performance in terms of accuracy is dependent on the given data sets. The following POLY function performs well with nearly all data sets, except high dimension ones :

$$POLY\ (x, z) = (x^T z + 1)^d$$

where d is the polynomial degree. The same performance is obtained with the Gauss RBF of the following form:

$$RBF\ (x, z) = \exp\ (-\gamma||x - z||^2)$$

where γ is appositive parameter controlling the radius. The Polynomial Radial basis Function (PRBF) as:

$$PRBF = ((1 + \exp(\omega))/v)^d$$

where $\omega = |x - z|$ and V=p*d is a prescribed parameter. Completely achieving a SVM with high accuracy classification therefore, requires specifying high quality kernel function,

Gauss RBF

Combine POLY, RBF, and PRBF into one kernel to become:

$$\mathrm{GRPF}(x, z) = \left(\frac{d + r.\exp(-||X - z||^r/(r.\sigma^2))}{r + d}\right)^{d+1}$$

$$\theta^0 = arg\ \min_{\theta} T(\alpha^0, \theta)$$

where σ is a statistic distribution of the probability density function of the input data; and the values of r(r >1) and d can be obtained by optimizing the parameters using the training data. The proposed kernel has the advantages of generality. However, The existing kernels such as PRBF and proposed Gaussian and polynomials kernel function by setting d and r in different values. For example if d =0, we get Exponential Radial when r= 1 and Gaussian Radial for r= 2 and so on. Moreover various kernels can be obtained by optimizing the parameters using the training data .GRPF depends on two parameters d and r, encoded into a Vector θ = (d, r). We thus consider a class of decision functions parameterized by α, b, θ:

$$f_{\alpha,b,\theta}(x) = sign(\sum_{i=1}^{l} \alpha_i y_i GPRF_\theta(x, z) + b)$$

and want to choose the values of the parameters α and θ such that w is maximized (maximum margin algorithm) and T is minimized with best kernel parameters. More precisely, for θ fixed, we want to have $\alpha^0 = \arg\max w(\alpha)$ *and choose* θ^0 such that $\theta^0 = \arg\min_\theta T(\alpha^0, \theta)$

When, θ is a one dimensional parameter with finite number of values and picks the one which gives the lowest value of the criterion T.When both T and the SVM solution are continuous with respect to h a better approach. They used an incremental optimization algorithm, one can train an SVMwith little effort when θ is changed by a small amount. However, as soon as h has more than one component computing T(α, θ) for every possible value of h becomes intractable, and one rather looks for a way to optimize θ along a trajectory in the kernel parameter space. In this work, we use the gradient of a model selection criterion to optimize the model parameters.

The procedure is changed as :

1. Initialize θ to some value.
2. Using a standard SVM algorithm, find the maximum of the quadratic form w
$\alpha^0 = \arg\max w(\alpha)$
3. Update the parameters h such that T is minimized. This is typically achieved by a gradient step
4. Go to step 2 or stop when the minimum of T is reached.

4. EXPERIMENTAL RESULTS

In this section we measure the performance of the system in terms of the precision, recall, Fmeasure with Gaussian Radial basis Polynomial Function (GRPF) SVM learning algorithms for news classification .Measuring these parameters show the results of the accuracy in terms of how classification is achieved better than the existing methods.

4.1 Precision

Precision value is calculated is based on the retrieval of information at true positive (TP) prediction, false positive (FP).In privacy preservation data precision is calculated the percentage of preserved data results returned that are relevant.

Precision =TP/ (TP+FP)

4.2 Recall

Recall value is calculated is based on the retrieval of information at true positive (TP) prediction, false negative (FP). In privacy preservation approach the data precision is calculated with percentage of positive results returned that are recall in this context is also referred to as the True Positive Rate (TP). Recall is the fraction of relevant instances that are retrieved,

Recall =TP/(TP+FN)

4.3 Fmeasure

Fmeasure is a measure of a test's accuracy. It includes both the precision p and the recall r of the test to compute the score p is the numeral of accurate results divided by the numeral of all returned results and r is the number of correct results divided by the number of results that should have been returned.

Fmeasure = 2.Precision.recall /(precision + recall)

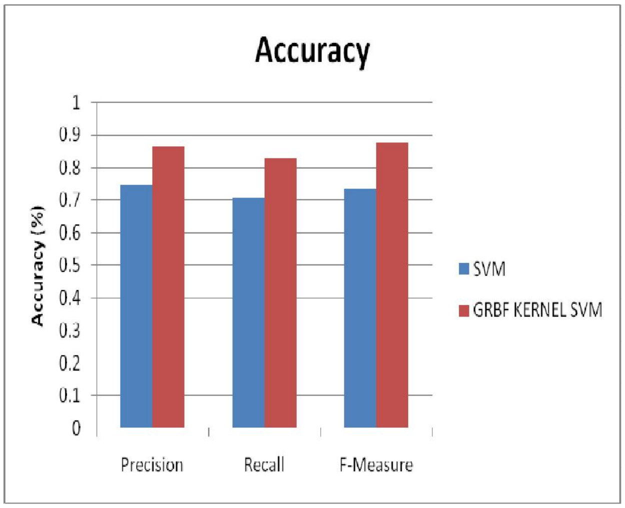

Figure 1: Prediction results of classification

In this Figure 1 measure the Prediction results of classification with news data .The system measures the Precision, Recall and Fmeasure results Support Vector machine and Kernel function called Gaussian Radial basis Polynomial Function (GRPF) with SVM for prediction of classification results .It shows that GRBF-SVM improves the classification result in terms of Precision, Recall and Fmeasure results.

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

In this paper Financial News Classification based solely on the technical and fundamental data analysis was performed by using SVM and GRBF-SVM classification. It is used to predict the Financial News based on the contents of relevant news articles which can be accomplished by building a prediction model which is able to classify the news as either rise or drop. It is capable of adapting to the dynamic changes in the financial business environment and can able to manage huge amount of data with classification methods. The prediction model applying all the types of news related to auto industry in general and the ones related to competitors and compare the results with the current prediction model .The experimental results showed that a GRBF-SVM can be a valid to predict the results. The results also prove that this technique provides high classification accuracy.

5.2 FUTURE WORK

In future work we apply other classification techniques such as fuzzy neural network, ANN for improving the classification accuracy then the existing system.

REFERENCES

[1] T. Hellström and K. Holmström, "Predicting the Stock Market," Technical Report Series IMATOM-1997-07, 1998.

[2] K. Kyong-jae and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," Expert Systems with Applications, vol. 19, 2000, pp. 125-132(8).

[3] T.S. Quah and B. Srinivasan, "Improving Returns on Stock Investment through Neural Network Selection," Expert Syst. Appl.,vol. 17, 1999, pp. 295-301.

[4] W. Huang, Y. Nakamori, and S. Wang, "Forecasting stock market movement direction with support vector machine," Computers & Operations Research, vol. 32, 2005, pp. 2513-2522.

[5] G. Rachlin, M. Last, D. Alberg, and A. Kandel, "ADMIRAL: A Data Mining Based Financial Trading System," 2007 IEEE Symposium on Computational Intelligence and Data Mining, 2007, pp. 720-725.

[6] Y. Zhai, A. Hsu, and S. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," Lecture Notes in Computer Science, 2007, pp. 1087-1096.

[7] R.P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news:The AZFin Text system," ACM Transactions on Information Systems, vol. 27, 2009, pp. 1-19.

[8] G. Fung, J. Yu, and W. Lam, "News sensitive stock trend prediction," Lecture Notes in Computer Science, vol. Volume 233, 2002, p. 481– 493.

[9] M. Mittermayer and G.F. Knolmayer, "NewsCATS: A News Categorization And Trading System," Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006, pp. 0-5.

[10] A. Mahajan, L. Dey, and S.M. Haque, "Mining Financial News for Major Events and Their Impacts on the Market," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2008, pp. 423-426.

[11] M. Mittermayer and G. Knolmayer, Text mining systems for market response to news: A survey, 2006.