Original Article

HIERARICAL CLUSTER BASED DATA PRE-PROCESSING FOR STOCK MARKET DATA PREDICTION

Mrs. D. Gokila[1], Dr.B.Azhagusundari[2]

[1] Research scholar, Department of computer science, NGM college, Tamilnadu, India.
[2] Assosciate professor, Department of computer science, NGM college, Tamilnadu, India.

DOI: 10.47750/pnr.2022.13.S08.389

Abstract

Financial area analysis are not limited to enterprise performance analysis. It merits examining as wide an area as conceivable to get the full impression of a particular enterprise. Stock market dataset content is a datum source that offers the prediction data of the ups and downs of growth in stock market, trading tasks, daily and timely status, and so on. Consequently, it merits investigating the news entry of up to date data. Mining the data and forecasting the data will be a challenging task due to huge volume of data , and doesn't give high precision. To beat this shortcoming, another equal data pre-processing algorithm in view of Hierarchical Clustering is proposed in this paper. This algorithm can decrease the size of information and runtime. This research using the proposed model will provide the best solution. The examination demonstrates the performance of our proposed preprocessing algorithm is better than existing.

Keywords: Stock market, Business decision, Dynamic environment, Forecasting, Optimization.

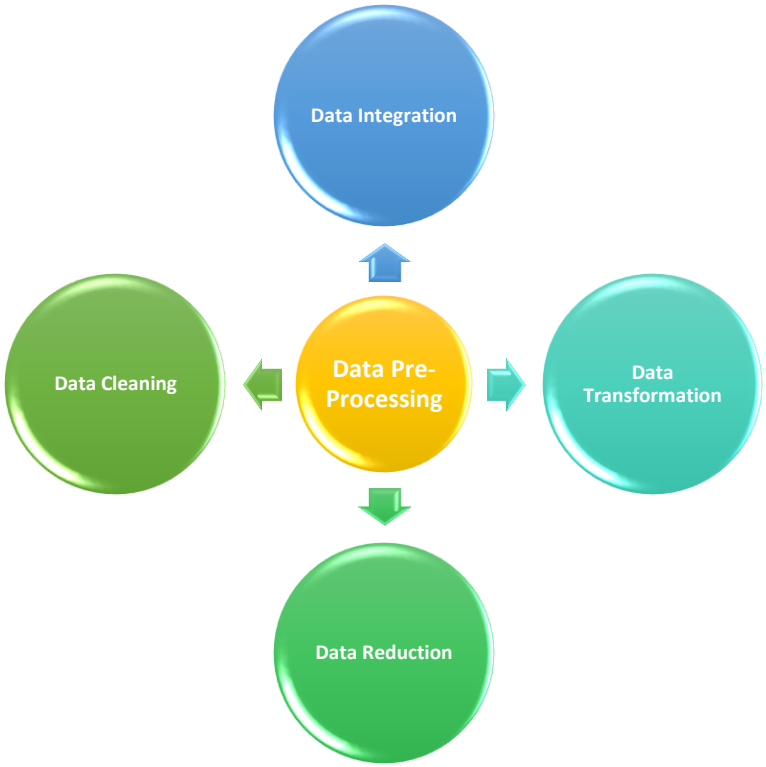
Introduction

One of the all time application in the exploration field is the stock market forecasting. As this is dynamic stage, despite the fact that number of endeavors have been made the disclosure of an exact and effective arrangement is as yet a test. The forecast trouble is straightforwardly relative to the models of the market dynamics. The examination field of stock market is exemplified with two convictions; Fundamental and Technical methodologies. The critical thought of central analysis lies in forecasting stock market development got from security's historical data, which is numeric, the idea of specialized analysis use the demonstrating techniques and diagrams to anticipate future stock patterns and values. Notwithstanding the above methodologies, Natural Language Based Financial Forecasting (NLFF) incorporates the literary data analysis thought about as the convenient produced data reports, news, public voice in web-based entertainment and web journals essentially influence the cost and pattern of a security share. The arrangement of assignments engaged with message analysis for sentiment extraction is: predefine the application pertinent arrangement of watchwords, distinguishing the appropriate machine learning technique for sentiment analysis.

Genuine data is untidy and is many times made, handled and put away by an assortment of people, business processes and applications. Thus, a data set might be missing individual fields, contain manual info blunders, or have copy data or various names to depict exactly the same thing. People can frequently distinguish and correct these issues in the data they use in the line of business, however data used to prepare machine learning or profound learning algorithms should be naturally pre-handled. A decent data pre-processing pipeline can make reusable parts that make it simpler to try out different thoughts for smoothing out business processes or further developing

consumer loyalty. For instance, preprocessing can further develop how data is coordinated for a proposal motor by further developing the age ranges utilized for sorting clients.

Figure 1. Data Pre-Processing



Preprocessing can likewise improve on crafted by making and changing data for more precise and designated business knowledge experiences. For instance, customers of various sizes, classes or districts might display various ways of behaving across areas. Preprocessing the data into the fitting structures could help BI groups mesh these experiences into BI dashboards. In a Customer Relationship Management (CRM) setting, data preprocessing is a part of web mining. Web utilization logs might be preprocessed to extricate significant arrangements of data called client exchanges, which comprise of gatherings of URL references. Client meetings might be followed to recognize the client, the sites mentioned and their request and the timeframe spent on every one. Once these have been pulled out of the crude data, they yield more valuable data that can be applied, for instance, to buyer exploration, marketing or personalization.

Stock Market

A stock market or security exchange or offer market or value market is the stage, where the economic exchanges were made by the arrangement of stock holders. At first the stocks ought to be enrolled under the exchange, implies stocks went into essential market. These protections can consider trading in secondary market. The stock market is free organization of e-trading done by the aggregate individuals known as purchasers and merchants. Stocks can be recorded under exchange in various ways, for the most part by the nation, where the organization is dwelled. A stock can be recorded under more than one index or exchange and can be exchanged across the world. Stocks can be arranged in various ways, in view of the sector, in light of the capital sum and so forth. At the point when the two gatherings partook in trading, the exchange activity is finished, when they settle on a price.

The trade request execution is done either by a stock exchange dealer for stock holder or straight by the stock holder through the stock office gave stage. The trouble with stock price assessment will stays same, assuming the choice of expectation calculation is inappropriate. The demonstration of assessing the past or future stock market pattern/stock price is named as stock market expectation. The occupation of stock market expectation tenders a

thought for the stock holders, can be utilized to realize the data in regards to stocks they hold and helps them in assessing the market prices and developments. The fundamental thing to be note in forecast is shrewd choice of variables list, as information and appropriate machine learning model leads the best results.

Data pre-processing

Data pre-processing, a part of data readiness, portrays any kind of processing performed on crude data to set it up for another data processing procedure. It has generally been a significant starter venture for the data mining process. All the more as of late, data pre-processing techniques have been adjusted for preparing machine learning models and AI models and for running inferences against them.

Data pre-processing changes the data into an organization that is all the more effectively and successfully handled in data mining, machine learning and different data science undertakings. The techniques are for the most part utilized at the earliest phases of the machine learning and AI advancement pipeline to guarantee accurate results.

Literature Survey

Mr. Rupesh and A. Kamble (2017) et.al proposed Short and Long Term Stock Trend Prediction using Decision Tree. First fair of this exploration is to smooth out the stock price pattern expectation for momentary using a couple of oscillators and pointers: Moving Average Convergence Divergence (MACD). It is seen that using appropriate pre-processing technique and Machine learning model, it is doable to additionally foster precision speed of transient pattern expectation. Applying Pre-processing and afterward using mix of data can yield an unrivaled Accuracy rate in Short term Trades, while predicting for Long-term Trend of Stock this Technical markers are not satisfactory. Alongside a piece of this Technical data and Fundamental Data of the association, expecting Long term stock development is doable. For Long term Prediction its Debt to Equity, Net profit of pervious long term, Promoters holding, Dividend yield and PE proportion is used alongside Technical Factors. It is seen that using Fundamental and Technical Data, Long term Stock Prediction is Possible.

Kiaa et.al proposed a prediction mixture model for calculation of direction of stock movement by considering global markets along with historical data. The creators named their created model as hybrid supervised semi-supervised model (HyS3), as it incorporates semi endlessly supervised approaches. The semi supervised approach is for addressing the co operations of the objective market with the worldwide market, through a name proliferation, name spreading ideas and graphical portrayal utilizing graph based semi supervised learning(GSSL) network, created by the method for consistent Kruskal-based graph construction (ConKruG) calculation. Graph based semi supervised expectation is utilized for anticipating the names, is called as mark spreading. The supervised methodology, SVM is utilized to anticipate historical stock data. Both the objective as well as worldwide market's historical data took care of to SVM model. Exactnesses are determined and thought about by creators, against the singular models. The proposed hybrid model gives the exceptional performance.

Ping-Feng Pai and Wan-Ru Wei et.al proposed Predicting Movement Directions of Stock Index Futures by Support Vector Models with Data Pre-processing. On account of the idea of high-influence, liberal remuneration can be obtained by minimal capital endeavor. As such, examination of prospects prices turns out to be maybe the most intriguing focuses concerning financial market. Lately, by applying the construction hazard minimization rule, support vector machines (SVM) move toward has been one of the most power techniques to overseeing classification issues. In this assessment, trading information including specific pointers is used by SVM model to expect improvement headings of Taiwan stock index prospects prices. In view of data pre-process affects forecast exactness of SVM models; pre-handled data gives by different strategies are used to examine influences on expectation execution of SVM models. Exploratory results reveal that the SVM approach has the best exhibition when data are handled by scaling and differencing exercises. Along these lines, scaling and differencing are two fundamental data pre-processing techniques to SVM and BPN models when classification issues with various information credits are addressed.

Chi Ma (2019) et.al proposed The Hybrid Dynamic Stock Forecasting Model Based on ANN and SVR. As of late, mass hybrid time series models have yet been applied to settle intricacy nonlinear and financial issues. Incidentally, a few hybrid models have limits, and the exactness of free forecasting model ought to be gotten to the next level. There are three responsibilities in the examination. Most importantly, highlight processing and normalization for time series in data pre-processing is fundamental for the forecasting. Second, the difference in dynamic weighted inclination can deal with the exactness of hybrid forecasting model. Third, the sensible model mix has favoured robustness over the single prediction model. The fitting data pre-processing is finished to deal with the nature of the element and Applies ANN and SVR to check the variance of stock. It deals with the precision of prediction by dynamic change inclination. To survey the proposed hybrid and dynamic model, we used China stock exchange data as the datasets which from June 8, 2015 to May 26, 2016.The results clearly show that the offered model causes result to show up at a more careful prediction, and the precision can be achieved 79%.

Dmitry V and Zhora et.al proposed Data Pre-processing for Stock Market Forecasting using Random Subspace Classifier Network. Financial forecasting is a troublesome issue which attracts specialists from different fields. Since numerous makers propose new techniques to anticipate the market, it's reasonable to expect that the strong game plan is subtle. Furthermore, found consistency in the market lead shouldn't exist for a long time period, considering the way that market individuals will endeavor to make the most of that open door. To foresee the future with some degree of assurance it's more intelligent to guarantee the forecasting technique gives guessed that results should different time frames. These article assessments the utilization of irregular subspace classifier for anticipating the next day stock cost return. Different data pre-processing draws near, particularly for stock cost normalization, are suggested. Forecasting execution is tested for different time spans. Besides, the limit of the association to anticipate the cost change is viewed as inside the test set.

Proposed Methodology

The steps used in data pre-processing include the following:

Data profiling

Data profiling is the method involved with inspecting, analyzing and reviewing data to gather measurements about its quality. It begins with a study of existing data and its qualities. Data researchers distinguish data sets that are relevant to the main pressing issue, stock its huge traits, and structure a hypothesis of highlights that may be pertinent for the proposed examination or machine learning task. They additionally relate data sources to the important business ideas and consider which pre-processing libraries could be utilized.

Data preparation provides the extracting data for the feature selection. The scoping of characteristics conducts feature selection in the extracting dataset, which assigns the next stage a profile. Third stage performs feature combinations by incorporating distribution of attribute values to get the attribute solution sets with accuracy. The filtering of missing number compares the performance of individual solution to obtain the finished dataset by determining the optimal subset and missing number. It is used to provide the data for further model of poverty rating.

Data reduction

Raw data sets frequently incorporate excess data that emerge from describing peculiarities in various ways or data that isn't pertinent to a specific ML, AI or examination task. Data reduction utilizes techniques like head part analysis to change the crude data into an easier structure reasonable for specific use cases.

Data transformation

Here, data researchers contemplate how various parts of the data should be coordinated to seem OK for the objective. This could incorporate things like organizing unstructured data, consolidating striking factors when it seems OK or distinguishing significant reaches to zero in on.

Data enrichment

In this progression, data researchers apply the different element engineering libraries to the data to impact the ideal transformations. The outcome ought to be a data set coordinated to accomplish the ideal harmony between the preparation time for another model and the required compute.

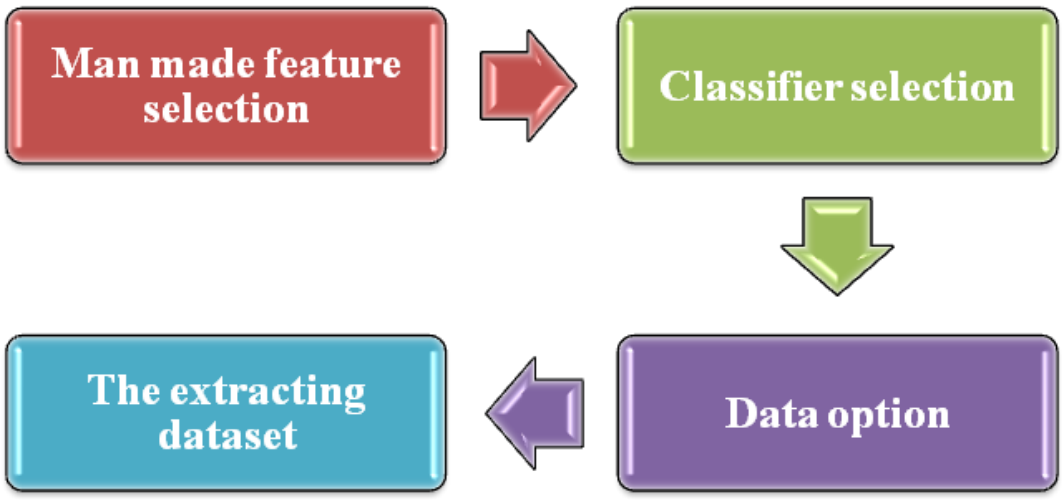
Data validation

At this stage, the data is parted into two sets. The main set is utilized to prepare a machine learning or profound learning model. The subsequent set is the testing data that is utilized to measure the exactness and robustness of the subsequent model. This subsequent advance recognizes any issues in the hypothesis utilized in the cleaning and element engineering of the data. In the event that the data researchers are happy with the results, they can push the pre-processing errand to a data engineer who sorts out some way proportional it for creation. In the event that not, the data researchers can return and make changes to the manner in which they executed the data purifying and highlight engineering steps.

Data provisioning

Manual feature selection, classifier choosing, data opting. Manual feature selection first exports data from the database, after that, with the principle of irrelevant features deleted, important features are selected artificially from the primitive features, and the original data are formed out of the essential feature in the exported data. Judged by correctness, classifier choosing takes the highest accuracy model in the original data, and the feature selection takes it to learning. In order to facilitate the experiment, the picking data uses the reduced loss count, in which the range is determined by reference to expert experience, and the data is selected from the original data as the extracting dataset based on the maximum missing number.

Figure 2. Data provisioning



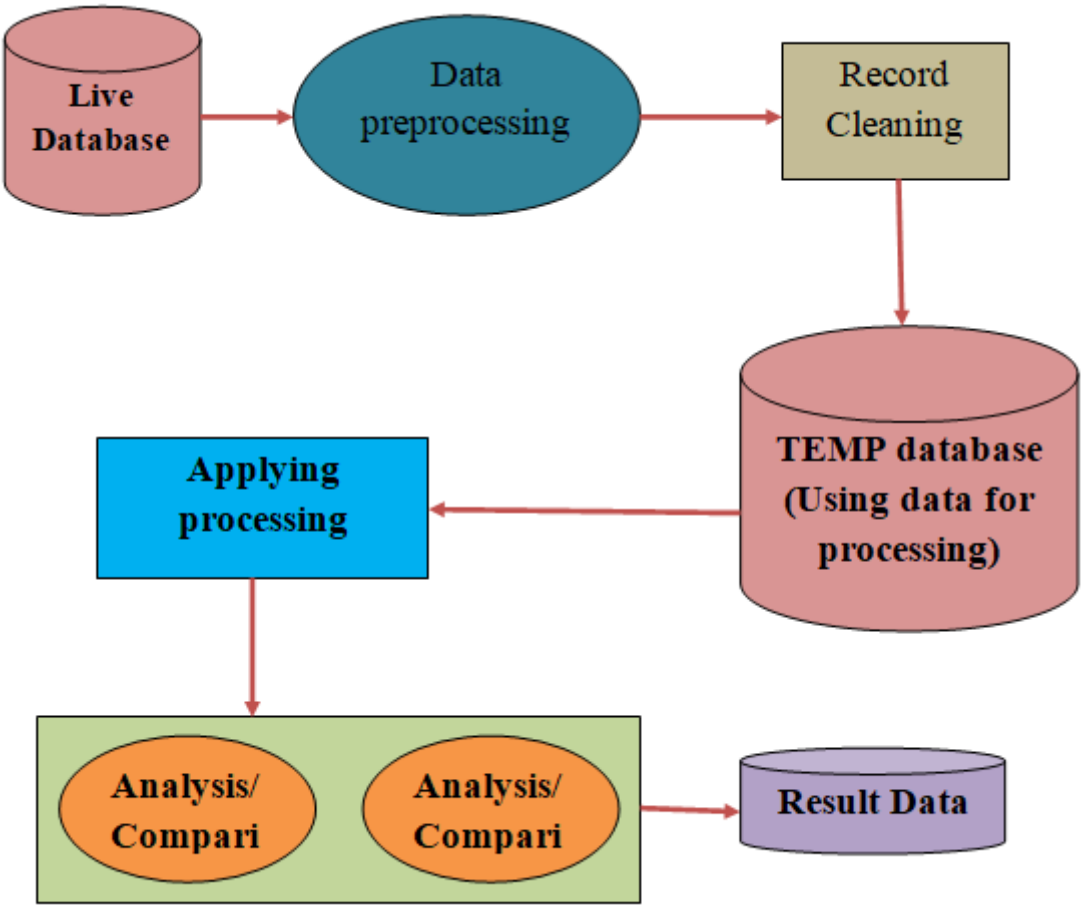
Missing number screening

The filtering of missing number starts with adjusting the classifier parameters and takes out the solution from the attribute solution sets sequentially to form a new dataset. Then it gets subsets of data with different missing number from this dataset by iterating over the range of missing number. Besides, a threshold is defined to determine the maximum number of missing for each solution. Comparing the maximum missing number of solutions, the optimal number is obtained to delete some of the solutions. Lastly, the best dataset is obtained with accuracy.

Characteristic combining

The characteristic combination sorts the attributes according to the degree of uniform distribution on attribute values. The uniform attribute set is obtained by selecting attributes with a threshold, and the uniform combined set is generated by deleting repeated features from the uniform attribute set, depending on a subset of features. Besides, it divides the transitional feature sets into a coarse subset, which basically represents the extracting dataset, and a refined subset, which represents it to a higher degree. When the added number is determined, the uniform combined set of corresponding is freely composed to derive the feature combinatorial set by its number, iterating over them to append their element to the initial solution to form a new set of features, and after iteration, the best feature set is added to the optimal feature sets by comparing all the feature sets obtained. The first stage is the lower bound, and the second stage is the upper bound. It is used to identify the subsets of feature in the two-stage optimal feature sets as the attribute solution sets under a threshold, and the process uses the dataset with the smallest missing number.

Figure 3. Real-time data cleaning



Data Cleaning or Data Cleansing

Data cleaning is part of data pre-processing. Data pre-processing has many activities one of it is data cleaning. Imperfect, incorrect, Incomplete, inaccurate or irrelevant parts of the data are identified in data cleaning process. These types of dirty data can be replace, modify or delete by the specific techniques. Data cleaning is also called data cleansing. Following are the steps for the data cleaning process;

- Select datasets
- Merge datasets into one datasets (if required)
- Identify Errors.
- Standardize process
- Scrub for duplicates
- Validate accuracy

Data cleaning can be achieved by many ways like adjusting the missing values, removing the duplicate row or removing the unwanted column.

Dimension Reduction

It removes redundant features

- Step by step Forward Selection
- Step by step Backward Selection
- Combination of forwarding and Backward Selection

Data Integration

Data integration means merging the two or more datasets in to one data set. Some of the application generates the database based on time interval; it requires merging if we want to process all the data at a time. For example, financial account system may generate the data yearly but if we want to perform analysis on 10 years then it requires merging 10 years dataset into one dataset that is called data integration. It also includes the process of merging data from dissimilar sources into a distinct, unified view. It integrates data at single place which are coming from multiple places. It may require to data conversion process to make unified format for each data.

Hierarchical clustering

There are two major methods of clustering -- hierarchical clustering and k-means clustering. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. The basic principle behind hierarchical clustering is as follows: if there are n input points or data items, we start with n clusters where each cluster has a single point. From there on, the "closest two clusters are identified. The two closest clusters are merged, resulting in a reduction in the number of remaining clusters is q where q is the target number of clusters and could be a part of the input.

Algorithm : Hierarchical clustering data preprocessing Algorithm

Begin

Input $n; c; \delta$

Build n *linked list and head of every; list denote an data point* $d(d \in n)$

For $k = 1$ *to* c ***do***

Build adjacent matrix $D(k)$

end

For *all processors* P_i *Where* $1 \leq i \leq p$ ***do***

Search $D(k)_{i,j}$ *in the* $D(k)$

If $0 < D(k)_{i,j} \leq \delta$ ***then*** *append the list of I to the back of list of j*

Alter data of i *row and* i *column into an infinite value*

i *Alter data of j row and j column into an infinite value*

end

end

Merge the lists that have a same head of list

end

Experiment Results

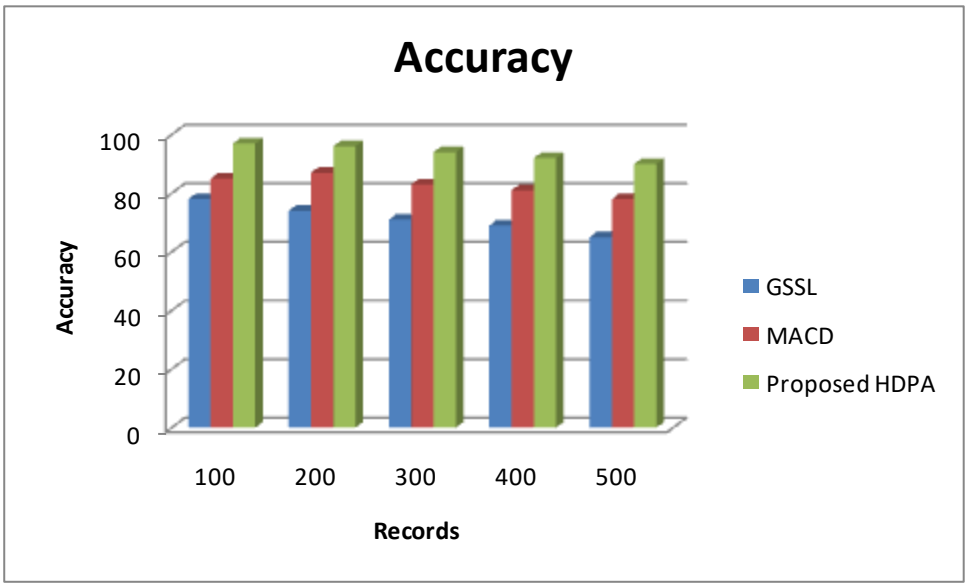
Accuracy

Table 1 Comparison Table of Accuracy

Records	GSSL	MACD	Proposed HDPA
100	78	85	97
200	74	87	96
300	71	83	94
400	69	81	92
500	65	78	90

The Comparison table 1 of Accuracy Values explains the different values of existing algorithms (GSSL, MACD) and proposed HDPA. While comparing the Existing algorithm (GSSL, MACD) and proposed HDPA, provides the better results. The existing algorithm values start from 65 to 78, 78 to 85 and proposed HDPA values start from 90 to 97. The proposed HDPA gives the great results.

Figure 4. Comparison chart of Accuracy



The Figure 4 Shows the comparison chart of **Accuracy** demonstrates the existing1, existing 2 (GSSL, MACD) and proposed HDPA. X axis denote the No. of Records and y axis denotes the Accuracy in percentage. The proposed HDPA values are better than the existing algorithm. The existing algorithm values start from 65 to 78, 78 to 85 and proposed HDPA values start from 90 to 97. The proposed HDPA gives the great results.

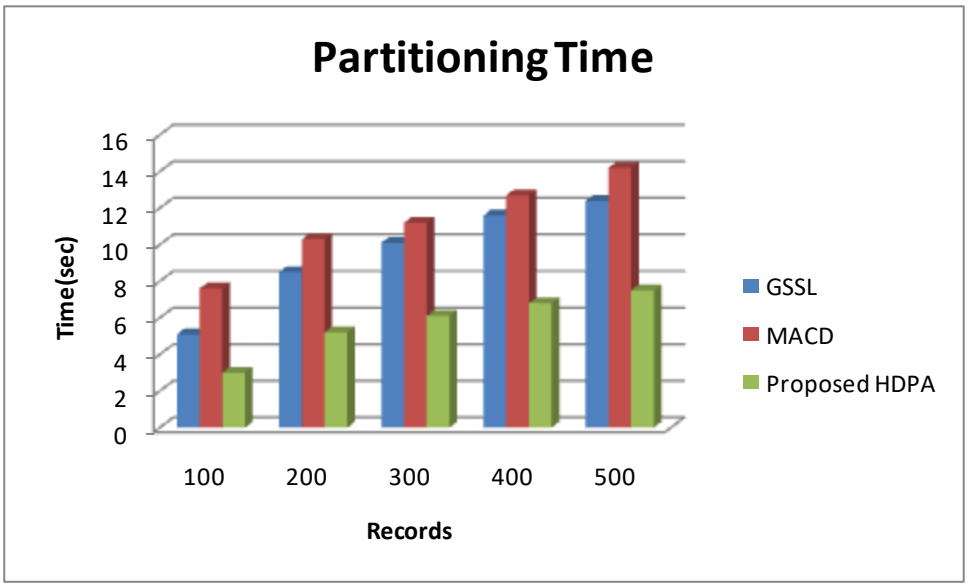
Partitioning time

Table 2. Comparison Table of partitioning time

Records	GSSL	MACD	Proposed HDPA
100	5.1	7.6	3.0
200	8.5	10.3	5.2
300	10.1	11.2	6.1
400	11.6	12.7	6.8
500	12.4	14.2	7.5

The Comparison table 2 of partitioning time Values explains the different values of existing algorithms (GSSL, MACD) and proposed HDPA. While comparing the Existing algorithm (GSSL, MACD) and proposed HDPA, provides the better results. The existing algorithm values start from 5.1 to 12.4, 7.6 to 14.2 and proposed HDPA values start from 3 to 7.5. The proposed HDPA gives the great results.

Figure 5. Comparison chart of partitioning time



The Figure 5 Shows the comparison chart of partitioning time demonstrates the existing1, existing 2 (GSSL, MACD) and proposed HDPA. X axis denote the No. of Records and y axis denotes the partitioning time in sec. The proposed HDPA values are better than the existing algorithm. The existing algorithm values start from 5.1 to 12.4, 7.6 to 14.2 and proposed HDPA values start from 3 to 7.5. The proposed HDPA gives the great results.

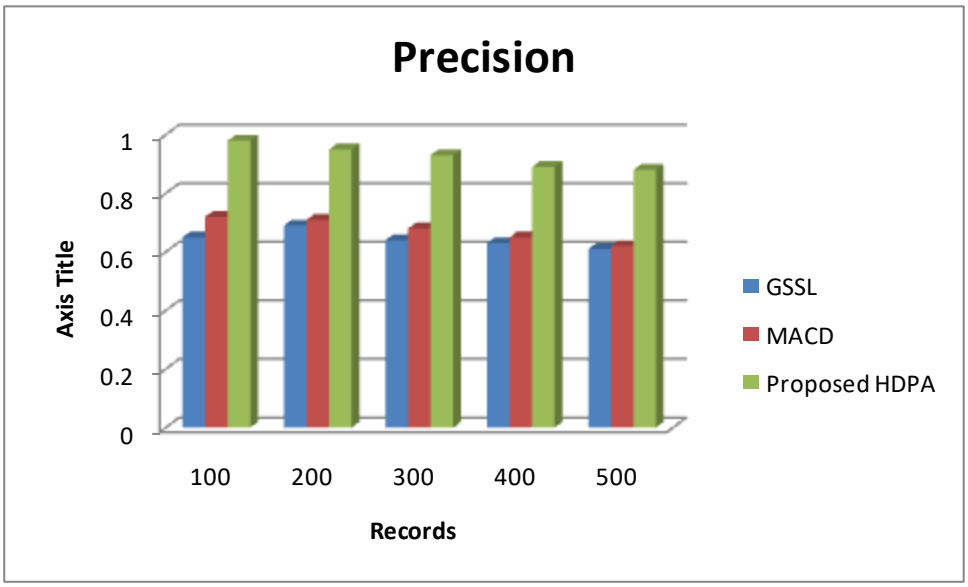
Precision

Table 3. Comparison table of precision

Records	GSSL	MACD	Proposed HDPA
100	0.65	0.72	0.98
200	0.69	0.71	0.95
300	0.64	0.68	0.93
400	0.63	0.65	0.89
500	0.61	0.62	0.88

The Comparison table 3 of Precision Values explains the different values of existing algorithms (GSSL, MACD) and proposed HDPA. While comparing the Existing algorithm (GSSL, MACD) and proposed CH-MFA, provides the better results. The existing algorithm values start from 0.61 to 0.65, 0.62 to 0.72 and proposed HDPA values start from 0.88 to 0.98. The proposed HDPA gives the great results.

Figure 6. Comparison chart of precision



The Figure 6 Shows the comparison chart of precision demonstrates the existing1, existing 2 (GSSL, MACD) and proposed HDPA. X axis denote the No. of Records and y axis denotes the Precision value. The proposed HDPA values are better than the existing algorithm. The existing algorithm values start from 0.61 to 0.65, 0.62 to 0.72 and proposed HDPA values start from 0.88 to 0.98. The proposed HDPA gives the great results.

Conclusion

This research focuses on pre-processing; flow research works in the area are for the most part focused on similar analysis of the customary AI and deep learning methods for remarks grouping. In this paper we proposed hierarchical based clustering for pre-processing the stock market data. Albeit the algorithm can successfully decrease the time and the size of the dataset of hierarchical clustering, another clustering algorithm, for example, the K-means algorithm likewise has a wide scope of utilizations, how to broaden the extent of the algorithm will be the subsequent stage to work. The proposed hierarchical cluster based preprocessing algorithm gives incredible results.

References

1. L. Zhao and L. Wang, "Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm," 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, 2015, pp. 93-98, doi: 10.1109/BDCloud.2015.19.

2. Z. Zhang, Y. Shen, G. Zhang, Y. Song and Y. Zhu, "Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 225-228, doi: 10.1109/ICSESS.2017.8342901.

3. R. A. Kamble, "Short and long term stock trend prediction using decision tree," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 1371-1375, doi: 10.1109/ICCONS.2017.8250694.

4. N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction," 2017 2nd International Conference for Convergence in Technology (I2CT), 2017, pp. 1199-1202, doi: 10.1109/I2CT.2017.8226316.

5. Ping-Feng Pai and Wan-Ru Wei, "Predicting movement directions of stock index futures by support vector models with data preprocessing," 2007 IEEE International Conference on Industrial Engineering and Engineering Management, 2007, pp. 169-173, doi: 10.1109/IEEM.2007.4419173.

6. U. Pasupulety, A. Abdullah Anees, S. Anmol and B. R. Mohan, "Predicting Stock Prices using Ensemble Learning and Sentiment Analysis," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 215-222, doi: 10.1109/AIKE.2019.00045.

7. M. Billah, S. Waheed and A. Hanifa, "Stock market prediction using an improved training algorithm of neural network," 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), 2016, pp. 1-4, doi: 10.1109/ICECTE.2016.7879611.

8. K. Ryota and N. Tomoharu, "Stock market prediction based on interrelated time series data," 2012 IEEE Symposium on Computers & Informatics (ISCI), 2012, pp. 17-21, doi: 10.1109/ISCI.2012.6222660.

9. S. Kalra and J. S. Prasad, "Efficacy of News Sentiment for Stock Market Prediction," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 491-496, doi: 10.1109/COMITCon.2019.8862265.

10. Z. Yixin and J. Zhang, "Stock Data Analysis Based on BP Neural Network," 2010 Second International Conference on Communication Software and Networks, 2010, pp. 396-399, doi: 10.1109/ICCSN.2010.12.

11. C. Ma, Y. Ning, H. Jin and J. Wu, "The Hybrid Dynamic Stock Forecasting Model Based on ANN and SVR," 2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS), 2019, pp. 715-718, doi: 10.1109/ICICAS48597.2019.00155.

12. B. S. Reddy, "Prediction of Stock Market indices — Using SAS," 2010 2nd IEEE International Conference on Information and Financial Engineering, 2010, pp. 112-116, doi: 10.1109/ICIFE.2010.5609262.

13. Dadabada Pradeepkumar, Vadlamani Ravi, Forecasting Financial Time Series Volatility using Particle Swarm Optimization trained Quantile Regression Neural Network, (2017), http://dx.doi.org/10.1016/j.asoc.2017.04.014.

14. T. Tantisripreecha and N. Soonthomphisaj, "Stock Market Movement Prediction using LDA-Online Learning Model," 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2018, pp. 135-139, doi: 10.1109/SNPD.2018.8441038.

15. Hegazy O, Soliman OS, Salam MA. A machine learning model for stock market prediction. arXiv preprint arXiv:1402.7351. 2014 Feb 28.