Probability PSO Search Feature Selection for Data Stream Mining for current market trend

Snehal V.Rode[1], Ram B.Joshi[2]
[1]*Savitribai Phule Pune University,*
[2]*Professor, Savitribai Phule Pune University,*
Indira College of Engineering and Management, Pune, Maharashtra, India

***Abstract*— Big Data is a technology which is used to store and process the exponentially increasing dataset that contains information in structured way, semi structured way and mostly unstructured way. It deals with 3 V's: Volume, Variety and Velocity for processing of data. Particle Swarm Optimization (PSO) is a computational search and optimization method which is used to categorize and analysis of data based on feature selection. Feature selection has been popularly used to increase the processing load in inducing a data mining model. But, when the large database having high dimensionality come into picture features subset also increases tremendously in size, leading to an intractable demand in computation. In order to tackle this problem which is mainly based on the high-dimensionality and streaming format of data feeds in Big Data, a novel lightweight feature selection is proposed. The feature selection is designed particularly for mining streaming data on the fly, by using accelerated particle swarm optimization (APSO) type of swarm search that achieves enhanced analytical accuracy within reasonable processing time.**

***Index Terms*— Big Data, Particle Swarm Optimization, Feature Selection, Swarm Intelligence**

I. INTRODUCTION

People tend to invest in stock of a company due to its high rate of return. However, stock prices are continuously changing and most investors including experienced investors fail to understand the market trends. Investors are often influenced by crowd psychology and tend to go with majority decision of buying or selling stock which may be against the current market trend.

Hence, an effective financial system is required which can make objective decision of whether to buy or sell the stock. There are many prediction methods available for making stock prediction such as chartist method, regression analysis and neural networks. This project application uses the PSO for feature selection and data mining for modeling the stock because of its ability to learn from the given training data and then generalize a rule.

The principal strength with the network is its ability to find patterns and irregularities as well as detecting multi-dimensional non-linear connections in data. The model is trained using the historical daily stock quotes of at-least two years. The historical stock quotes can be fed directly through the Yahoo Finance Server using internet or from the local database. Once the model is successfully trained, the model can be used to predict the value of share in future. This allows the investors to reduce the risk of investing, and enables easier decision making resulting in maximum financial gain.

This survey paper points out the shortage that exists in current traditional statistical analysis in the stock, then makes use of PSO for feature trend selection and then uses Datamining algorithm to predict the stock market by establishing a three-tier structure of the neural network, namely input layer, hidden layer and output layer. After building the data pre-processing set before data mining, lots of widely used stock market technical indicators. Finally, we get a better predictive model to improve forecast accuracy.

In survey paper, data mining technology is applied to stock market in order to research the trend of price, it aims to predict the future trend of stock market and the fluctuation of price that are occur in price. The challenge rests on several stringent requirements in data stream mining: First of all, the amount of data feed is potentially infinite, and the data delivery is continuous like a high-speed train of information. The processing hence is expected to be real-time and

instantly responsive. Feature selection is a heuristic process which retains only the significant features as an optimal subset of the full features, representative enough to induce an accurate classification model for pattern recognition.

II. LITERATURE REVIEW

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behaviour of bird flocking or fish schooling.

There are traditional methods available for the feature extraction. However it is recently reported that many proposed methods are limited to one or more of the following constraints in their designs

1. The size of the resultant feature set is assumed fixed. Users are required to explicitly specify the maximum dimension for feature subset. Although the number of combinations reduces from 2k to k !/(k −s)! s! where there is a maximum of k features in the original dataset and s is the upper limit of the subset (for $s \leq k$), the major drawback is that users may not know in advance what would be the ideal size of s
2. The feature becomes minimal. By the principle of removing redundancy, the feature set may shrink to its most minimal size.
3. The feature selection methods are custom designed for some particular classifier and optimizer. Although an exhaustive computing method may be used for finding the most appropriate feature subset, this is quite impractical for data streams which are usually in very high dimensions and their amount may accumulate to infinity.

There is large amount of heart related data present, which is in unstructured format. Hence by analyzing the data and formatting it into structured manner helps for making the decision. For diagnosing the disease there are many ways in which heart related diseases can be diagnosed and treatment can be provided. Different approaches have different aspects in diagnosing the diseases. By using the neural network approach the accuracy secured was around 80- 90% but the hidden layers description cannot be evaluated. In fuzzy logic approach the weighted rules are generated initially and then the fuzzy rule decision is provided and the accuracy obtained id around 79.05%. In naive bayes classification approach helps in predicting whether the patient is prone to heart disease or not and depicting the risk factor for heart attack. The accuracy observed for naive bayes approach was around 90%. Similarly by using Support vector machines concept the accuracy was achieved around 80% approx. While as by using particle swarm approach the accuracy is increased.

III. PROBLEM SPECIFICATION

Till now PSO system is exist only for the ECG system but as the people are more interested in investment of stock so by taking stock dataset the PSO for feature selection and data mining is done for modelling the stock because PSO will increase the feature selection of data so that people will predicate more accurately where the market will go in future and what are the areas where people can invest. Here main challenge is amount of data feed is tremendous and also data delivery require at high speed hence processing is expected to be real time and instantly responsive.

IV. PROPOSE METHODOLOGY

Stock market prediction is an important area of financial forecasting, which is of great interest to stock investors, Still now PSO system is exist in ECG system but here PSO is used regarding prediction of stock market, stock traders and applied researchers. Main issues in developing a fully automated stock market prediction system are: feature extraction from the stock market data, feature selection for highest prediction accuracy, the dimensionality reduction of the selected feature set and the accuracy and robustness of the prediction system. In this paper, an PSO, B-PSO decision tree-adaptive neuro-fuzzy hybrid automated stock market prediction system is proposed. The proposed system uses technical analysis (traditionally used by stock traders) for feature extraction and decision tree for feature selection. Dimensionality reduction is carried out using fifteen different dimensionality reduction techniques. The dimensionality reduction technique producing the best prediction accuracy is selected to produce the reduced dataset. The reduced dataset is then applied to the adaptive neuro-fuzzy system for the next-day stock market prediction.

V. ARCHITECTURE

Pre-Processing Filter

The Data collected by the sensor is filtered using High Pass and Low Pass Filter. The filtered data is further given for Feature extraction Process

Feature Extraction Normalization Process

The Filtered data is normalized using the below feature extraction methods

Standard deviation of the NN

The simplest variable to calculate is the SDNN that is the square root of variance. Since variance is mathematically equal to total power of spectral analysis, SDNN reflects all the cyclic components responsible for variability in the period of recording. In many studies, SDNN is calculated over a 24 hours period and thus encompasses both short-term high frequency variation, as well as the lowest frequency components seen in a24-hours period, as the period of monitoring decreases, SDNN estimates shorter and shorter cycle lengths. It should also be noted that the total variance increases with the length of analysed recording. Thus SDNN is not a well-defined statically quantity because of its dependence on the length of recording period. Thus, in practice, it is inappropriate to compare SDNN measures obtained from recordings of different durations. A short-term recording are used in this work. Calculation of standard deviation is below shown in equation.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}, \quad (1)$$

Where { $x_1, x_2 \ldots, xn$} is a sample. The denominator N-1is the no of degrees of freedom in the vector

Standard deviation of differences between adjacent NN intervals

The most commonly used measures derived from interval differences include the standard deviation of differences between adjacent NN intervals. Calculation of standard deviation is show in above equation.

Root mean square successive difference of intervals

The most commonly used measures derived from interval differences include the square root of the mean squared differences of successive NN intervals. Calculation of root mean square is show in equation.

$$x_{\text{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} \quad (2)$$

Equation III.I: equation for calculating X_{rms}

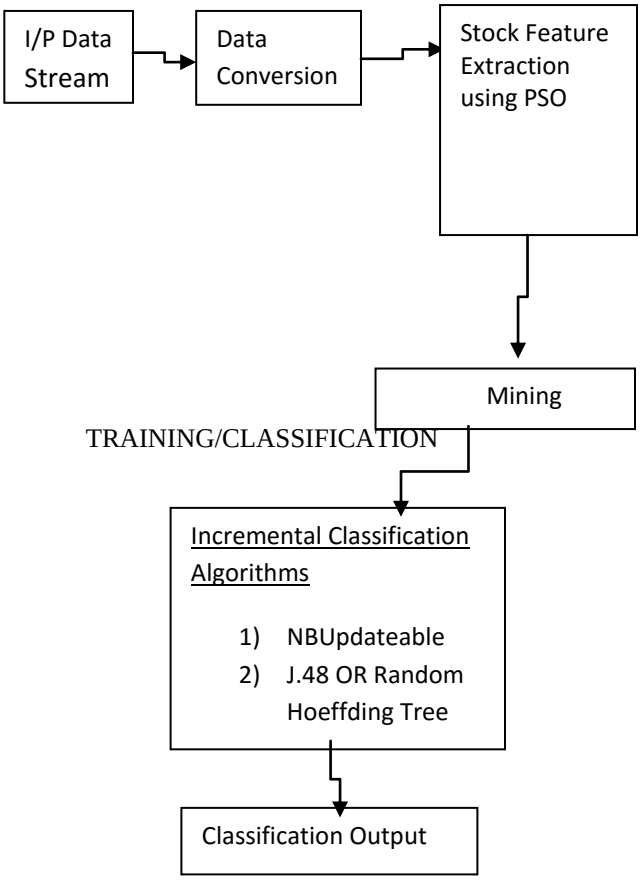

Figure I: System Architecture for PSO

VI. APPLICATION

Investment in stock of a company is due to its high rate of return. However, stock prices are continuously changing and most investors including experienced investors fail to understand the market trends. So this system can be used to predict the value of share in future. This allows the investors to reduce the risk of investing, and enables easier decision making resulting in maximum financial gain. This model will help in making decision whether to buy or sell the stock.

VII. CONCLUSION

Previously PSO is used for feature selection in the ECG system but with this PSO and incremental classification algorithm feature selection can be calculated for stock market so that investor will understand the market trends using PSO we can more accurately specify the value of share in future so accordingly people can decide whether to buy or sell the share.

REFERENCES

[1] Ping-Feng Pai, Tai-Chi Chen, "Rough set theory with discriminant analysis in analyzing electricity loads", Expert Systems with Applications 36 (2009), pp.8799–8806

[2] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data IEEE TRANSACTIONS ON JOURNAL NAME, MANUSCRIPT ID

[3] S. Ray, R.H. Turi, Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation, Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99) , Calcutta, India, 137-143, 1999.

[4] Hart, P. E., Nilsson, N. J., & Raphael, B., 1968. A Formal Basis For The Heuristic Determination Of Minimum Cost Paths. Systems Science and Cybernetics, IEEE Transactions on, 4(2), 100-107.

[5] Hereford, J. M. (2006). A distributed particle swarm optimization algorithm for swarm robotic applications. Paper presented at the Evolutionary Computation, 2006. CEC 2006. IEEE Congress on.

[6] Eberhart, R., & Kennedy, J. (1995). A new Optimizer using particle swarm theory. Paper presented at the Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on.

[7] S. Ray, R.H. Turi, Determinati on of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation, Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Calcutta, India, 137-143, 1999.

[8] I.H. Witten, E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann (2005), J.S Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” *Neurocomputing—Algorithms, Architectures and Applications*, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: SpringerVerlag, pp. 227-236, 198

[9] I.H. Witten, E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufmann (2005), J.S Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” *Neurocomputing—Algorithms, Architectures and Applications*, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: SpringerVerlag, pp. 227-236, 1989.