Critical Analysis of History Bits Algorithm with Attribute Ranking Strategies for Stock Market Prediction Accuracy Evaluation

Nitin Sakhare[1,*] ,Divya Midhun[2], Dharmesh Dhabliya[3]
[1]Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, India
[1]Lincoln University College, Malaysia.
[2]Lincoln University College, Malaysia.
[3]Department of Information Technology, Vishwakarma Institute Technology, Pune, India.
nitinsakhare4@gmail.com, divya@lincoln.edu.my, dharmeshdhabliya@gmail.com

Abstract: Prediction of stock market trading decision is a classical problem. Different machine learning models can be used to solve this prediction problem and help traders to get best return on investment. In an attempt to solve this prediction problem, we have developed a novel History Bits based algorithm. This algorithm accepts 75 technical indicators prioritized using attribute ranking strategies and transformed into buy and sell trading signals. All 75 indicators are partitioned into certain number of groups based on their rankings and bias is added as per their groupings. There are many attribute ranking strategies like Pearson correlation coefficient, information gain, gain ratio, OneR, relief evaluator and, symmetrical uncertainty evaluator. Different ranking strategies may produce different importance levels of attributes impacting the trader’s decision. In order to overcome this problem we propose an ensemble based attribute ranking strategy to collect combinatorial effect of all attribute ranking strategies. For the experimentation, a real dataset of NIFTY 50 index for a period of 20 years is used. Prediction accuracy of History Bits algorithm using different ranking strategies is compared. The proposed algorithm is quite stable with prediction performance over different attribute ranking strategies.

Keywords: Stock Market; Pearson correlation coefficient; information gain, gain ratio; OneR; relief evaluator; symmetrical uncertainty evaluator.

Introduction

The Chiaroscuro problem of stock market movement analysis and forecasting is not an easy issue to understand and study. Technical analysis is one of the most popular methods of making predictions about the future market trends among the other existing techniques. Technical indicators are technical indicators that are statistical calculations based on statistics of price, opening, and closing values of stock contracts, which is important in the analysis. Technical analysts use these indicators to predict the future trends of a price by analyzing past trends in a stock index.

Although intra-day traders use technical indicators to hold short-term changes in the market, long-term investors use them to determine the best moments of entering and leaving the market. Data cleansing and transformation are also required prior to the use of statistical or machine learning model to make predictions. These tasks involve missing data, deleting duplicate records, and normalizing to maintain consistency and reliability of the data [1].

In this research, dimensionality reduction is avoided deliberately in order to eliminate the risk of underfitting, which can take place in case important predictors are dropped. Rather than the extraction of redundant or irrelevant attributes by means of conventional feature extraction techniques, the study aims at addressing various types of attribute ranking approaches in order to rank the relevant features [2]. Many ranking and feature selection methods have been effectively used in a variety of applications in text mining, medical diagnostics, defense systems, agriculture and economics. Since the methods continue to acquire greater significance, research and interpretation of their efficacy in enhancing predictive performance becomes very important [3-5].

Pehlivanli, Baris and Gulay proposed a committee decision methodology which integrates a number of ranking techniques, including: t-statistics, Fisher score, ReliefF and Effective Range-based Gene selection, in order to improve the stability and interpretability of selected attributes. In a comparable manner, Drotar, Gazda, and Smekal suggested an ensemble-based feature selection model and proved that employing an optimal set of indicators enhanced the accuracy of prediction of the Istanbul Stock Exchange when compared to the utilization of all available features. Pehlivanli also compared ten feature selection methods, metrics of their stability, and similarity, and found that univariate methods are better than multivariate methods on high dimensional data. A four stage feature selection process was also suggested by the same researcher to find out compact and high impact feature subsets.

Rakkeitwinai, Lursinsap, Aporntewan and Mutirangura used Principal Component Analysis (PCA) to improve the performance of prediction with the help of the removal of irrelevant features. Liu and Yu developed feature selection principles in different data mining scenarios. Solomon, Sharif, Alazab, and Al-Nemrat came up with a variance-based attribute ranking technique, which was found to solve class imbalance problems, whereas Somol and Novovicova compared the accuracy performance of Information Gain and Pearson Correlation Coefficient on attribute ranking [6]. Fangyi et al. subsequently suggested a fuzzy interpolative reasoning based ranking system to enhance the discovery of pertinent features [7]. More recently, in the biomedical activity, Huda, Yearwood, Jelinek, Hassan, Fortino and Buckland have shown that ensemble classification coupled with feature selection yielded the greatest diagnostic speed and accuracy in the case of oligodendroglial brain tumors [8]. Another strategy proposed by Hamid, Baboli and Hamid to use is Classifier-Based Clustering Strategy (CBCS), using decision trees as a base classifier and ensemble partitioning using k-means [9].

During the experimental analysis of this research, technical indicators are the input attributes of the proposed History Bits Algorithm. An amount of 75 indicators was taken into consideration when engaging in technical analysis. As some indicators are more significant in trading decisions, the algorithm is oriented to the clustering and prioritization of the most significant indicators to enhance the accuracy of the decision [10]. The weights of higher-ranked indicators have a direct influence on the end result of the trading. These key indicators were not only identified but also using six popular attribute ranking strategies availability- Pearson Correlation Coefficient, Information Gain, Gain Ratio, OneR, Relief Evaluator and Symmetrical Uncertainty Evaluator [11,12]. WEKA tool was used to rank and evaluate and attribute ranking mechanism based on ensembling the results of all the single strategies was developed so that there would be a balanced and unbiased ranking of the technical indicators.. History Bits algorithm is then evaluated on the basis of stock market dataset prediction accuracy using different ranking strategies. The reminder of the paper is organized as follows:

Section 2 describes the research data used for experimentation. Section 3 gives the overview of the History Bits algorithm and different ranking strategies. Section 4 gives the ranker wise brief evaluation the History Bits algorithm based on different prediction accuracy parameters. Section 5 concludes the experimental work performed in this paper.

Research data

The experimentation employs a comprehensive real-world dataset derived from the NIFTY 50 index of the Indian stock market. This dataset encompasses 4,977 instances spanning a 20-year period (approximately 2003–2023) and features 75 technical indicators as predictive attributes. Sourced from the official NSE India website (www.nseindia.com), the raw data is presented in the conventional OHLCV (Open, High, Low, Close, Volume) format, which forms the foundation for deriving the technical indicators through standard computational methods.

These indicators are subsequently transformed into binary trading signals—indicating buy or sell recommendations—using a trend deterministic framework that analyzes momentum and directional trends in price movements. To generate final prediction labels, a Wisdom of the Crowd mechanism is applied, leveraging majority voting across multiple indicator ensembles to classify outcomes into five discrete categories: strong buy, buy, hold, sell, and strong sell. The resultant dataset, comprising the original 4,977 instances augmented with these 75 engineered features, facilitates robust machine learning-based forecasting of market directions while mitigating overfitting through ensemble aggregation.

Table 1. Technical Indicators

Group A	Group B	Group C	Group D
Simple Price Average (SPA)	Stochastic Momentum (%D Line)	Triple Smoothed Exponential Average (TSEA)	Dual Exponential Smoothing Average (DESA)
Median Price Average (MPA)	Stochastic Relative Strength Index (Stoch RSI)	Adaptive Volume-Linked Average (AVLA)	Volume-Weighted MACD (VW-MACD)
Smoothed Moving Mean (SMM)	Williams Percent Range (%R)	Bandwidth of Bollinger Bands (BBW)	Market Momentum Oscillator (MMO)
Exponential Price Average (EPA)	Ultimate Momentum Index (UMI)	Standard Pivot Calculations (SPC)	Price Rate Variation (PRV)
Weighted Price Average (WPA)	Awesome Momentum Oscillator (AMO)	Fibonacci Pivot Reference (FPR)	Relative Strength Measure (RSM)
Hull Smoothing Average (HSA)	Mass Volatility Index (MVI)	Volume Force Index (VolFI)	Inverse Fisher Mapping (IFM)
Triangular Price Average (TPA)	Vortex Strength Index (VSI)	Directional Movement Strength (DMS)	True Volatility Range (TVR)
Triple Smoothed Oscillator (TSO)	Sure Thing Signal (STS)	Volume Weighted Price Average (VWPA)	Average True Volatility (ATV)
Volume Adjusted Price Mean (VAPM)	True Strength Oscillator (TSO)	Smoothened Price Mean (SPM)	Parabolic Stop-Reversal Signal (PSR)
Kaufman Efficiency Measure (KEM)	Typical Market Price (TMP)	MACD Trend Divergence (MTD)	Money Flow Momentum (MFM)
Adaptive Kaufman Average (AKA)	Accumulation–Distribution Index (ADI)	Percentage Price Oscillator (PPO)	On-Balance Volume Indicator (OBVI)
Zero-Lag Smoothing Average (ZLSA)	Chaikin Volume Oscillator (CVO)	Weighted Volume Average (WVA)	Volume-Weighted OBV (VWOBV)
Wave Trend Strength Oscillator (WTSO)	Volume Price Correlation (VPC)	Bull–Bear Strength Index (BBSI)	Buy–Sell Pressure Ratio (BSPR)
Price Standard Deviation (PSD)	Volume Zone Strength Oscillator (VZSO)	Ease of Market Movement (EMM)	Normalized Buy–Sell Pressure (NBSP)
Ichimoku Cloud Framework (ICF)	Price Zone Strength Oscillator (PZSO)	Commodity Channel Measure (CCM)	Chande Momentum Strength (CMS)
Vector Magnitude Indicator (VMI)	Elder Force Strength Index (EFSI)	Coppock Trend Curve (CTC)	Chandelier Exit Level (CEL)
Squeeze Market Momentum (SMMI)	Cumulative Force Oscillator (CFO)	Twiggs Money Flow (TMF)	Q-Value Stick Indicator (QSI)
Directional Motion Index (DMI)	Stochastic %K Signal (S%K)	Adaptive Price Zone Index (APZI)	Finite Volume Element (FVE)
Fisher Transform Indicator (FTI)	Bollinger Band Range (BBR)	—	—

Method, Experiments and Results

This paper presents a new algorithm known as the History Bits Algorithm which aims at improving the accuracy of the decision prediction of the stock market. The basic principle is prioritization and classification of a complete set of 75 technical indicators using a systematic process of prioritization based on attributes.

Most traditional ranking methods, like Information Gain, Pearson Correlation Coefficient, Gain Ratio, OneR, Relief Evaluator, and Symmetrical Uncertainty Evaluator are usually prone to bias when used in isolation. In order to overcome this weakness, an ensemble based ranking framework is suggested. This process integrates the results of the various assessors in that, the results are combined and a more balanced and credible result is achieved. The ensemble method has the advantage of effectively reducing the impact of individual evaluator bias and decreasing the overall ranking error through the integration of multiple views which results in a more objective prioritization of indicators [1315].

Upon acquiring the ranks obtained by unification, the History Bits Algorithm, initially thought of by Sakhare and Joshi, subdivides the indicators into five distinct groups, which each has 15 indicators in descending order of significance. All groups are then given a weight factor in relation to the level of priority so that the indicators with greater levels of priority have more weight in the final decision that the model will trade.

A. History Bits Algorithm

The proposed History Bits algorithm works as follows

Step 1- More weight is assigned to the technical indicators with higher ranks.

Weights are initialized to the indicators using formula given in equation (1)

$$y = w * x + b \tag{1}$$

$$y = \varphi(\sum_{i=1}^{n} W_i x_i + b) \tag{2}$$

$$y = \varphi(w^T + b) \tag{3}$$

Step 2- if number of ‘sell’ calls are more than ‘buy’ calls then the bit 0 is assigned to the corresponding group, else it is assigned as 1.

Introduction of ambiguity in the second step of the proposed approach may sometimes cause slight variation in the results of trading decisions. An example is where the Relief Evaluator creates rankings of indicators which are then partitioned into five groups, each of which is one bit of a 5-bit pattern. Those indicators that belong to the first group have higher rankings than those indicators that belong to the second ones and so forth. This hierarchical group, however, can also introduce boundary ambiguities when there are indicators that are close to the boundary between two groups and have differing priorities.

The worst-case scenario is more of an exception than a rule, and happens when the number of indicators that give a buy signal slightly outnumber the number that give a sell signal. In this case the decision outcome of the resulting group will prevail and hence affect the ultimate prediction on the trading. Though these boundary conditions are not common, the possible effect of these is alleviated through the implementation of an average-based aggregation procedure when analyzing the accuracy of the algorithm.

Even in the worst case, as can be empirically observed of the resulting bit patterns, less than one out of five indicators in each of the first three groups yields a signal counter to the majority trend. This proves that the algorithm is resistant to such edge cases so as to maintain consistent performance even when indicators behave slightly differently.

Step 3 – Bit Pattern Aggregation

The individual binary outputs obtained from each indicator group are combined to form a unified bit sequence. This consolidated representation reflects the cumulative trading signals derived from all groups, where each bit corresponds to one indicator cluster's decision outcome.

Step 4 – Five-Bit Representation Formation

Since the indicators are divided into five prioritized groups, the resulting sequence naturally constitutes a five-bit pattern. Each bit encapsulates the collective behavior of one group—yielding a compact yet information-rich encoding of the entire indicator set.

Step 5 – Label Generation for Predictive Modeling

The constructed five-bit pattern is then utilized to generate prediction labels for subsequent machine-learning models. These labels act as the target variables for supervised learning algorithms, enabling the History Bits Algorithm to integrate seamlessly with predictive frameworks for trading-decision forecasting.

Table 2.Interpretation of Bit Patterns

Bit pattern	Calls
[11111]	Strong_buy
[11110]	Strong_buy
[11101]	Strong_buy
['11011]	Strong_buy
[10111]	Strong_buy
[01111]	Buy
[11100]	Strong_buy
[11010]	Buy
[10110]	Buy
[01110]	Hold
[11001]	Buy
[10101]	Hold
[01101]	Hold
[10011]	Hold
[01011]	Hold
[00111]	Hold
[11000]	Hold
[10100]	Hold
[10010]	Hold
[10001]	Hold
[01100]	Hold
[01010]	Hold
[01001]	Sell
[00110]	Sell

[00101]	Sell
[00011]	Strong_sell
[10000]	Sell
[01000]	Strong_sell
[00100]	Strong_sell
[00010]	Strong_sell
[00001]	Strong_sell
[00000']	Strong_sell

In this paper different attribute ranking strategies are used to evaluate and analyze the prediction performance of the proposed History Bits algorithm.

B. Attribute Ranking Strategies

The attribute ranking stage is important in the process of determining the most powerful technical indicators of the History Bits Algorithm. Six ranking methods, which include Pearson Correction Coefficient, Information Gain, Gain Ratio, OneR, Relief Evaluator, and Symmetrical Uncertainty Evaluator, are used in this paper to estimate the relative importance of attributes using different evaluation criteria. Both approaches prioritize the indicators as per its own statistical or information-theoretic viewpoint giving rise to a list of priorities of the most to the least important indicators.

The attribute priority of the History Bits Algorithm is a very important element because the relative priority of the attributes directly influences how the group forms and the performance of the prediction. Nevertheless, when using individual ranking strategies different ranking orders are obtained resulting in differences in attributes-priority and possibly inconsistencies between model results [16]. In order to curb these prejudices, an ensemble-based ranking strategy is suggested. It involves the combination of the results of several evaluators into a compound ranking, which guarantees a more balanced, stable, and objective ranking of indicators. The analysis of prediction capability of the History Bits Algorithm during the individual and ensemble-based ranking conditions could give more information about the trustworthiness and the resilience of the offered approach.

C. Pearson Correlation Coefficient

Correlation Coefficient of Pearson is a statistical statistic which is used to measure the level of linear association between two variables using their covariance (Hall, 1999). The coefficient value is between +1 and -1 in which the value magnitude and sign show, respectively, the strength and direction of correlation.

PCC values interpretation is as follows:

- Perfect Correlation: The coefficient value of + 1 or -1 indicates a positive or negative linear relationship between the two variables.
- High Correlation: The coefficient between +-0.5 and +-1.0 is a strong relationship.
- Moderate Correlation: The quality of correlation lies between +-0.30 and +-0.49 which depicts that the correlation is moderate.
- Low Correlation: Coefficient values of less than +-0.29 imply that there is a weak linear relationship.

- No Correlation: The value of 0 shows that there is no linear relationship between the two variables. The Pearson Correlation Coefficient of variables, X and Y is mathematically calculated as:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sqrt{\Sigma_{i=1}^{n}(y_i-\bar{y})^2}}} \quad (4)$$

D. Information Gain

Information Gain (IG) is an entropy measurement of evaluation that is used to determine the contribution of a specific attribute to the prediction of the target variable. It is a response of the decrease in uncertainty on the output with the known value of a given feature. The value of information gain of 0 means that the attribute does not give predictive information, and the value of information gain of 1 gives the maximum predictive information. Attributes that succeed in splitting the dataset into homogeneous subsets provide better information gain and features that have no relations to the target class provide information gain that is either very small or zero. Basically, entropy is a measure of impurity or randomness of the dataset, and information gain is a measure of how much impurity a selection based on a specific attribute decreased [17].

Let X be the set of training instances such that $x = \{x_1 x_1, x_2 x_2, x_3 x_3, ..x_n x_n\}$ and Y is the prediction label. Each $x_a \in val(a) x_a \in val(a)$ is the value of the attribute in X.

For a value 'a' taken by attribute xa, the information gain of an attribute xa is defined using Shannon entropy.

Consider:

$$s_a(v)s_a(v) = \{x \in Xx \in x |\, x_a \in val(a) x_a \in val(a)\} \quad (5)$$

as a set of training instances of X such that value of attribute 'x' is 'v'. Information gain of X for attribute 'a' is the difference between H(X) and H(X|a), where H(X) is known as priori Shannon entropy and H(X|a) is known as conditional entropy.

$$H(x|a) = \sum_{v \in va^1(a)} \frac{|S_a(v)|}{|X|} H\big(S_a(v)\big) \quad (6)$$

$$IG(X,a) = H(X) - H(X|a) \quad (7)$$

Information Gain proved to be a good strategy to identify the relevance of the attribute. However it doesn't work well when attributes hold to many distinct values [18]. For example, the NIFTY 50 dataset considered for experimentation in this paper has 75 attributes each having distinct values for 4977 instances causing the problem of overfitting. To overcome this issue, Ross Quinlan proposed another ranking strategy based on gain ratio for considering the attributes with lower information gain value.

E. Gain Ratio

In order to minimize the bias for multivalued attributes Ross Quinlan proposed another feature selection technique known as gain ratio. Gain ratio is defined as the ratio of the information gain to the intrinsic information.

Let X be the set of all features and S be the set of all training instances. Consider a function value (x, a) such that

$$x \in Sx \ \ and \ a \in xa \in x$$

The function value (x, a) provides the set of all possible values of a where $a \in x a \in x$ and information gain for a is defined as:

$$IG(s,a) = H(S) - \frac{|\{x \in s | value(x,a) = v\}|}{|s|} . H(|\{x \in s | V(x,a) = v\}|) \tag{8}$$

F. OneR

OneR calculates the ranking of the features using the rule based learning algorithm. This algorithm constructs simple yet accurate rules by associating the value of the feature with the majority class. Features of the dataset are sorted as per the quality of the rules. Quality of the rules is determined by accuracy of the rule. For all the rules constructed total error is calculated by evaluating it against the entire dataset and the rule with them smallest error is selected. As a ranker, OneR will sort the attributes based on the error rates.

Pseudocode:

```
    for each feature in the dataset
     {          for each value of the feature


    {compute the values of prediction   classes
     prediction label = most frequent prediction label
    construct a rule such that feature=value=prediction label
                }
                computer the error rate for the rule
     }
  select the rule having lowest error rate
```

G. Relief Evaluator

A filter based attribute ranking method is called the Relief Evaluator which measures the relevance of features in terms of their capability to separate neighboring examples. It was originally intended to be used in binary problems of classification with numerical features, and calculates a feature relevance score that may be used to rank attributes with respect to their predictive relevance [19].

The algorithm works by finding nearest neighbor instances to each point of data. It then compares the variation of feature values between the instances that are of the same category (nearest hit), and other instances which belong to different categories (nearest miss). When a feature depicts similar values representing the instances of the same class, its score goes down, which presupposes a low discriminative power. On the other hand, when a feature changes more significantly among the instances of different classes, the score is higher, which indicates the higher predictive relevance. By this means, the Relief Evaluator is an effective way to estimate the contribution of each feature to the separation of classes, and as such, it is an effective and computationally simple way to rank features in a classification task..

Algorithm:

Input: - ‘n’ training instances

‘f’ feature count

‘t’ training instances to update W such that t Ɛ n.

Steps:

$initially\ w_i = 0$

for i=1 to t

select the instance arbitrarily

find near hit (the nearest instance of the same class) and near miss (the nearest instance of the different class)

update W such that

for i=1 to f

$w_i = W_i - (x_i - near\_hit)^2 + (x_i - near\_miss)^2$

end for

end for

$Return\ w_i$

H. Symmetrical Uncertainty Evaluator

Symmetrical uncertainty is another useful attribute ranking strategy with filter based feature selection mechanism. It uses information gain based theoretic measure to evaluate the goodness of the attributes. This evaluator has symmetric information gain property which states that aggregate of the information given by one feature about another feature is effectively same as that of information given by later feature to the former feature. Hence SU (x,y) is same as that of SU (y,x) where x and y are independent of each other. Unlike information gain, symmetric uncertainty is not influenced by multivalued attributes and values are normalized [20].

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \tag{9}$$

IG(X|Y) indicates the information gain of feature X for class Y and H(X) and H(Y) indicates the entropy of X and Y respectively [21].

I. Ensemble Based Ranking Strategy

You cannot know which feature selection technique is best suitable for dataset under consideration. Subsequently, it is a smart thought to attempt various diverse attribute determination procedures on your dataset and thus make a wide range of perspectives on your dataset [22]. Analyzing the results will give an idea of which view could result in best performance. Compare the results to get an idea of which view of specific attributes of the dataset results in the best performance [23,24]. In our research work considering the numerical nature of our dataset we have also considered correlation based feature selection which best suites. Learner based feature selection technique is well suited for nominal and categorical data. Information gain based feature selection does not work well in case attributes have diversified values. This is may result in overfitting. To avoid these issues we have used correlation based feature selection method for indicator selection. However to avoid the biasing effect of any particular ranking strategy and other disadvantages a novel ensemble based ranking strategy is proposed which considers the aggregation of ranking results of different attribute ranking methods [25,26,27].

Results

Technical indicators are the features calculated by examining historical data of a stock index based on the volume, price, open and close of a contract. We used these technical indicators as input attributes for our algorithm. Different indicators have different impact on the trading decision. Therefore we analyze the importance of these indicators with the prediction class using different ranking strategies [28]. In the second phase of the experimentation prediction performance of a novel History Bits algorithm is evaluated against different ranking strategies. For the experimental work we have used dataset of NIFTY 50 index of Indian Stock market having 4977 instances with 75 indicators

Table 3. Performance evaluation of History Bits algorithm based on Pearson Correlation Coefficient

Performance Metric	**Recorded Value**
Total Sample Count	4,977
Accurate Predictions	4,213
Misclassified Records	764
Overall Classification Accuracy	**84.64 %**
Cohen's Kappa Coefficient	0.7399
Mean Absolute Deviation (MAD)	0.3453
Root Mean Squared Deviation (RMSD)	0.9334
Macro-Averaged Precision	0.83
Macro-Averaged Recall	0.85
Macro-Averaged F1 Score	0.83
Weighted Matthews Correlation Index (WMCI)	0.74

Table 4. Performance evaluation of History Bits algorithm based on Information Gain

Evaluation Criterion	**Measured Output**
Total Sample Size	4,977
Correct Predictions	4,367
Misclassified Samples	610
Overall Predictive Accuracy	**87.74 %**
Cohen's Kappa Index	0.7929
Mean Absolute Deviation (MAD)	0.2656
Root Mean Square Deviation (RMSD)	0.8083
Average Precision (Weighted)	0.86
Average Recall (Weighted)	0.88
Average F-Score (Weighted)	0.87
Weighted Matthews Correlation Index (WMCI)	0.79

Table 5. Performance evaluation of History Bits algorithm based on Gain Ratio

Assessment Metric	**Recorded Outcome**

Total Data Samples	4,977
Correctly Predicted Records	4,358
Incorrectly Predicted Records	619
Overall Model Accuracy	**87.56 %**
Cohen’s Kappa Coefficient	0.7889
Mean Absolute Deviation (MAD)	0.2660
Root Mean Square Deviation (RMSD)	0.8000
Weighted Average Precision	0.85
Weighted Average Recall	0.88
Weighted Average F1 Score	0.86
Weighted Matthews Correlation Index (WMCI)	0.79

Table 6. Performance evaluation of History Bits algorithm based on OneR

Evaluation Indicator	Observed Result
Total Data Instances	4,977
Accurate Classifications	4,312
Misclassified Entries	665
Overall Model Accuracy	**86.63 %**
Cohen’s Kappa Value	0.7751
Mean Absolute Deviation (MAD)	0.3066
Root Mean Square Deviation (RMSD)	0.8874
Weighted Precision Score	0.85
Weighted Recall Score	0.87
Weighted F1-Measure	0.86
Weighted Matthews Correlation Index (WMCI)	0.77

Table 7. Performance evaluation of History Bits algorithm based on Relief Evaluator

Performance Indicator	Measured Outcome
Total Number of Samples	4,977
Correct Predictions	4,312
Incorrect Predictions	665
Overall Classification Accuracy	**86.63 %**
Cohen’s Kappa Coefficient	0.7751
Mean Absolute Deviation (MAD)	0.3066
Root Mean Square Deviation (RMSD)	0.8874
Weighted Precision Score	0.85
Weighted Recall Score	0.87
Weighted F1 Score	0.86
Weighted Matthews Correlation Index (WMCI)	0.77

Table 8. Performance evaluation of History Bits algorithm based on Symmetrical Uncertainty Evaluator

Evaluation Aspect	Observed Metric
Total Data Samples	4,977
Correct Predictions	4,367
Incorrect Predictions	610
Overall Accuracy Rate	**87.74 %**
Cohen's Kappa Value	0.7929
Mean Absolute Deviation (MAD)	0.2656
Root Mean Square Deviation (RMSD)	0.8083
Weighted Precision Score	0.86
Weighted Recall Score	0.88
Weighted F1 Score	0.87
Weighted Matthews Correlation Index (WMCI)	0.79

Table 9. Performance evaluation of History Bits algorithm based on Ensemble Based Ranking Strategy

Evaluation Metric	Recorded Outcome
Total Data Samples	4,977
Correct Predictions	4,348
Misclassified Samples	629
Overall Model Accuracy	**87.36 %**
Cohen's Kappa Coefficient	0.7889
Mean Absolute Deviation (MAD)	0.2835
Root Mean Square Deviation (RMSD)	0.8480
Weighted Precision Score	0.85
Weighted Recall Score	0.88
Weighted F1-Measure	0.86
Weighted Matthews Correlation Index (WMCI)	0.79

From the performance evaluation of History Bits algorithm based on different raking strategy it is clear that ensemble based ranking strategy not only overcomes the biasing effect of any particular ranking strategy but also shows the prediction accuracy as high as other ranking strategies.

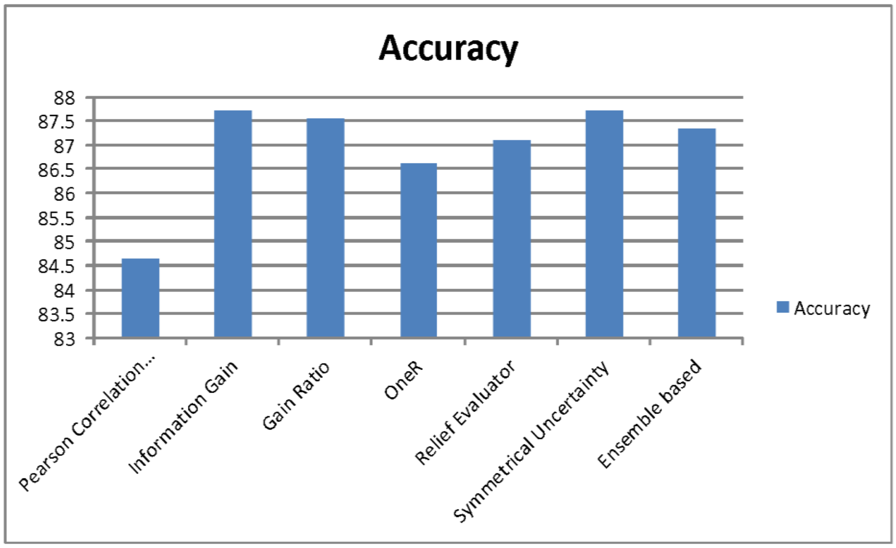

Figure 1. Analysis of History Bits Algorithm Using Different Attribute Ranking Strategies

Conclusions

Technical analysis is a very important part in the stock market prediction and there is a great variety of technical indicators that can assist in the given process. Such indicators serve as input variables in stock market datasets, which normally contain the parameters of open, high, low, and close prices. Based on these parameters, diverse derived indicators may be calculated to reflect bury or sell signals to guide intelligent trading choices by investors and algorithms. The indicators have different degrees of importance as each one has a contribution in the prediction of the class labels. Thus, prioritization and sorting of indicators according to their relative importance can be of significant help in raising the level of prediction. With various attribute ranking strategies available, it will be necessary to evaluate the efficiency of the suggested History Bits Algorithm with reference to various ranking strategies. Pearson Correlation Coefficient, Information Gain, Gain Ratio, OneR, Relief Evaluator, Symmetrical Uncertainty Evaluator, and proposed Ensemble-Based Ranker were used in this study in order to identify the significance of each indicator. Each method of ranking was tested regarding the accuracy of its predictions in the model. These findings indicate that the Ensemble-Based Ranker had better predictive accuracy than single strategies. Also, the History Bits Algorithm had a consistency in its results of predictions under all the ranking methods, which indicates its high strength and adaptability.

The future research can include to continue with this analysis through incorporation of dimensionality reduction algorithms like Principal Component Analysis (PCA), Backward Feature Elimination, or Forward Feature Construction. A combination of these approaches with the History Bits Algorithm can also help to improve computational efficiency and prediction reliability.

References

1. Pehlivanli A.C., Baris A., & Gulay G. (2016) Indicator Selection with Committee Decision of Filter Methods for Stock Market Price Trend in ISE, Elsevier, Applied Soft Computing, Volume 49, Pages 792-800. https://doi.org/10.1016/j.asoc.2016.09.004J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

2. Drotar, P., Gazda, J., & Smekal, Z. (2015) An experimental comparison of feature selection methods on twoclass biomedical datasets, Computers in Biology and Medicine, Volume 66, Pages 1-10, https://doi.org/10.1016/j.compbiomed.2015.08.010
3. Rakkeitwinai, S., Lursinsap, C., Aporntewan, C., & Mutirangura, A. (2015) New feature selection for gene expression classification based on degree of class overlap in principal dimensions, Computers in Biology and Medicine Volume 64, Pages 292-298 https://doi.org/10.1016/j.compbiomed.2015.01.022
4. Liu, H., & Yu, L. (2005) Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering 17 (4) Pages 491-502 10.1109/TKDE.2005.66
5. Ashraf, M., Chetty, G., & Tran, D. (2013) Feature selection techniques on thyroid, hepatitis, and breast cancer datasets. International Journal on Data Mining and Intelligent Information Technology Applications(IJMIA),vol. 3, no. 1, pp. 1-8, https://doi.org/10.1007/978-3-642-34478-7_34
6. Somol, P., & Novovicova, J. (2010) Evaluating Stability and Comparing the output of feature selectors that optimize feature subset cardinality. IEEE Transaction on Pattern Analysis and Machine Intelligence, pp.1921- 1939, https://doi.org/10.1109/TPAMI.2010.34.
7. Fangyi, L., Changjing, S., Ying, L., Jing, Y., & Qiang, S., (2018) Fuzzy Rule Based Interpolative Reasoning Supported by Attribute Ranking. IEEE transaction on Fuzzy Systems, Vol. 26, Issue 5, pp. 2758-2773 10.1109/TFUZZ.2018.2812182
8. Huda, S., Yearwood, J., Jelinek, H., Hassan, M., M., Fortino, G., & Buckland, M. (2016) A Hybrid Feature Selection with Ensemble Classification for Imbalance Healthcare Data: A case study for Brain Tumor Diagnosis. IEEE ACCESS 2016, vol. 4, pp. 9145-9154, 10.1109/ACCESS.2016.2647238
9. Hamid, P., Baboli, M., M., & Hamid, A., R.(2015) Proposing a classifier ensemble framework based on classifer selection and decision tree. Elsevier, Engineering Applications of Artificial Intelligence, Vol. 37. , pp.34-42 10.1016/j.engappai.2014.08.005
10. Joshi, S., & Sakhare, N.(2015) History Bits based novel algorithm for classification of structured data, 2015 IEEE International Advance Computing Conference (IACC), Banglore, pp. 609-612. 10.1109/IADCC.2015.7154779
11. Syed, I., & Waseem, S. (2012) A Feature Subset Selection Method Based on Symmtric Uncertainty and Ant Colony Optimization. 2012 International Conference on Emerging Technologies 10.1109/ICET.2012.6375420
12. Gustavo, S., C., Miguel, G.,T., Santiago, G.,G., Schaerer, C., E., & Federico, D. (2019) A Multivariate Approach to the Symmetrical Uncertianty Measure: Application to Feature Selection Problem. Elsevier, Information Sciences, vol. 494, https://doi.org/10.1016/j.ins.2019.04.046
13. Altidor, W., Khoshgoftaar, T., M., Hulse, V., & Napolitano, A. (2011). Ensemble Feature Ranking Methods for Data Intensive Computing Applications. Handbook of Data Intensive Computing, Springer, Newyork, pp. 349-376, https://doi.org/10.1007/978-1-4614-1415-5_13
14. Pes, B. (2020) Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Computing & Applications,Volume 32, Pages 5951–5973. https://doi.org/10.1007/s00521-019-04082-3
15. Pinar, Y., (2015) Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. International Journal of Machine Learning and Computing, Vol. 5(4): Pages 258-263 DOI: 10.7763/IJMLC.2015.V5.517.

16. Liu, H., & Setiono, R. (1995) CHI2: Feature selection and discretization of numeric attributes. 7th IEEE International Conference on Tools with Artificial Intelligence, 10.1109/TAI.1995.479783

17. Sakhare, N., & Joshi, S. (2014) Classification of Criminal Data Using J48 Algorithm", International Journal of Data warehousing and Mining. Vol. 4., pp. 167-171.

18. Sakhare, N., Shaik, I., Kagad, S., Kapadwanjwala, T., Malekar, H., Dalal, M. (2020) Stock Market Prediction Using Sentiment Analysis. International Journal of Advanced Science and Technology. Volume 29, Issue 4s pp. 1126 - 1133. http://sersc.org/journals/index.php/IJAST/article/view/6664.

19. Khalid, S., Khalil, T., & Nasreen, S. (2014) A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. Science and Information Conference (SAI) London, UK, pp. 372-378. 10.1109/SAI.2014.6918213

20. Lee, I., Lushington, G., & Visvanathan, M. (2011) A filter-based feature selection approach for identifying potential biomarkers for lung cancer. Journal of clinical Bioinformatics, vol. 1, no. 11, pp. 1-8, https://doi.org/10.1186/2043-9113-1-11

21. Hall, M., A., & Smith, L., A. (1999) Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, pp. 235-239, 1999.

22. Pes, B. (2017) Feature Selection For High-Dimensional Data: The Issue of Stability. 26th International Conference on Enabling Technologies: Infrastruture for Collaborative Enterprizes, WETICE, Poznana, Poland, pp. 170-175 10.1109/WETICE.2017.28

23. Novakovic, J., Strbac, P., & Bulatovic, D. (2011) Toward optimal feature selection using ranking methods and classification algorithms Yugoslav Journal of Operations Research, vol. 21, no. 1, pp. 119-135, http://yujor.fon.bg.ac.rs/index.php/yujor/article/view/364

24. Tsymbal A., & Puuronen S. (2000) Local Feature Selection with Dynamic Integration of Classifiers. In: Raś Z.W., Ohsuga S. (eds) Foundations of Intelligent Systems. ISMIS 2000. Lecture Notes in Computer Science, vol 1932. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-39963-1_4

25. Saeys, Y., Abeel T., & Van de Peer Y. (2008) Robust Feature Selection Using Ensemble Feature Selection Techniques. Machine Learning and Knowledge Discovery in databases, Lecture notes in Computer Science, Vol. 5212, Springer, Berlin, pp.313-325 https://doi.org/10.1007/978-3-540-87481-2_21

26. Kuncheva, L., I., Smith, C.,J., Syed, Y., Phillips, C.,O., & Lewis, K.,E. (2012) Evaluation of Feature Ranking Ensmbles For High Dimensional Biomedical Data. A case study"; IEEE 12th International Conference on Data Mining Workshops,pp.49-56, 10.1109/ICDMW.2012.12

27. Kohavi, R., & John, G. (1997) Wrappers for feature subset selection. Artificial Intelligence Volume 97, Issues 1–2, Pages 273-324 https://doi.org/10.1016/S0004-3702(97)00043-X

28. Sakhare, N., & Shaik, I. (2019) Performance Analysis of Regression Based Machine Learning Techniques for Prediction of Stock Market Movement. International Journal of Recent Technology and Engineering, Vol.7, issue 6S, pp. 655-67