JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A SYSTEM FOR STOCK MARKET ANALYSIS AND REAL TIME PREDICTION USING SUPERVISED LEARNING TECHNIQUES

Devi.P.P[1], Aarthy.N[2]

pp.devi@aalimec.ac.in[1], aarthy.n@aalimec.ac.in[2]

Asst.Professor[1,2],Department of Computer Science and Engineering[1,2]

Aalim Muhammed Salegh College Of Engineering[1,2], Chennai[1,2], India[1,2]

Abstract : Generally, predicting how the stock market will perform is one of the most difficult things to do. It can be described as one of the most critical process to predict that. This is a very complex task and has uncertainties. To prevent this problem in One of the most interesting (or perhaps most profitable) time series data using machine learning techniques. Hence, stock price prediction has become an important research area. The aim is to predict machine learning based techniques for stock price prediction results in best accuracy. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict the stock price Index value by prediction results in the form of stock price increase or stable state best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given dataset with evaluation of GUI based user interface stock price prediction by attributes. dataset with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

IndexTerms - Stock Price Prediction, SMLT, Dataset

I.INTRODUCTION

A stock or share (also known as a company's "equity") is a financial instrument that represents ownership in a company or corporation and represents a proportionate claim on its assets and earnings. Stock and financial markets tend to be unpredictable and even illogical. Due to these characteristics, financial data should be necessarily possessing a rather turbulent structure which often makes it hard to find reliable patterns. Modelling turbulent structures requires machine learning algorithms capable of finding hidden structures within the data and predict how they will affect them in the future

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. They learn from previous computations to produce reliable, repeatable decisions and results.

When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. In this project we compare various Supervised machine learning algorithms and select the best one out of them, The most important factor of this project is to compare the performance of various Supervised machine learning algorithms from the given dataset and evaluate their accuracy of all the algorithms and will use that algorithm to predict whether the price will increase or not and with the help of GUI it will easy to input the sample data.

II.EXISTING SYSTEM

They are used Fuzzy rough theory can describe real-world situations in a mathematically effective and interpretable way, while evolutionary neural networks can be utilized to solve complex problems. Combining them with these complementary capabilities may lead to evolutionary fuzzy rough neural network with the interpretability and prediction capability. Their modified FRNNs were utilized to predict the stock price. To optimize the modified FRNN models, we encoded the structure, MFs, and parameters as variables, constructing FRNN models with multi objective optimization by simultaneously considering the prediction precision and network simplicity as objectives.

Therefore, the integration of rough neurons, the consequence node enhancement, and the application of the interval type-2 fuzzy set were beneficial to performance improvement. Through experimentation, we verified that the proposed PCMLIA-ADE was superior to the other algorithms. Additionally, via comparison to the conventionally optimized LSTM network, the FRNNs generated through the evolutionary framework were more precise and interpretable. To search for the FRNN models with low prediction errors and simple structures, on the basis of PCMLEA, the optimizer of SHADE was utilized to replace JADE, denoted as PCMLEA-SHADE. Then, another version of PCMLEA without the crossover operator was proposed, denoted as PCMLEA-NO-XOR. Moreover, based on PCMLIA, by replacing SBX with DE and utilizing adaptive parameters, PCMLIA-DE and PCMLIA-ADE were presented, respectively. Through inherent distributed parallelism, the optimization time was greatly reduced.

III.PROPOSED SYSTEM

Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy. Supervised learning uses a training set to teach models to yield the desired output. By Comparing the models like logistic regression, decision tree algorithms, random forest, Naïve bayes, K nearest neighbor and SVM, The algorithm with best accuracy is used to predict the stock market price. The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done. These reports are to the investigation of applicability of machine learning techniques for stock price prediction in operational conditions.

A. ARCHITECTURE DIAGRAM

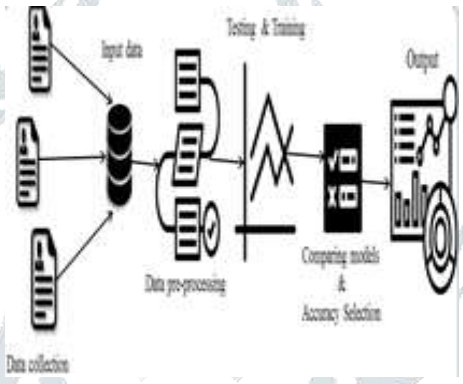

fig 1: architecture diagram

The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict the stock price Index value by prediction results in the form of stock price increase or stable state best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation of GUI based user interface stock price prediction by attributes.

B. DATA VALIDATION PROCESS

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Importing the library packages with loading given dataset. To analysing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process.

C. EXPLORATION DATA ANALYSIS OF PRE-PROCESSING AND VISUALIZATION PROCESS

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end. Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots.

D.ACCURACY CALCULATION

False Positives (FP):A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP):A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN):A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

E. PREDICTION RESULT BY ACCURACY

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

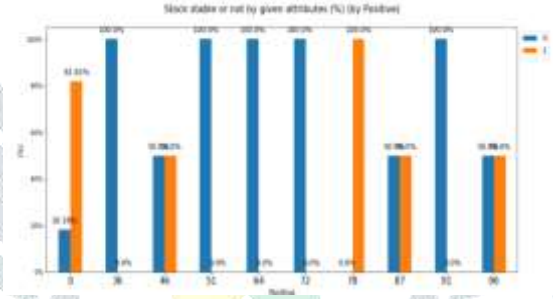

fig 2: visualization - histogram

F. ACCURACY

The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct. (When the model predicts default: how often is correct?)

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

- Recall = TP / (TP + FN)
- Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score :is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

fig 3: gui output

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

IV CONCLUSION

In this paper various supervised machine learning techniques like logistic regression, decision tree, random forest and support vector machine were compared with the dataset and the algorithm with the best accuracy will be selected, the accuracy will be calculated based on True positive, True negative, False positive and False negative of the particular algorithm and sample data will be fed into the application and with the help of GUI the organization can easily find whether their stock will rise or not.

REFERENCES

[1] Bin Cao , Member, IEEE, Jianwei Zhao , Zhihan Lv , Senior Member, IEEE, Yu Gu , Member, IEEE,Peng Yang , and Saman K. Halgamuge , Fellow, "Multiobjective Evolution of Fuzzy Rough NeuralNetwork via Distributed Parallelismfor Stock Prediction" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 28, NO. 5, MAY 2020.

[2]Xianghui Yuan, Jin Yuan, Tianzhao Jiang, and Qurat UL Ain ,"Integrated Long-term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market",Received December 22, 2019, accepted January 16, 2020, date of publication January 24, 2020, date of current version February 6, 2020.

[3]Xie Chen, Deepu Rajan, Chai Quek,"A Deep Hybrid Fuzzy Neural Hammerstein-Wiener Network for Stock Price Prediction", Jul 2020 International Joint Conference on Neural Networks (IJCNN), MAY 2019.

[4]Jiannan Chen, Junping Du, Feifei Kou,"Prediction of Financial Big Data Stock Trends Based on Attention Mechanism" Stock Market One-day ahead Movement Prediction Using Disparate Data Sources [J]. Expert Systems with Applications, 2017, 79(2): 153–136.

[5]RahmaFirsty Fitriyana, Brady Rikumahu, Andry Alamsyah ,"Principal Component Analysis to Determine Main Factors Stock Price of Consumer Goods Industry"IEEE Trans. Fuzzy Syst., vol. 27, no. 7, pp. 1347–1361, Jul. 2019.