IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Financial Market Predictions: Generative Vs Discriminative Methods

P.Misra[1], S.Chaurasia[2*]

[1]Department of Computer Science, University of Lucknow, Lucknow, India
[2*]Department of Computer Science, University of Lucknow, Lucknow, India

[*]*Corresponding Author: siddharth515@gmail.com, Tel.: +91-94101-10598*

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/Jul/2018

Abstract— Prediction of stock market is acclaimed by many as one of the most challenging areas for machine learning. The existence of quant industry that makes use of artificial intelligence based computational methods to predict the market provide enough evidence contrary to Efficient Market Hypothesis and Random Walk theory. Recent research on the financial market has focused on machine learning based approaches where instead of specifying the rules, learning algorithms are employed to make use of existing data. Financial markets provide one of the most organized data sets where data from each tick is recorded. Both generative and discriminative class of machine learning techniques have been explored in search of improved accuracy. Even with the abundance of structured financial data, complex, chaotic and nonlinear nature of the market that can ride on public emotions keeps the scope for probabilistic generative methods. This paper discusses the usability of machine learning techniques from both the classes: generative and discriminative along with the characteristics of data that enables them.

Keywords— Discriminative, Financial market, Generative, Machine Learning, Sentiment Analysis, Supervised Learning

I. INTRODUCTION

Based on the inputs utilized, predictions in the world of finance can be divided into two categories: First deals with historical financial numbers and second analyzes texts for the prediction tasks. For first, historical data of price and volume for different securities, indexes, currency is used as input to statistical and machine learning (ML) techniques to forecast the future. Second, pacts with text mining methods to gauge sentiments, information from reports, to make predictions. Text analysis gets value from behavioural economics which says public-mood has an impact on markets. Information available online in the form of news, blogs, comments on social media platforms is mined to get information used for prediction.

ML Methods applied in both the categories can be classified into one of the two classes: Generative or Discriminative. Generative methods model how the data was generated to classify the input data stream. It tries to answer based on generation assumptions, which category is most likely to generate an appropriate data stream. Discriminative method does not focus on how the data was generated; rather it just tries to classify the input data.

Algorithms dealing with numbers and text implement the two classes differently. While, for the numeric data it is quite straightforward, for text it needs to be interpreted. Qualitative information available in the text needs to be converted to quantitative which can then be utilized by algorithms. One interpretation is given by [1] with their relative advantages of the two approaches for sentiment analysis. At the very basic level, the word-based n-gram (generative) model and the character-based tagging (discriminative) model are two approaches in the literature. The former gives excellent performance for the in-vocabulary (IV) words; however, it handles out-of-vocabulary (OOV) words poorly. On the other hand, though the latter is more robust for OOV words, it fails to deliver satisfactory performance for IV words. These two approaches behave differently due to the unit they use (word vs character) and the model form they adopt (generative vs discriminative).

This paper first presents the general viewpoints irrespective of the domain it is being applied to. More notably, discussion relates to data that is numeric and structured. Paper compares both classes of methods when applied to the world of finance for prediction and forecasting using two very different approaches based on the inputs they take - Structured quantitative data or unstructured text data from various media. Paper surveys the current literature for the methods that are being used for financial predictions as it indicates the focus on panaches that are likely to give at par results which have been achieved in the field of research. The review viewpoint taken is unique, as to the best of our knowledge, there has been no study done which tried to focus on utilization of specific class of method for market prediction in both structured quantitative finance or unstructured sentiment analysis.

II. General Discussion

Generative classifiers learn a model of joint probability p(x,y), of input x and label y, and make the predictions by using Bayes rules to calculate p(y|x) and then pick the most likelihood label. Discriminative classifiers model the posterior p(y|x) directly or learn a direct map for the input x to class labels [2].

Published literature indicates wider acceptance for discriminative learning for classification tasks. Still, there have been results where generative learning is shown to be at par in the various experimental setting. Long[3] declares discriminative learning as clear winner with its title. The conclusion drawn in the paper are:

- There exists a class of distribution, parameterized by d (dimension of samples) such that there is a discriminative algorithm that can learn the correct classifier with only 2log(2/d) samples, while the number of samples required for any generative classifier is at least d.
- Since the requirements of generative learning are stronger than those of discriminative learning, it follows that in the framework used discriminative learning is strictly "easier" than generative learning.

As mentioned by the authors, their experimental setup is simple and artificial. Examples in this paper showed the limitation of algorithms rather than a whole class of generative algorithms.

In an earlier paper, [2] mentions the prevailing consensus on the adoption of discriminative methods for classification but argues on both the approaches being two sides of the coin by being complementary instead of being competitive.

Paper compares one discriminative algorithm, i.e. logistic regression with another generative algorithm, i.e. Naive Bayes. The two algorithms form "generative-discriminative" pair. Naive Bayes maximizes the total joint likelihood, $\sum^{n}_{i=1}$logP(xi,yi) over the samples, while logistic regression maximizes the total conditional likelihood, $\sum^{n}_{i=1}$logP(yi|xi) over the same parametric model.

Which approach will be more efficient depends on scenario that needs to be modeled. The conclusion made in Ng's paper are:

- The generative model has a higher asymptotic error (as the number of training examples becomes significant) than the discriminative model.
- The generative model may approach its asymptotic error much faster than the discriminative model possibly with many training examples that is only logarithmic, rather than linear, in the number of parameters.

III. Financial viewpoint

In this section, we do a broad classification of ML classifiers into either discriminative or generative classifier based on the way they approach the learning in the domain of finance.

A discriminative classifier creates the model by depending on the observed data. It hardly makes any assumption on the distribution but depends on the quality of the data. For e.g. Logistic Regression.

A generative classifier learns the mechanism that generates the data by estimating the assumptions and distributions. It then uses this knowledge to predict unseen data because it assumes the model that was learned captures the real model for e.g. Nave Bayes classifier.

Most of the methods can be classified in either category as they have apparent implementation approach for e.g. Naive Bayes, Logistic Regression. In contrast, some of the ML method categorization may depend on their implementation viz. ANNs. Though popular variants of ANNs are discriminative, e.g., feedforward, but these can be modelled in generative style too, e.g. Bayesian regularized NN as used by [4]. In a nutshell, if the algorithm cares about the distribution of Y, it is generative, if not, then it is discriminative. Table 1 makes an attempt to group the two categories of methods as Discriminative and Generative. This list is in no way exhaustive and only tries to group the popular ML techniques in the two groups. It is inspired from [5] where a comparison is presented between deep models.

Generative models are specified as probabilistic graphical models, which offer rich representations of the independence relations in the dataset. Discriminative models focus on modelling the boundary between classes. Thus, if the same computational power is given, a discriminative model tends to yield more complex representations of the boundary than a generative model.

When dealing with non-stationary distributions, the test data may be generated by different underlying distributions than the training data; generative methods will have an edge as they try to model the data generation process. It is easier to detect distribution changes and update a generative model accordingly than do this for a decision boundary like in a discriminative method viz. SVM.

In general, generative models outperform discriminative models on smaller datasets as their generative assumptions place some structure on the model that prevent over-fitting. In contrast, if data is in abundance and training data provides a good representation of test set, discriminative methods should outperform the generative methods.

Table 1. High-level comparison on the basis of attributes for Generative and Discriminative methods

Attributes	Discriminative	Generative
Interpretation	Harder	Easy (generative 'story')
Scalability	Easier	Harder
Accuracy of results	Better when data is in abundance	Better when not much data is present
Tackling uncertainty	Hard	Easy
Empirical Goal	Classification, Feature Learning	Classification (via Bayes rule), latent variable inference
Approach	Focuses on boundary between classes	Learns independent relations in data
Boundary Representation	More complex representations	Less Complex
Data Generation	Cannot generate	Can generate
Anticipation capability	Cannot anticipate unseen data, i.e. stateless	Can anticipate the input not yet seen, i.e. stateful
Perform with respect to inputs	Static inputs like images, numbers	Dynamic inputs like videos, texts from sentences, speech recognition
Evaluation	End performance	On almost every intermediate quantity
Examples	LR, LoR, Feed forward ANN, SVM, DT	NB, HMM, NDA, RBM, BN-ANN, DBN

Abbreviations used in table: ANN - Artificial Neural Network, LR - Linear Regression, LoR - Logistic Regression, SVM - Support Vector Machine, DT - Decision Tree, NB - Naive Bayes, HMM - Hidden Markov Model, NDA - Normal Discriminant Analysis, RBM -Restricted Boltzmann Machine, BN-ANN – Bayesian regularized ANN, DBN - Deep Belief Network

A. *Discriminative Methods*

Financial market prediction problems can be expressed as an attempt to find a relationship between an output y and a set of D inputs x where $x = x_1; x_2 \dots x_D$, i.e. $y = f(x)$.

If y represents a future asset return or price observation at some point in the future, the function f could be learned from in-sample training data so that when new unseen (out-of-sample) data is presented, a new prediction can be made. Both regression where $y \epsilon R$ and classification where $y \epsilon \{-1,+1\}$, e.g. a return is positive or negative, would be useful to investigate [6].

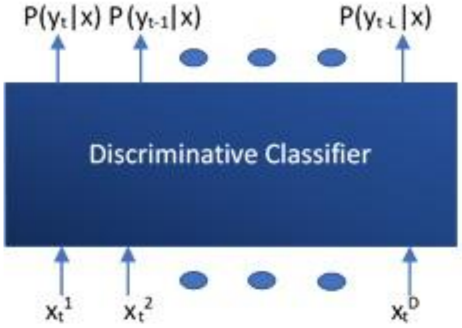

Figure1. Discriminative Classifier

x could be composed of exogenous variables or L lags of y; T time steps into the future so that:

$$y_{t+T} = f(x_t)$$

where

$$x_t = x_t^1 \dots x_t^D \ , \ y_t; y_{t1 \dots} y_{t-L}$$

The discriminative classifier is trained using all the training examples of different classes as $y = y_{t1 \dots} y_{t-L}$, $x_t = x_t^1 \dots x_t^D$ to form a single discriminative classifier as P(y|x). Artificial Neural Network (ANN) and Support Vector Machine (SVM) are among the most popular ones applied where structured financial data is involved.

B. *Generative Methods*

For financial prediction, if the assumption is that price action is the result of noisy and chaotic observations in the market, the need can be to generate the data that can probabilistically present a representation for scenarios. Moreover, this also signifies that it is possible to generate points with low probability p(x) which would otherwise be difficult to observe.

Generative models specify a joint probability distribution over observation and label sequences. are used in machine learning for either modelling data directly (i.e., modelling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function [7].

$$\text{Given } x_t = x_t^1 \dots x_t^D \quad \text{we find}$$

$$P(x|y) \ ; y = y_{t1 \dots} y_{t-L}$$

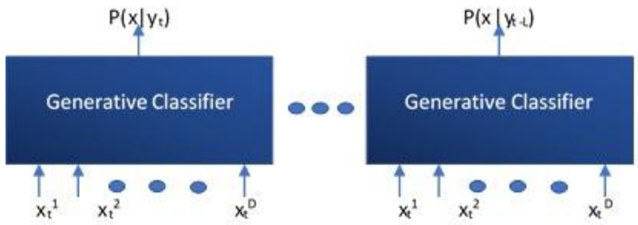

Figure2. Generative Classifier

In figure 2, L probabilistic models are trained independently. Generative means that the model produces data subject to the distribution via sampling.

ML-based methods that can probabilistically generate the data come under this category for e.g. Naive Bayes, Bayesian Networks etc. deep generative methods have already shown good results in the field for information retrieval. [8] provides a good overview of few of these methods for information retrieval from texts and speech recognition.

IV. Literature survey

The topic of generative vs discriminative classifiers is quite an old one with unending debate for their superiority. We shall survey the literature in the context of the financial domain by dividing it into two parts:

- When the input comprises of historical quantitative data, such as price movements of stocks or currency. Inputs are well structured, and researchers have used technical and fundamental analysis for price prediction of stock, a group of stocks, currency etc. The survey focuses on the ML techniques that have been applied in quantitative finance to make the movement prediction.
- When the input comprises of unstructured text and sentiment analysis and text mining techniques are used to do the prediction task. Text can be from any source, e.g. social media blogs, News channels, annual reports etc.

A. Literature Survey: Structured Quantitative Input

For market prediction, classifiers based on supervised learning are profoundly used. Since in the field of quantitative finance there is no dearth of data, we hardly look for the generative properties of these classifiers. Still, there have been few researchers who have explored probabilistic properties of generative classifiers to model the instances that do not occur with high probability.

[9]compared multiple ML techniques to predict the movement direction of the market. He approached the problem in two ways, first without much of the pre-processing and later optimizing the data presentation to these techniques. In the second effort, generative NB (Multivariate Bernoulli) came out with best results. An observation that can be made from author's attempt is that both generative and discriminative model need pre-processing and if the input to them is given in an optimal way, their results can be improved to provide considerable accuracy in prediction.

Surveying the literature including various reviews for research in finance using computational methods, it was found that ANN followed by SVM are the most popular method that has been employed and researched. Various favours of ANN have been tried, and occasionally it belonged to generative class. In the past, few reviews have been published that exclusively covered ANN based approach for market prediction, e.g. [10] and other more generic reviews like [11] and [12] had a considerable focus on ANN based approaches by giving a number of layers used and other details of architecture used.

Though predominantly ANN applied in financial prediction are discriminative in nature, there are few probabilistic generative variants also available in the literature for e.g. [4] used daily prices and other financial indicators as input to Bayesian regularized ANN to predict the one-day future closing price of individual stocks.

B. Literature Survey: Unstructured Textual Input

Rationale behind the application of sentiment and text analysis is taken from behavioural economics which says news, blogs and social media comments may present a picture of sentiments of the larger world for a financial event or stock. The media do not report market status only, but they actively create an impact on market dynamics based on the news they release [13].

In case of textual data, the aim is to solve text classification problem: given a text; we try to predict a class based on the real-world continuous data associated with texts meaning. This may be in the form of positive, negative or neutral sentiments with varying degrees.

There are three primary sources which are used to analyze the broader sentiments.

i. News headlines/articles from financial news agencies and analysts. viz. media sentiments.
ii. Company annual reports press releases and corporate disclosures. viz. corporate sentiments.
iii. Social media comments, blogs viz. Twitter, Facebook comments, blogs of financial analysts viz. public sentiments.

Generative models are built to capture the interaction between all the variables of a system, in order to probabilistically synthesize possible classes. In the context of text mining, it describes how prospective each topic is, and how probable is the word given in the topic. This is how it says documents are generated by the word. A topic is the result of some distribution of words, and words arise because of the topic in the document. Generative models classify the document of words W into topic T by maximizing the joint likelihood: P(T, W) = P(W|T)P(T).

Discriminative model describes how likely a topic is when the set of words are given. It does not say anything about how likely the words or topic are by themselves. The task is to model P(T|W) directly such that T that maximizes this. Discriminative approaches are not concerned with P(T) or P(W) directly.

[14] observes SVM as the most used ML followed by NB for text mining and ANN, K-NN has not drawn much attention with regards to analyzing unstructured text data. In line with above observation[15] in his exhaustive survey showed that SVM and NB are the two most widely used approaches which are used for stock market prediction by processing texts. Notable reviews in text mining and its application to financial predictions are [14], [16] which talks about the impact of various sources of texts on financial forecasting, features selection techniques and algorithms used in the field. In both the reviews there is a mixed blend of algorithms from both generative and discriminative classes. Hjek [17] observes ANN with regularization and dropout and NB to outperform other methods and especially when higher dimensions were involved suggesting a high variance in text data.

The available input data is pre-processed to be fed into an ML algorithm. For the textual data, this means a transformation of subjective content to a representative objective form which can be processed by the classifier. We observe that there is greater acceptance and usage of generative class methods for text inputs in comparison to numerical inputs.

V. Conclusion

Data from literature re-affirms our understanding of implementation bias towards discriminative methods when quantitative, i.e. structured historical data is used in the financial domain. The primary reason for an inclination towards non-generative classifiers narrows down to the abundance of data for almost any granularity that may be required for predictions. Another reason can be theoretically computational expensiveness of generative methods which first work at generating the out of sample data and then works at prediction task.

Search for improved accuracy in prediction motivates researchers at exploring uncharted avenues. Newer methods that are extensively computational with the support of high performing hardware are being explored. ANN has been considered as a most sought-after method for employing in finance, and with deep learning, it has become the go-to approach. Independent generative methods employing Bayesian and probabilistic approach may have lost its shine in a financial context, but generative ANNs are still active and may show promising results. [4] used a Bayesian regularized network which assigns a probabilistic nature to the network weights, allowing the network to automatically and optimally penalize excessively complex models. Such usage of probabilistic methods reduces the potential for over-fitting and over-training, improving the prediction quality and generalization of the network.

Sentiment analysis is employed on texts that can be in the form of news, analyst reports or even social media comments. [14], [15] in their work claimed SVM and Naive Bayes classifiers to be most favoured methods by researchers, though methods like ANN, K-Nearest Neighbours (k-NN), fuzzy logic show promising potential for textual classification and sentiment analysis in other fields but are very under-researched in the context of market prediction.

An integral part of sentiment analysis is to make sense of the message being conveyed within the line. This becomes especially true with social media and news reports where there is no set pattern of text. Hence, it is more open to probabilistic methods in compare to numerical analysis which is carried out over enormous financial data where features used are derived from explicit formulas which use technical and fundamental concepts from financial engineering. Consequently, we observe more utilization of generative methods for deriving sentiments. In line with above observation [18] perceive that NB classifiers often perform well in practice for sentiment analysis especially for sentiment-polarity classification.

As evident from comments of various researches and survey of the literature, research in both numeric and textual field has grown in different directions owing to the property inhibited by input data. While research in the numeric field is quite mature and still searches for improvements, research on sentiment aspect is still young. Better results can be observed when instead of isolation, capabilities of the generative-discriminative pair are used. Another survey [19], though not specific to finance, classifies deep learning models as generative and discriminative and stresses on their complementary nature. This pairing can further be extended when both numeric and textual aspects are applied simultaneously. Not much of the work has been done in a financial domain where hybridization of two different approaches in forms of generative-discriminative pairs has been done and can be explored in future researches. The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

References

[1] K. Wang, C. Zong, and K.-Y. Su, "Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation," *ACM Trans. Asian Lang. Inf. Process.*, vol. 11, no. 2, pp. 1–41, 2012.

[2] A. Y. Ng and M. I. Jordon, *On Discriminative vs. Generative* Classifiers*: A comparison of logistic regression and naive Bayes*. MIT Press, 2002.

[3] P. M. Long, R. A. Servedio, and H. U. Simon, "Discriminative learning can succeed where generative learning fails," *Inf. Process. Lett.*, vol. 103, no. 4, pp. 131–135, 2007.

[4] J. L. Ticknor, "A Bayesian regularized artificial neural network for stock market forecasting," *Expert Syst. Appl.*, vol. 40, no. 14, pp. 5501–5506, 2013.

[5] L. Deng and N. Jaitly, "Deep Discriminative and Generative Models for Pattern Recognition," in *Handbook Of Pattern Recognition And Computer Vision (5th Edition)*, Fifth., C. C. Hau, Ed. World Scientific Press, 2015, pp. 1–26.

[6] T. S. B. Fletcher, "Machine learning for financial market prediction," p. 207, 2012.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag Berlin, Heidelberg ©2006, 2006.

[8] J. Xu, H. Li, and S. Zhou, "An Overview of Deep Generative Models," *IETE Tech. Rev.*, vol. 32, no. 2, pp. 131–139, 2015.

[9] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock *market* index using fusion of machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2162–2172, 2015.

[10] B. K. Wong and Y. Selvi, "Neural network applications in finance: A review and analysis of literature (1990–1996)," *Inf. Manag.*, vol. 34, no. 3, pp. 129–139, 1998.

[11] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market forecasting techniques - Part II: Soft computing methods," *Expert Syst. Appl.*, vol. 36, no. 3 PART 2, pp. 5932–5941, 2009.

[12] R. C. Cavalcante, R. C. Brasileiro, V. L. F. Souza, J. P. Nobrega, and A. L. I. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions," *Expert Syst. Appl.*, vol. 55, pp. 194–211, 2016.

[13] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," *Expert Syst. Appl.*, vol. 54, pp. 193–207, 2016.

[14] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.

[15] B. S. Kumar and V. Ravi, *A survey of the applications of text mining in financial domain*, vol. 114. Elsevier B.V., 2016.

[16] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artif. Intell. Rev.*, pp. 1–25, 2017.

[17] P. Hájek, "Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns," *Neural Comput. Appl.*, vol. 29, no. 7, pp. 343–358, 2018.

[18] Q. Li, J. Wang, F. Wang, P. Li, L. Liu, and Y. Chen, "The role of social *sentiment* in stock markets : a view from joint effects of multiple information sources," *Multimed. Tools Appl.*, 2016.

[19] J. *Padmanabhan* and M. J. Johnson Premkumar, "Machine Learning in Automatic Speech Recognition: A Survey," *IETE Tech. Rev.*, vol. 32, no. 4, pp. 240–251, 2015.

Authors Profile

Puneet Misra is an Assistant Professor of Computer Science in the Department of Computer Science at the University of Lucknow, Lucknow, U.P., India. He received bachelor degree (1995) in Physics and Maths and a dual Master's degree in Electronics and Computer Applications, and a PhD degree (2003) from the University of Lucknow. He is currently engaged in research areas which include Soft computing, Artificial Intelligent Systems, human-computer interaction and issues related to cybercrime and its prevention policies etc.

Siddharth is a dual masters in computer science. He received his MTech degree from BITS Pilani, India, in System Software and MCA degree from BHU, Varanasi, India. Presently, he is pursuing his doctoral studies in the Department of Computer Science at University of Lucknow, India. He has more than 13 years of experience in IT industry. His research interests include artificial intelligence, machine learning and data mining for time series data. He is particularly interested in the field of machine learning and its application to the field of finance.