Forecasting Daily Demand of Orders Using Random Forest Classifier

Ahmed Alsanad[1],*

[1] Department of Information Systems, King Saud University, Riyadh 11543, Saudi Arabia

Summary

In logistics companies, forecasting daily demand of orders is crucial in scheduling and planning tasks of supply chain to meet the consumer needs on time, improving the efficiency and reducing the costs. Even though there are some machine learning techniques have been used for predicting daily demand orders of products in logistics companies for supply chain, the choice of the most appropriate forecasting method remains a significant concern. In this paper, we investigate the application of random forest (RF) for predicting the daily demand orders of products in short time interval. We chose RF classifier in our methodology because it is a sophisticated machine learning technique that faces the strain between over-fitting and under-fitting circumstances. The methodology is evaluated on a real database of a Brazilian logistics company collected during 60 days. The RF classifier is trained on this collected dataset using 10-folds cross validation mode to predict the daily demand of orders of 6 days 10 times. The experiment show the ability the proposed classifier to predict the daily demand of orders with a high accuracy result compared to the baseline classifiers in the state-of-the-art.

Key words:

Forecasting, Daily demand of orders, Logistics companies, Supply chain, Machine learning, Random forest classifier.

1. Introduction

Forecasting daily demand of orders in logistics companies is a significant component for the supply chain management process. It plays an important role in scheduling and planning tasks of supply chain to meet the consumer needs on time, improving the efficiency and reducing the costs from the point of production to the point of consumption. Forecasting activity of products consumption in logistics companies makes it a critical bottleneck process of the supply chain. It is used to monitor, coordinate, and manage the resources required to move products in a reliable, timely and smooth cost effective manner [1]. Subsequently, the supply chain management controls the movement of products to satisfy customer needs. While the logistics companies achieve a high productivity by forecasting and predicting the daily demand of orders for products and services [1].

In order to improve the business strategy that responds to changes in demands, a robust and accurate model for predicting daily demand of orders is strongly required [2]. The usage of prediction methods to guess future trends in finance and business is known as Effective Market Hypothesis [3]. In the state-of-the-art, there are few techniques have been proposed to predict orders, prices and costs of products from the historical data.

Li et al. [4] pointed out that the sensitivity of products' prices is subjected to some external conditions. These external conditions comprise daily quotes for prices of products and service like nature gas, crude oil, gold, cotton and corn in two foreign currencies (JPY, EUR). In the experiments of this work, 2666 of trading data from U.S stocks are collected in the interval of 01/01/2000 to 10/11/2014. The dataset contains daily open, close, highest, and lowest prices, as well as the stock volume. From the external variables mentioned above and the historical data of stocks, the features are derived. In the results of experiments, logistic regression (LR) achieves a highest rate of 55.65%.

Dai and Zhang [5] used a number of 1471 instances captured from the daily stocks in the interval between 1/9/2008 and 11/8/2013 as a training dataset. Several machine learning methods are used for training the prediction model. The methods used are LR, quadratic discriminant analysis (QDA), and support vector machine (SVM) classifiers. They are utilized to predict the short term and long term prices for the next day and the next n days, respectively. The results of the next day prediction were ranging from 44.52% to 58.2%. Nevertheless, the prediction method of long term price achieved improved results, especially, if the time period was 44. In this work, the SVM attains the highest accuracy result with 79.3%.

Devi et al. [6] proposed a hybrid model combines cuckoo search with Gaussian kernel SVM. Cuckoo search algorithm is used as a method for initializing the SVM's parameters. Giacomel et al. [7] proposed a prediction model for trading agent using neural network ensemble. The model is used to predict if a stock is going to fall or rise. The evaluation of this model was on two datasets: The Brazilian stock market and the North American. Boonpeng and Jeatrakul [8] applied a One-against-One (OAO) and One-against-All (OAA) neural network model to predict Sell, hold or Buy data. The performance of this model is also compared with the traditional neural network. The historical data of seven years, from 03/01/2007 to 29/08/2014 is collected from stock exchange of Thailand

Manuscript received April 5, 2018
Manuscript revised April 20, 2018

(SET) and used as a training data. The accuracy results of OAA-NN were higher than traditional NN and OAO-NN methods, achieving an accuracy result of 72.50%.

Ferreira et al. [9] introduced a study for forecasting delay demands of orders using artificial neural network (ANN). ANN is a powerful model for regression and classification tasks inspired from the function of biological neurons [3]-[5]. However, ANN faces the over-fitting and local minima problems, as well as it has many parameters which need to be optimized. In this paper, we propose an effective prediction model for forecasting daily demand of orders using random forest (RF) classifier. The model is able to learn from historical data and face the over-fitting and under-fitting problems.

The subsequent parts of the paper is as follows: Section 2 explains the relationship between data analysis and machine learning as well as a background of RF classifier. The proposed methodology is introduced in section 3. Section 4 presents the experimental results and discussion. Finally, section 4 demonstrates the conclusion and future work.

2. Data Analysis and Machine Learning

Data analysis is an effective method to solve many problems in several domains. For example, Gumaei et al. used the data analysis method to estimate the cost of software products in Saudi Arabia software industry [10, 11]. Nowadays, modern scientific methodologies depend on the data analysis and machine learning to form an integrative effective part. They offer an automated procedure for building a prediction model which uses the past observations and patterns in data to provide a notion about the problem. Machine learning is a subfield of artificial intelligence (AI) as shown in Figure 1.

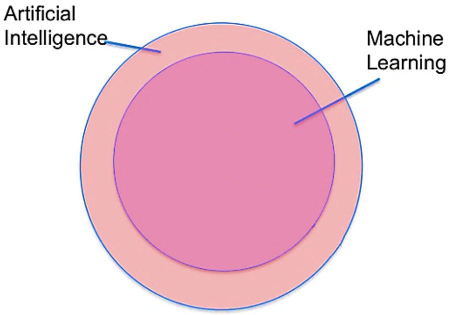


Fig. 1 AI and machine learning disciplines.

It learns composite relations in data, aiming to build models that produce correct predictions and extract the knowledge in a more perceivable way. In the next subsection, we give a background on one of the most important methods of machine learning which is used in the proposed methodology of this study.

2.1 Random Forest Classifier

Random Forest (RF) is an effective algorithm used for the purpose of regression and classification. It follows the ensemble learning concept [12]. The basic idea of the method is to build a set of decision trees. RF uses in its work two bases of randomness are:

1. Each decision tree is individually and randomly built on a different sample of the training dataset.

2. Throughout the construction process of each tree, a subset of m instances is randomly chosen from the original training dataset and the best split according to these m instances is utilized. For a new case c, the prediction of the RF is constructed by the decision trees which are aggregated. In RF that includes N decision trees, the forest output probability of the class label y for a new case c given a feature vector x is computed using the following ensemble learning model:

$$P(y/x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y/x) \qquad (1)$$

It averages the output probability of decision trees generated from the random samples of the dataset. Figure 2 visualizes the ensemble learning model of RF.

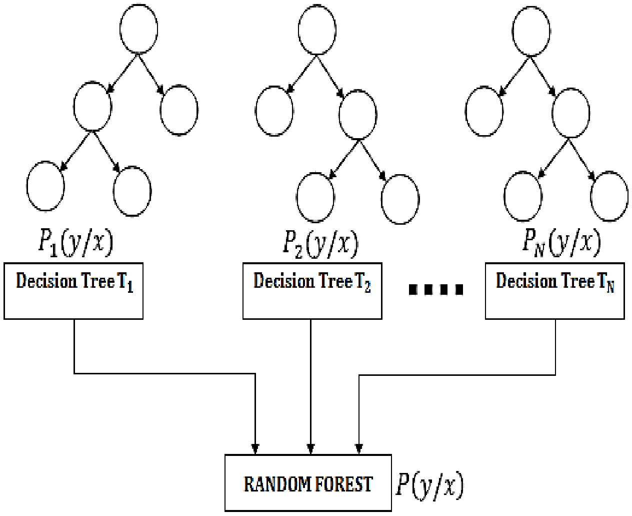
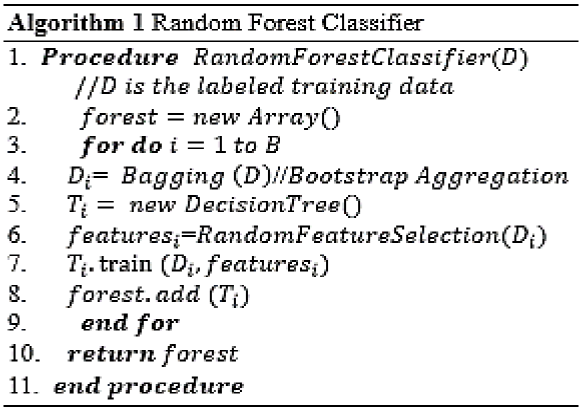


Fig. 2 An ensemble learning model of RF.

Each decision tree in RF divides the dataset down into smaller subsets, building a subtree with leaf nodes of decision nodes. A decision node has two or more branches with leafs. Each category in the dataset is represented by the leaf node. Since the decision trees in RF depend on random data, they can lack meaning and may be noisy. Therefore, the RF averages these decision trees to build a low variance model. The unfitting decision trees can cancel out each other and the useful trees are used to create the resulted model. Effective approaches based on RF have been appeared in a wide range of real-life applications, such as network intrusion detection [13], image classification [14], and neuroimaging [15]. Algorithm 1 describes the steps of RF classifier.



Algorithm 1 Random Forest Classifier

```
Procedure $RandomForestClassifier(D)$
    //D is the labeled training data
    $forest = new\ Array()$
    for do $i = 1$ to $B$
        $D_i$= Bagging (D)//Bootstrap Aggregation
        $T_i$ = new DecisionTree()
        $features_i$=RandomFeatureSelection($D_i$)
        $T_i$.train ($D_i$, $features_i$)
        forest.add ($T_i$)
    end for
    return forest
end procedure
```

In this study, we investigate the application of RF for forecasting daily demand of orders in logistics companies.

3. Proposed Methodology

The proposed methodology aims to build a robust prediction model that attains high accuracy for forecasting daily demand of orders. It involves three main steps as presented in Figure 3.

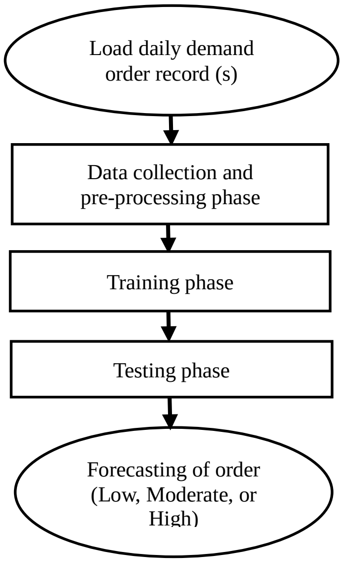

Load daily demand order record (s)
Data collection and pre-processing phase
Training phase
Testing phase
Forecasting of order (Low, Moderate, or High)

Fig. 3 Proposed methodology for orders forecasting.

In the following subsection, we explain each phase of the proposed methodology.

3.1 Data Collection and Pre-processing Phase

In this phase of our proposed methodology, we use the daily demand forecasting orders dataset collected in [9]. The features used in this dataset are the week of the month (first week, second, third, fourth or fifth week), ‘Day of the week (Monday to Friday)’, ‘Non-urgent order’, ‘Urgent order’, ‘Order type A’, ‘Order type B’, ‘Order type C’, ‘Fiscal sector orders’, ‘Orders from the traffic controller sector’, ‘Banking orders (1)’, ‘Banking orders (2)’, ‘Banking orders (3)’, and ‘Target (Total orders)’. We categorize the ‘Target (Total orders’) into three classes according to the total number of orders. The class High is for the total number of orders which is more than 400 orders. The class Moderate is for the total number of orders which is less than 400 orders and more than 250 orders. The class Low is for the total number of orders which is less than 250 orders. The distribution of the collected dataset is shown in Figure 4.

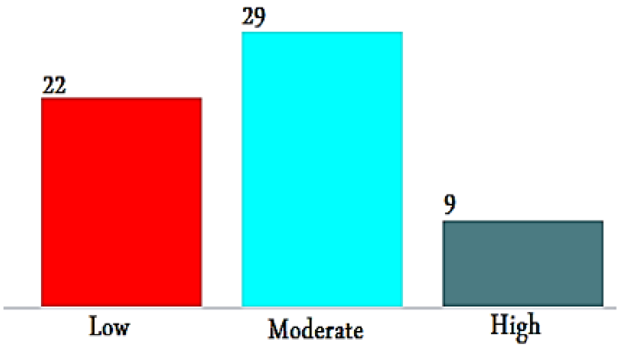


Fig. 4 Distribution of orders in the dataset.

3.2 Training Phase

Here, we train the RF classifier on the collected dataset using 10-folds cross validation mode. In 10-folds cross validation, the dataset can be divided into 10 folds and the RF classifier is trained using 9 out of 10 folds 10 times. The remaining 1 fold is used in the testing phase. The result of this phase is 10 trained RF classifier models.

3.3 Testing Phase

In testing phase, every time of 10 times we train the RF classifier on the 9 folds, we test it on the remaining fold. The result of these 10 times will be averaged using weighted average method to get the final result of the trained RF classifier.

4. Experimental and Discussion

4.1 Implementation Tool Description

Waikato Environment for Knowledge Analysis (WEKA) [16] is employed to implement our methodology. It is a popular tool for machine learning and data mining fields.

4.2 Evaluation Measurements

Three common evaluation measurements are used to evaluate the results. These measurements are the precision (PR), recall (RE), and the accuracy (ACC).

$$PR = TP/(TP + FP) \quad (2)$$

$$RE = TP/(TP + FN) \quad (3)$$

$$ACC = (TP + TN)/(TP + FP + TN + FN) \quad (4)$$

, where FP and FN are the false positive and negative rates, whereas TP and TN are the true positive and negative rates.

4.3 Results and Comparisons

The results computed for evaluation are based on the precision, recall, and the accuracy of the testing phase. Five baseline classifiers are adopted to prove the effectiveness of the proposed classifier. These five classifiers are: (1) Bagging of decision trees (BDT), (2) artificial neural network (multilayer perceptron (MPL)) used in [9], (3) linear support vector machine (LSVM), (4) decision tree (J48), and (5) K-nearest neighbors (KNN). Table 1 and Figures 5-7 show the results of PR, RE, and ACC for the proposed classifier compared to the baseline classifiers.

Table 1: Results of TF-IDFT representation with the eight classifiers.

Classifier Model	Accuracy (%)	Weighted Avg. of Precision	Weighted Avg. of Recall
BDT	81.667	0.822	0.817
ANN []	85	0.851	0.850
LSVM	81.667	0.819	0.817
DT	86.667	0.868	0.867
KNN	73.333	0.739	0.733
Proposed RF	**88.333**	**0.894**	**0.883**

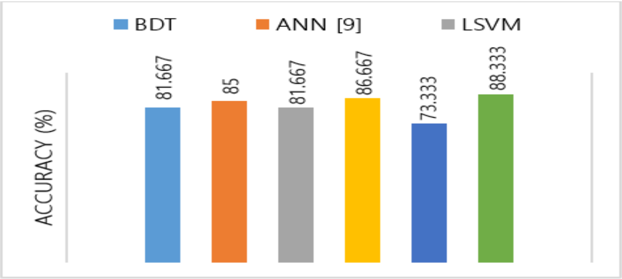


Fig. 5 Accuracy of proposed RF compared to other classifiers.

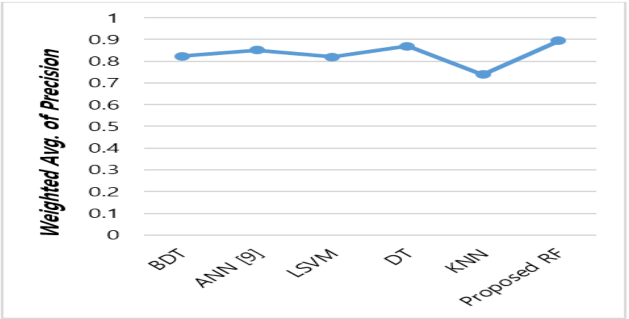


Fig. 6 The Average of weighted avg. precision of proposed RF compared to other classifiers.

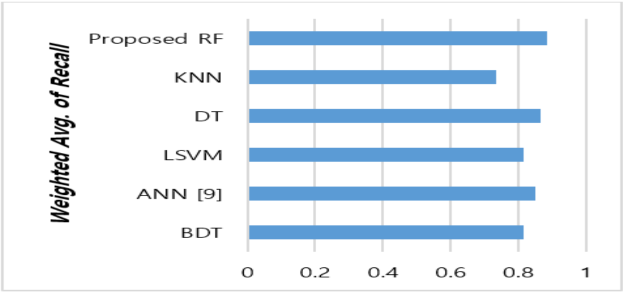


Fig. 7 The Average of weighted avg. recall of proposed RF compared to other classifiers.

From Table 1 and Figures 5-7, we see that RF classifier achieves the highest accuracy result of 88.333%, compared to all other baseline classifiers. Moreover, it attains the highest precision and recall results with 0.894 and 0.883, respectively, against the other classifiers. They also show that the DT classifier achieves the second highest accuracy result with 86.667%, compared the baseline classifiers.

5. Conclusion and future work

In this paper, we proposed an effective methodology to use RF classifier for predicting the daily demand of orders in logistics companies. RF classifier can use a number of decision trees and make the prediction probability output by averaging the prediction outputs of all decision trees. It normally achieves much better accuracy result than using a decision tree classifier as a lone. To evaluate the proposed classifier, we used 10-folds cross validation for training and testing.

The experiment is implemented using WEKA tool. Precision, recall, and accuracy are employed as evaluation measurements to assess the output results of the classifier. Five baseline classifiers are adopted to compare the effectiveness of the RF classifier for predicting the daily demand of orders. The experimental results demonstrated the ability and robustness of RF classifier compared to the baseline classifiers in the state-of-the-art.

In the future work, we will propose a new SL method to classify Arabic text based on the effect of IDFT method proved in this paper. In future work, we solve the problem of K-most demanding products investigated in [17] using the proposed algorithm in [18] and RF classifier proposed in this work.

Acknowledgment

The author is thankful to the Deanship of Scientific Research, College of Computer and Information Sciences (CCIS) at King Saud University for funding this research.

References

[1] Christopher, M. (2016), “Logistics & supply chain management”, Pearson UK.

[2] Kelley, E. K., and Tetlock, P. C. (2013), “How wise are crowds? Insights from retail orders and stock returns”, The Journal of Finance, Vol. 68, No. 3, pp. 1229-1265.

[3] Degutis, A., and Novickyte, L. (2014), “The efficient market hypothesis: a critical review of literature and methodology”, Ekonomika, Vol. 93, No. 2, pp. 7.

[4] Li, H., Yang, Z., and Li, T. (2014), “Algorithmic Trading Strategy Based On Massive Data Mining”, Stanford University.

[5] Dai, Y., and Zhang, Y. (2013), “Machine Learning in Stock Price Trend Forecasting”, Stanford University.

[6] Devi, K. N., Bhaskaran, V. M., and Kumar, G. P. (2015, March), “Cuckoo optimized SVM for stock market prediction”, In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on (pp. 1-5). IEEE.

[7] Giacomel, F., Galante, R., and Pereira, A. (2015, December), “An algorithmic trading agent based on a neural network ensemble: a case of study in North American and Brazilian stock markets”, In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on (Vol. 2, pp. 230-233). IEEE.

[8] Boonpeng, S., and Jeatrakul, P. (2016, February), “Decision support system for investing in stock market by using OAA-neural network”, In Advanced Computational Intelligence (ICACI), 2016 Eighth International Conference on (pp. 1-6). IEEE.

[9] Ferreira, R. P., Martiniano, A., Ferreira, A., Ferreira, A., & Sassi, R. J. (2016), “Study on daily demand forecasting orders using artificial neural network”, IEEE Latin America Transactions, Vol. 14, No. 3, pp. 1519-1525.

[10] Gumaei, A., Almaslukh, B., and Tagoug, N. (2015), “An Empirical Study of Software Cost Estimation in Saudi Arabia Software Industry”, International Journal of Soft Computing and Engineering (IJSCE), Vol. 4, No. 6, pp. 2231-2307.

[11] Jasmine, K. S., and Vasantha, R. (2008), “An automated environment for design based performance prediction of component based software products”, International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 116-120.

[12] Breiman, L. (2001), “Random Forests. In Machine Learning”, Vol. 45, No. 1, pp. 5-32.

[13] Zhang, J., Zulkernine, M., and Haque, A. (2008), “Random-forests-based network intrusion detection systems”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 38, No. 5, pp. 649-659.

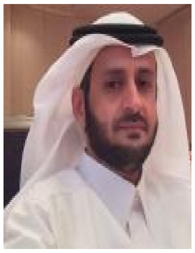
[14] Du, P., Samat, A., Waske, B., Liu, S., and Li, Z. (2015), “Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features”, ISPRS Journal of Photogrammetry and Remote Sensing, 105, pp. 38-53.

[15] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., and Alzheimer's disease Neuroimaging Initiative. (2013), “Random forest-based similarity measures for multi-modal classification of Alzheimer's disease”, NeuroImage, 65, pp. 167-175.

[16] WEKA, “Data Mining Software in Java”, 2017. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka. [Accessed: 15- March- 2018].

[17] Gumaei, A., Sammouda, R., and Al-Salman, A. S. (2017), “An Efficient Algorithm for K-Rank Queries on Large Uncertain Databases”, International Journal of Computer Science and Network Security (IJCSNS), Vol. 17, No.4, pp. 129.

[18] Bang, S., & Kalavadekar, P. N. (2014). “Determining K-most demanding products using data mining technique”, International Journal of Computer Science and Network Security (IJCSNS), Vol. 14, No. 6, pp. 18.



AHMED ALSANAD is an Assistant Professor of Information System Department and chair member of Pervasive and Mobile Computing, CCIS, at the King Saud University, Riyadh, KSA. He received his Ph.D. degree in Computer Science from De Montfort University, Unit Kingdom in 2013. His research interests include Cloud Computing, Health Informatics, ERP and CRM. He has authored and co-authored more than 12 publications including refereed IEEE/ACM/Springer journals, conference papers, and book chapters.