Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 7763 – 77474

Research Article

Predicting Stock Market Dividends Using Ranking and Data Mining Techniques

Mahshad Najafi[a*], Hamid Paygozar [b]

[a] Master of Information Technology Engineering E-Commerce Orientation, Khomein University, Arak, Iran. Email: mahshad_najafi@hotmail.com, ORCID: 0000-0002-2687-5345, Corresponding author
[b] PhD in Computer Engineering - Artificial Intelligence,Amirkabir University of Technology,Tehran,Iran, Email: paygozar.hp@gmail.com

Article History: Received: 14 July 2020; Accepted: 2 January 2021; Published online: 5 February 2021

Abstract: Data mining is one of the growing sciences and is very suitable for the analysis of databases. Data mining is used in many sciences, such as business intelligence, shopping cart analysis, and medicine. The main data mining algorithms are 4 categories, that 2 of their main categories are attribute ranking and classification algorithms. So far, no research has been presented regarding the use of ranking in classification algorithms. In this research, we present a method for predicting dividend market price using ranking and data mining techniques. First, a new method for classifying data has been presented, then ranking algorithms has been used as the input of this method. Afterwards, by implementing the above approach on 10 databases and obtaining the accuracy of each model according to the model inputs (attribute ranking algorithm) and placing the accuracy obtained in each stage in the data envelopment analysis model, we have evaluated and ranked the attributes. We have used stock market data to make this research applicable and based on the approach, we have predicted the ratio of change in companies' dividends in 2015 according to the companies' data. The results indicated high accuracy of the proposed approach and its high speed.

Keywords: dividend change ratio; market price; ranking; data mining

1. Introduction

Discussions related to accounting information introduce earnings as a sign of an institution's ability for cash dividends, because investors believe that earnings figure represents the foundation of economic events that in turn reflect the ability to pay and distribute earnings. The question that arises here is whether, apart from the earnings figure, other information and figures of the financial statements that accrual accounting presents them are related to the nature of the cash dividend or not? Theoretically, assuming a complete capital market, as long as there are investment plans with a return of more than their cost required for the company, the company will use the retained earnings to finance them, and after financing all the investment opportunities of the company, the earnings, if remained, will be distributed in cash among the shareholders. Obviously, if the investment opportunities of the company are more than their earnings, the institution for financing them will use issuing new stocks or borrowing or a combination of both. From this point of view, the amount of cash earnings depends on the number of investment opportunities of the company and the costs required to implement them (Miller, 201

The relative acceptance of the efficient securities markets hypothesis, in scientific circles, changed the deductive and inductive reasoning basis of accounting theory from two aspects and about changing the purpose of preparing accounting information; the first is that an information is required to make economic resources allocated among producers and are optimally distributed. Optimal allocation of economic resources occurs when producers can create the highest gross domestic product with a certain amount of resources. Secondly, the information is required that investors by using it can maintain a selective set of securities with regard to risk-taking at desirable limit of output (in the securities price structure existing in the market) (Saghafi, 2013).

Growth and investment opportunities are one of the factors that investors consider along with other stock evaluation criteria such as higher dividends, free cash flows and stock risk. With the continuous development of the society's economy, there has been a rapid increase in the emergence of capital markets in the country. Today, investment in the stock market composes an important part of society's economy. For this reason, predicting stock price is of particular importance for shareholders to be able to obtain the highest return on their investment (Elson & Delen, 2014). The main purpose of investors in investing in the stocks of companies is to increase wealth, which is realized by obtaining stock returns. Therefore, evaluating the stock returns of various companies is the most important issue that investors face in the capital market.

There is no doubt that investment in the stock market composes an important part of the country's economy, and undoubtedly, the highest ratio of the capital is exchanged through stock markets around the world, and the national economy is strongly affected by stock market performance. In recent years, the world financial markets have always faced significant fluctuations and uncertainties, in a way that the existing uncertainty related to the return on invested assets has made many investors and financial analysts worried (Asgharpour, 2012). As investors state, uncertainty is the most important factor in pricing any financial asset. Many studies have been performed regarding dividends on the stock exchange and predicting these earnings for the coming years is very important, but no comprehensive studies have been done so far about predicting investment earnings in the stock market. Using the company's profitability in previous years is effective in predicting dividends, but it is not that much efficient and by regression methods the amount of increase or decrease in profitability cannot be accurately determined (Bezale, Gholami, Faghih, & Ahmadi, 2011).

The increase in the number of financial crisis cases in public joint stock companies in recent years has attracted the attention of many investors and creditors to predict these cases through the information provided in the financial statements of companies in order to prevent losses due to them. But it is very difficult to predict these financial crises, especially in cases that earnings management is also involved. Many earnings management studies have only identified the factors that can affect earnings management and have examined the dependence of earnings management on these factors and not the use of these factors in predicting earnings management.

The need to understand large complex datasets and complete and rich information in the fields of business, science and engineering is more or less common. In the business world, the data of companies and customers are proposed as a strategic capital. The ability to extract useful knowledge and information existing in this data and the possibility of using this knowledge in today's competitive world is more than ever important. In general, the process of applying computer-based methodology, including new methods for obtaining knowledge and information, is called data mining.

Data mining is a repeatable process in which progress is made by exploration through automated or manual methods. Data mining is the most useful exploratory analytical scenario in which there is no predetermined notion and perception about the "remarkable" result that is obtained. In fact, data mining is the necessary search for finding new, valuable, and unobvious general information from large volumes of data. In other words, data mining is the interaction of cooperation between human being and computer. The best results are obtained by creating balance between the knowledge of experts in describing issues and goals and computer search capabilities (Chou, Lin, Chou, & Haung, 2014).

Data mining refers to the investigation and analysis of large amounts of data in order to discover meaningful patterns and rules. Data mining is mainly related to building models. A model is basically said to an algorithm or a set of rules that relates a set of inputs (usually in the form of contexts in an organization's database) to a specific purpose or destination. Regression,

neural networks, genetic algorithm, decision trees and nearest neighbor, visualization, the rule of collective establishment and discovery, and many other artificial intelligence techniques, statistical analyses, mathematical and optimization methods are techniques for modeling in data mining.

Classification seems to be one of the most common data mining duties. It is one of the duties of mankind; we constantly classify, categorize and grade in order to recognize and establish a relationship with the world. We divide living creatures into humans, animals, plants, and various races. Business issues such as turning analysis, risk management, and case targeting are included in the classification. Classification refers to assigning cases to groups based on a predictable attribute. Each case includes a set of attributes, one of which is called class attributes (predictable attributes). This duty involves finding a model that describes the class attribute based on a function of the input attributes. To teach the ranking model, you need the class value of the input cases in the training dataset.

Artificial Neural Networks (ANN), or more simply neural networks, are new computational systems and methods for machine learning, knowledge display, and finally the application of knowledge gained in order to predict output responses from complex systems. The main idea of such networks is (to some extent) inspired from the way the biological neural system function, to process data and information in order to learn and create knowledge. The key element of this idea is to create new structures for the information processing system. The system is made up of a large number of extremely interconnected processing elements called neurons that work with each other coordinately to solve a problem and transmit information through electromagnetic communications. In these networks, if one cell is damaged, other cells can compensate its absence, and also contribute to its reconstruction. These networks are able to learn. For example, by irritating tactile nerve cells, the cells learn not to go towards the hot object, and with this algorithm, the system learns to correct its error. Learning in these systems is performed as adaptive, that is, using examples, the weight of the electromagnetic communications changes in such a way that the system produces the correct response if new inputs are given (Hwangn & Yoon, 2014).

A neural network consists of a network of simple processing elements (neurons), which can exhibit a determined general complex behavior of the relationship between processing elements and element parameters. The main and inspiring source for this technique comes from the experiment of the central nervous system and neurons (axons, numerous branches of nerve cells, and junctions of two nerves), which forms one of the most notable elements of nervous system information processing. In a neural network model, simple nodes (broadly "neurons"), "PE" ("processing elements: or units") have been connected to each other to form a network of nodes. For this reason, it is referred as the neural networks term, while a neural network should not be self-compatible by itself, practical use of it is possible through algorithms, which have been designed to change the weight of communication in the network (in order to produce the desired signal).

An efficient and specific method to create classifiers or categorizers from data is to generate a decision tree. The decision tree representation uses the logical method extensively. There are many inductive algorithms of decision tree that are mainly described in machine learning and applied statistical literature. They are supervisory learning methods that create decision trees from a set of input-output samples. An educational sample system uses a decision tree of top-down strategy that creates a solution in one part of the search space. This method ensures that a simple tree, but not necessarily the simplest tree, will be found. A decision tree consists of nodes that have tested specificities. The outer branches of a node correspond to all possible test outputs

in the node. A simple decision tree for classifying samples with two input characteristics of X and Y has been presented in Figure 2-7. All samples with specificity values of X <1 and Y = B belong to the class or classification 2, while samples with value of X <1 belong to class 1 (with both values for characteristic Y). The samples are segmented in a non-leaf node in a tree structure along the branch, and each node obtains the child of a proportional subset of the samples. Decision trees that use two single-deviation components or sections have a simple layout that make it is simple relatively to use and in order to understand the resulting model. At the same time, they show a limitation on the significance of the model. In general, any limitation on displaying the tree can generally limit the form of performance and therefore the approximate power of the model. A well-known growing tree algorithm for generating decision trees is based on two deviant parts of Quinlans ID3, which is provided with an extended version called C4.5. Greedy search methods that include the growth and greed of decision tree structures, particularly in this algorithm may be used to search for exponential space of these models.

The ID3 algorithm starts with all the educational samples in the main node of the tree. An attribute or specificity is selected to divide these samples. For each value of the attribute, a branch is created and the corresponding subset of the samples that has the specificity value is determined by the branch and moves towards the newly created child node. The above algorithm is used recursively for each child node so that all samples are in one node of the class. Each path towards the leaf in the decision tree represents a classification rule. Note that critical decisions in such an algorithm of top-down decision generation is to select specificity and attribute in one node. Specificity attribute selection in ID3 and C4.5 algorithms is based on the minimization of an entropy evaluation of the information used in the examples in a node.

In this research, stock market data are classified using data mining algorithms and then the obtained results are evaluated using ranking algorithms to obtain the stocks with the highest earnings.

2. Suggested Method

The diagnostic system proposed in this research performs the diagnostic operation through three main steps:

1. Preparing data for processing through data normalization: In this step, the initial information is preprocessed and the samples are transformed into a structure that will be used in the next steps of the proposed algorithm.
2. Classifying data related to processed specificities using:

A) Decision Tree

B) K Nearest Neighbor (KNN) Algorithm

C) Multilayer Perceptron Neural Network

3. Voting for the classification results of each of the three algorithms: In this step, the results obtained from the three algorithms used in the previous step are combined to determine a final result regarding the sample being profitable or harmful. The purpose of applying the voting technique in the proposed method is to reduce the error of each one of the classification algorithms used.

The structure of the proposed method and the relationship between the components and steps of the model presented in this research have been displayed in Figure 1.

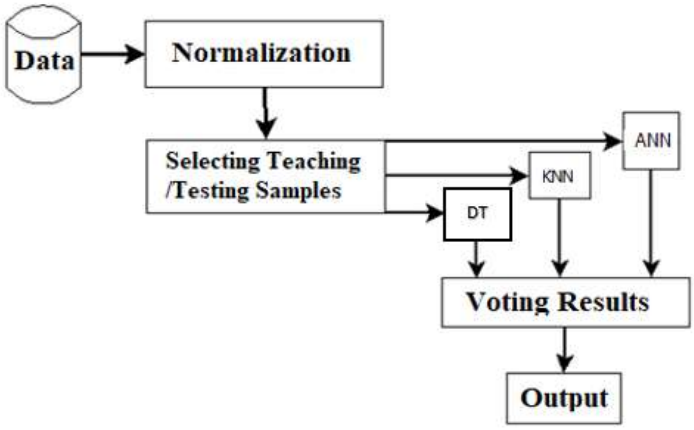

Fig. 1 structure of the proposed method and the relationship between components and steps.

3. Data Preprocessing

The data preprocessing step is the first step in the proposed method and is used to prepare the database for processing in the next steps. This step involves a set of operations to normalize the specificities. In order to normalize the data, we perform the following operations:

- Records with missing values in the database will be deleted.
- We quantify the nominal specificities of records in the database as numerical. For example, the "profit and loss" specificity in the database has the values of "profit" and "loss", that these values are replaced by the numbers one and two.
- After converting the nominal values, all values related to the database samples are normalized using the following equation (Elson & Delen, 2014):

$$\overrightarrow{N_i} = \frac{\vec{x} - \min(\vec{x})}{\max(\vec{x}) - \min(\vec{x})} \qquad (1)$$

In the above equation $\overrightarrow{x}$ is the specificity vector for various samples of data and $\overrightarrow{N_i}$ is the normalized vector of specificities. Also, min and max are the minimum and maximum value search functions in the input vector, respectively.

After doing the preprocessing operation, three various algorithms are used to classify the specificities and to predict the profit and loss.

Predicting Profit and Loss Using Decision Tree

One of the algorithms used in the proposed method in order to predict the profit and loss is the decision tree. Decision tree algorithm is one of the data mining methods which, although it does not require complex calculations to classify data, and understanding it is also easy, its accuracy is competitive with other classification methods, so it is widely used in issues related to classification.

Each decision tree path to a leaf is usually understandable. In this sense, a decision tree can explain its predictions, which is an important advantage. Decision trees pass a small number of times from data (maximally once for each tree level) and work well with many predictor variables. As a result, models are built quickly. If we allow the tree to grow without restriction, a longer time is spent for it to be built, which is unintelligent, but the problem is that they overfit with the data. In the proposed method, the J48 algorithm has been used to build the decision tree. The most important part of the J48 algorithm is the process of creating an initial decision tree

from a set of training examples. As a result, this algorithm creates a classifier in the form of a decision tree. A structure with two types of nodes: a leaf node that represents a class or a decision node that determines some tests to run on a value of one attribute and specificity with a branch and under a tree for each possible result of the test. This process usually continues until the sets are purified, meaning that all the samples are placed in a class and at this time the growth of the tree stops (Fayyad, Shapiro, & Smyth, 1996).

Predicting Profit and Loss Using KNN Algorithm

Another model used to distinguish profit and loss is the KNN classification algorithm. In this algorithm, the K parameter is determined based on various values after repeating the test, in a way to be able to achieve the highest ratio of accuracy. Accordingly, after performing various tests, the optimal parameter value for K was determined equal to 4. The Euclidean distance model has also been used to calculate distances.

Predicting Profit and Loss Using Neural Network

The third algorithm used in the proposed method for estimating dividends is a multilayer perceptron neural network. This neural network is a multilayer perceptron network with two hidden layers. The first hidden layer of this network has 15 neurons and its transmission function has been determined of Logarithmic Sigmoid type. The second hidden layer of this neural network has 10 neurons and its transmission function has been set as a triangular basis. Also, the number of input layer neurons is determined equal to the number of specificities of each sample (P) and the number of output layer neurons is equal to 2.

The Levenberg-Marquardt backpropagation algorithm has been used to teach the neural network. This algorithm performs the network learning operation by approaching the output error towards zero and based on the Jacobian matrix.

4. Determining the Final Output in the Proposed Method Using Aggregation

The last step in order to predict the profit and loss of companies' stocks in the proposed method is to use the voting technique. The purpose of the voting technique is to improve the accuracy of the classification algorithms relative to the case in which each one of the algorithms is used separately. This solution is called learning based on voting or Ensemble. In this method, various classification algorithms are used as combination and finally the final output of the system is determined using the voting of results. Each of the classification algorithms may have errors in classifying some samples; therefore, the purpose of voting-based techniques is to reduce the resulting error and increasing the accuracy in samples classification. It has been theoretically proven that the use of voting technique can improve the results (Dietterich, 2001). In the proposed method, we use a model of applying various algorithms.

Test Scenario

The implementation of the proposed model and testing it have been performed in MATLAB software environment. The database used to evaluate the performance of the proposed model has been collected from the website of Tehran Stock Exchange, which will be described in the next section. In order to increase the accuracy of the test results, we repeat the tests 20 times. In each test, we divide the data into two categories: training and testing. In this segmentation, 75% of the database data will be used to teach the proposed model and the remaining 25% will be used to test its performance in distinguishing profit and loss. It should be mentioned that the selection of training and testing samples is random and uniform. In order to evaluate the results obtained from extracting specificity by the proposed method, the Tehran Stock Exchange database has been used.

Database

The database used in this research has been collected through a set of data provided by the stock market and securities. This database contains the data of 371 companies. The output values in this database have been specified as numbers that represent the company's profit and loss. Positive numbers indicate profit and negative numbers indicate a company's loss. Regarding that the purpose of the proposed method is merely to predict the profit or loss of a company; in doing the tests, all output values in the database have been replaced by the numbers 1 and 2. In this case, the number 1 indicates loss and the number 2 indicates profit.

This database describes the specifications of each company through 9 specificities. The list of the specifications of each specificity of this database has been mentioned in Table 1.

Table 1. database specifications.

No.	Specificity	Mean	std
1	Sales Income	6863444	**23220068**
2	Final Cost	-4865903	**20129708**
3	Gross Earnings	1899741	**6538796**
4	Administrative and Sales General Costs	-550129	**2679670**
5	Other Operating Income and Costs	54439	**1076916**
6	Total Income or Operating Costs	-495689	**2917827**
7	Operating Earnings	1445159	**4774487**
8	Financial Cost	-335670	**1363089**
9	Other Non-Operating Incomes and Costs	169500	**859472**
10	Total Non-Operating Income or Costs	-166170	**1334299**
11	Earnings Before Tax	1279978	**4478976**
12	Tax	-73766	**370744**

In this database, 82% of the samples are in the profitable category and 18% are in the harmful category. The first step in implementing the proposed approach is information preprocessing. In this step, records with missing values are specified first. The database used contained 17 records with missing values that have been ignored during the process of tests. In the following, we will express the results of implementation and evaluation of the proposed method.

5. Results Obtained from Implementation

As mentioned, the training and testing steps were repeated 20 times and each repetition time new data has been used as test data. Figure 2 shows the neural network convergence diagram in 20 repetition of the test.

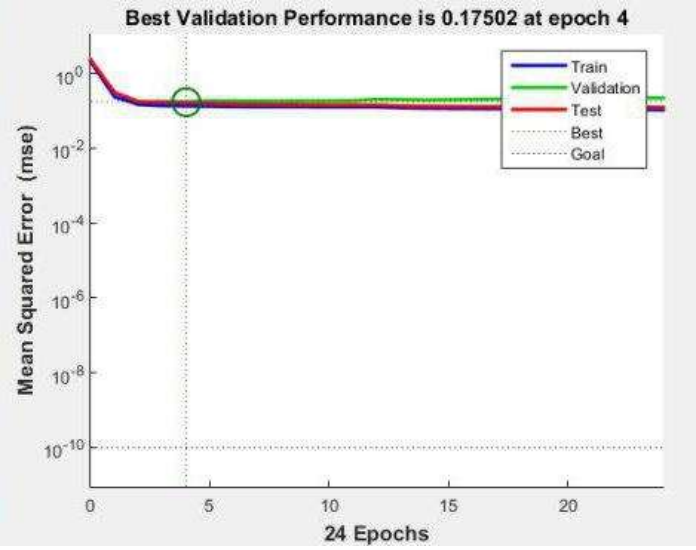

Fig. 2. Neural Network Convergence Diagram.

Figure 3 Shows The Regression Diagram Representing The Performance Ratio Of The Neural Network.

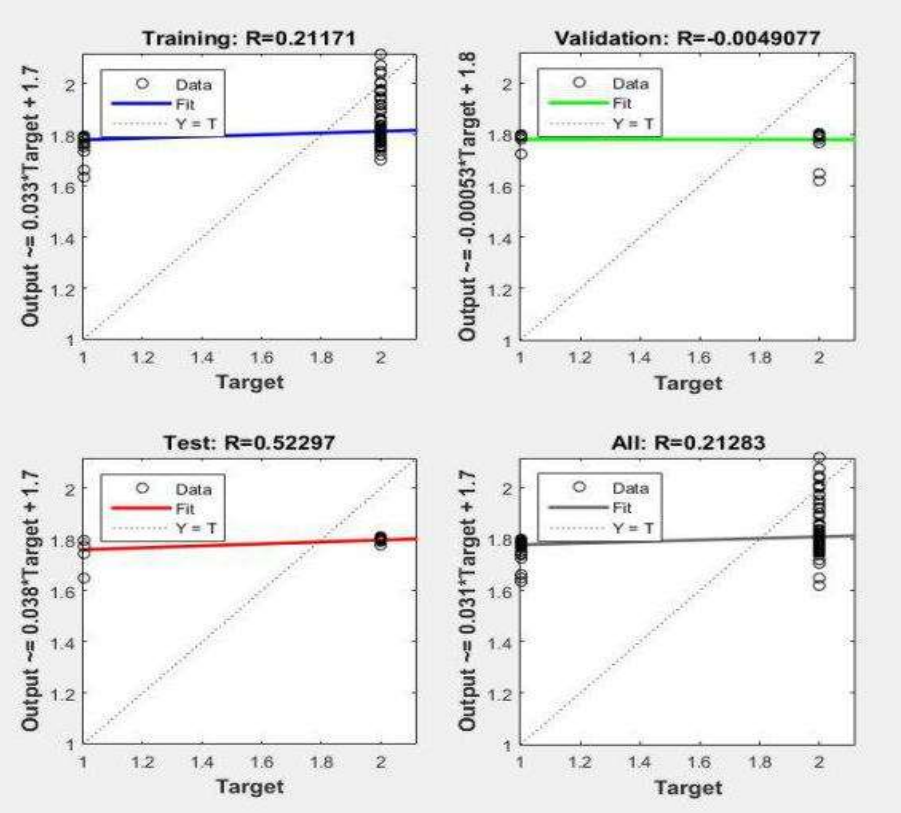

Fig. 3. neural network regression diagram.

The results obtained from this test show that using the voting technique, companies' profit and loss can be distinguished with an average accuracy of 99.43%. The results related to the accuracy of each of these algorithms have been shown in Table 2.

Table 2. Results summary of the prediction of each of the algorithms used in the proposed model.

Title	Accuracy Mean	Minimum Accuracy	Maximum Accuracy	Standard Deviation
ANN Neural Network	98.88	96.63	100	**1.153**
Decision Tree	99.39	97.75	100	**0.7025**
KNN	90.56	85.39	96.63	**3.289**
Proposed Method (Voting)	99.44	97.75	100	**0.5292**

The results shown in Table 2 indicate that the proposed method also has a more limited range of changes during various repetitions in addition to the higher accuracy mean. Limitation in the range of changes of the accuracy of an algorithm in its various repetitions is considered as an advantage. By limiting the range of accuracy changes in an algorithm, its reliability will also increase.

Figure 3 shows the confusion matrix resulting from the profit and loss distinction in 20 various repetitions for the proposed method. In the displayed confusion matrix, the number 293 in the first row and column indicates the number of harmful samples tested that have been correctly distinguished by the classification algorithm as harmful. This number is identified in the confusion matrix as TN. The number 9 in the second row and the first column indicates the number of harmful samples tested that were incorrectly distinguished by the algorithm as profitable. This number is identified in the confusion matrix as FP. The second row and the second column of the confusion matrix, which shows the number 1,477, specifies the profitability samples that are correctly classified by the algorithm and are identified as TP samples. Also, the number 1 in the first row and the second column indicates the number of profitability samples that have been incorrectly distinguished by the classification algorithm as harmful. This number is identified in the confusion matrix with the name of FN.

In the displayed confusion matrices, the number 1 indicates the harmfulness of the sample and the number 2 specifies its profitability. The results related to the confusion matrix of other classification algorithms have been shown in Figures 4a, b, and c.

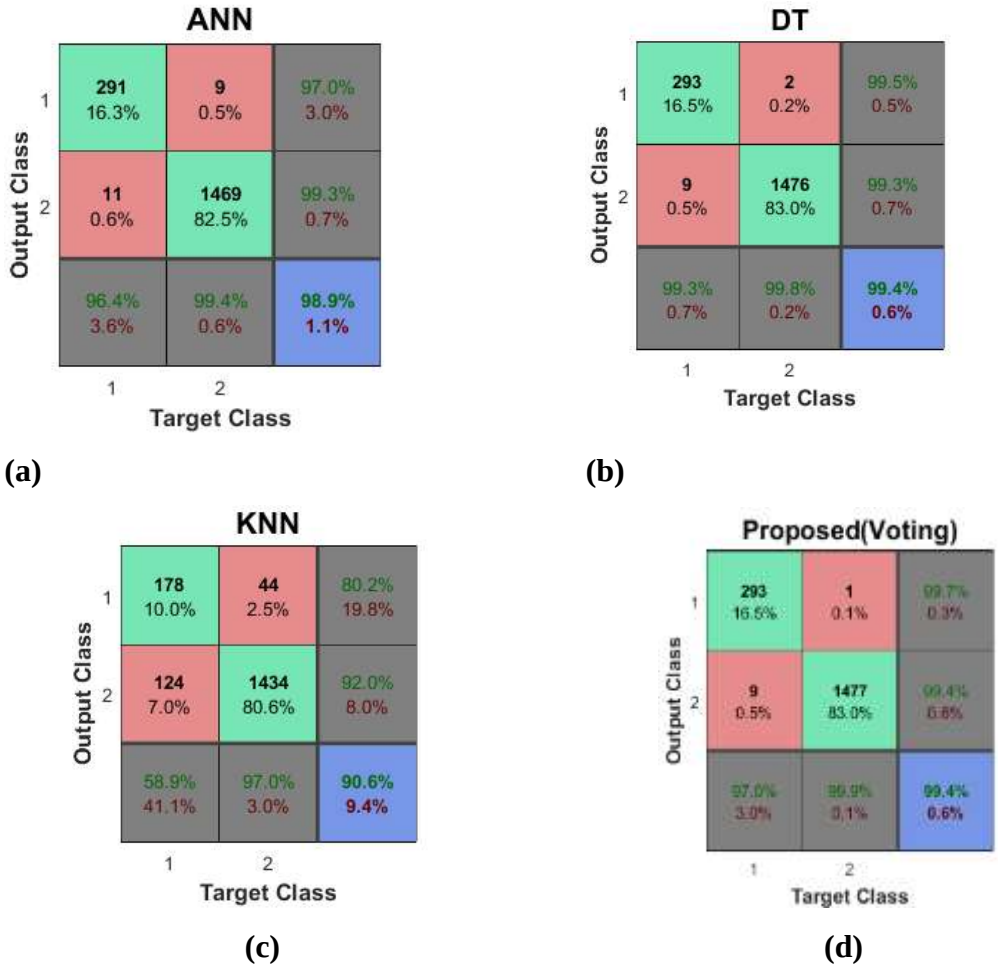

Fig. 4. confusion matrix of classification algorithms for 20 repetitions of test: (a) neural network (b) decision tree (c) k nearest neighbor, (d): proposed method.

Table 3 compares the test results obtained from the proposed algorithm for profit and loss distinction with the results of each of the classification algorithms. In this table, the criteria of sensitivity and specificity have been compared. The sensitivity criterion is for measuring the total ratio of profitability samples that have been correctly distinguished as profitable and are calculated as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

In the above equation, TP is the number of profitable samples that have been correctly distinguished and FN is the number of profitable samples that have been distinguished as harmful samples.

The specificity criterion is used to measure harmful samples that have correctly been classified. This criterion is calculated as follows:

$$pecificity = \frac{TN}{TN+FP} \quad (3)$$

In the above equation, TN is the number of harmful samples that have been correctly distinguished and FP is the number of harmful samples that have been distinguished as profitable. Finally, the accuracy criterion indicates the percentage of harmful and profitable samples that have been correctly distinguished by the algorithm. This criterion is calculated using the following equation:

$$Accuracy = 100 \times \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Table 3. comparison of the results of each of the algorithms used in the proposed model.

Title	Specificity	Sensitivity	Accuracy
Ann Neural Network	0.9926	0.9700	**98.87%**
Decision Tree	0.9899	0.9986	**99.39%**
KNN	0.9204	0.8018	**90.56%**
Proposed Method (Voting)	0.9939	0.9966	**99.43%**

These results in addition to the values of Area Under Curve (AUC) criterion for each of the classification algorithms have been shown graphically in Figure 5. The AUC criterion specifies the level below the ROC diagram for each classification algorithm.

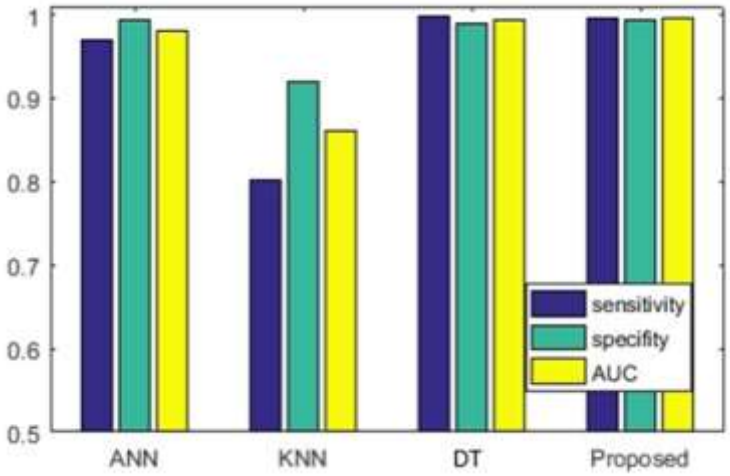

Fig. 5 comparison of the values of sensitivity, specificity and level below the roc diagram.

As the results obtained from the confusion matrix as well as Table 3 show, using the proposed method can improve the prediction ratio accuracy of each of the algorithms used.

6. Conclusion

In this research, a new method based on machine learning was presented to predict companies' profit and loss by processing their financial information. The prediction operation in the proposed method is done through three steps:

- Preparing data for classification using their normalization
- Classifying normalized data using:

- Multilayer Perceptron Neural Network
- J48 Decision Tree
- KNN Classification Algorithm

• Voting the results of each of the algorithms of previous step and determining the final output

MATLAB software has been used to implement the proposed method. Also, to test the performance of the proposed method, a database consisting of financial information of 371 companies in the stock market has been used. In doing the tests, the criteria of accuracy, sensitivity and specificity have been evaluated. The results of the tests showed that using the proposed system, the profitability or harmfulness of the company can accurately be predicted with an mean accuracy of 99.4%.

The results obtained from these tests show that the proposed method using the voting technique has better accuracy in predicting profit and loss than the algorithms discussed in this research.

Finally, it is suggested that the performance of the proposed system be examined using other similar databases. In future work, by testing other classification algorithms, the accuracy of the proposed method in predicting profit and loss can be improved. The proposed solution can be used for other applications, including the diagnosis of diseases based on people's medical information. It is also suggested that the use of clustering algorithms such as differential clustering or hierarchical clustering be investigated to develop the current research.

References

1. Asgharpour, D. M. (2012). *Multi-Criteria Decision Makings*, Vol. 7. Tehran: University Press.
2. Bezale, A., Gholami, P., Faghih, N., & Ahmadi, M. (2011). Approach to Rank Classification Algorithms with the Help of DEA. In: *Second Conference on Advanced Topics in Computer Sciences*, Zanjan Postgraduate University of Basic Sciences, Iran.
3. Chou, T., Lin , C., MChou, W., & Haung, P. (2014). Application of the Promethee technique to determine deression outler location and flow direction in DEM. *Journal of Hydrology*, *287*, 49-61 .
4. Dietterich, T. G. (1857). Ensemble methods in machine learning. *Kittler Multiple Classifier Systems, LNCS, 1857*, 1–15.
5. Elson, L., & Delen, D. (2014). *Advanced Data Mining*, Translated by Jandaghi, Gh., Hashemi, A., & Shadkam, A. University Jihad of Kerman Province.
6. Fayyad, U., Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge discovery in Databases. *AI Magazine*, *17*(3), 37-54.
7. Hwangn, C. L., & Yoon, K. (2014). *Multople Attribute Decision Making.* Springer Verlag.
8. Miller, M. Y. (2011). Dividend Policy, Growth and Determination of Stock Value. *Financial Researches*, *3*(3).
9. Saghafi, A. (2013). The Hypothesis of Effective Market of Securities and Its Impact on Accounting. In: *Proceedings of Auditor, Auditing Organization*.