khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Analysis of Slow Moving Goods Classification Technique: Random Forest and Naïve Bayes

Deny Jollyta*[1], Gusrianty[1], Darmanta Sukrianto[2]
[1]Program of Informatics
Sekolah Tinggi Ilmu Komputer Pelita Indonesia
Pekanbaru, Indonesia
[2]Program of Informatics
AMIK Mahaputra Riau
Pekanbaru, Indonesia
*deny.jollyta@lecturer.pelitaindonesia.ac.id

Abstract-Classifications techniques in data mining are useful for grouping data based on the related criteria and history. Categorization of goods into slow moving group or the other is important because it affects the policy of the selling. Various classification algorithms are available to predict labels or class labels of data. Two of them are Random Forest and Naïve Bayes. Both algorithms have the ability to describe predictions in detail through indicators of accuracy, precision and recall. This study aims to compare the performance of the two algorithms, which uses testing data of snacks with labels for packaging, size, taste and category. The study attempts to analyze data patterns and decides whether or not the goods fall into the slow moving category. Our research shows that Random Forest algorithm predicts well with accuracy of 87.33%, precision of 85.82% and recall of 100%. The aforementioned algorithm performs better than Naïve Bayes algorithm which attains accuracy of 84.67%, precision of 88.33% and recall of 92.17%. Furthermore, Random Forest algorithm attains AUC value of 0.975 which is slightly higher than that attained by Naïve Bayes at 0.936. Random Forest algorithm is considered better based on the value of the metrics, which is reasonable because the algorithm does not produce bias and is very stable.

Keywords: slow moving; random forest; naïve bayes;

1. Introduction

Goods can be classified based on its circulations over a certain period of time and goods with very slow circulation are called slow moving goods [1]. Slow moving goods have be stored in warehouses in large quantity. Slow moving goods are materials that circulate with the speed of one item within a year [2]. Classification problems associated with slow moving goods occur due to lack of analysis of previous data [3]. Analysis can be conducted using classification algorithms of data mining. Classifications create patterns through analysis of the closeness of labels or attributes that construct item data. The resulting patterns are the predictions of slow moving goods.

In this study, Random Forest and Naïve Bayes are the classification algorithms that are used, which work on data of packaged snacks. Both algorithms were chosen because they can produce accurate predictions with descriptions that highly agree with actual situations. Many studies have been carried out that relate to the two algorithms for classification. In [4], Random Forest was used to analyzing multispectral images by classifying points in images. Taxonomy of Random Forest algorithm has been described in [5] through several parameters such as the base of classifications, size division, number of tracks, combination of strategies, number of attributes, criteria, cut-off ability, additional classifications, and number of datasets used in training phase. In addition, Random Forest algorithm has highlighted the advantages and benefits in prediction on large datasets [6].

The ability of Naïve Bayes algorithm has been tested in various data predictions including to predict the behavior of the purchase on transaction time [7]. The pattern shows that more buyers make transactions in the afternoon, particularly on Sundays. Naïve Bayes algorithm has been used to group blogger data [8] and banking product marketing data [9] - [10] to assist banks to find potential customers. The performance of Naïve Bayes algorithm has been compared with other classification algorithms such as K-Nearest Neighbor (KNN) algorithm and Decision

Tree [11] to group data of school students who consume alcohol. Despite the differences, Naïve Bayes' performance has shown better accuracy than the other two algorithms.

The results of previous studies in using Random Forest and Naïve Bayes algorithms motivate an attempt to observe both algorithms in classification of slow moving goods. The result is valuable for decision makers to implement policies related to such goods. A comparison is required for a clear picture of the performance of both algorithms.

2. Theory

a. Random Forest Algorithm

Random Forest algorithm is an ensemble model that was created and developed by Tin Kam Ho [12]. It belongs to supervised learning and works based on calculations of various models to obtain results [6]. As an ensemble model, Random Forest is able to build decision trees and uses its rules for the calculation of the final result, following formula (1) [13].

$$hj(X,\Theta j)$$
(1)

Having processed training data, predictions are obtained from the average results of all trees, using formula (2).

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$
(2)

In its various applications, Random Forest algorithm is widely used for its advantages, such as better accuracy, resistance to various disturbances, speed, and convenience in implementation [14].

b. Naïve Bayes Algorithm

Naïve Bayes algorithm is a simple probabilistic classification technique based on the application of the Bayes theorem with strong assumptions [15]. Naïve Bayes is applied to a limited number of data to get the appropriate parameters of classification. Naïve Bayes formula is expressed using formula (3) [11].

$$P(H \mid E) = \frac{P(H)\prod_{i=1}^{a} P(E_i \mid H)}{P(E)} \quad (3)$$

P (H|E) = data probability with vector E in class H.
P (H) = initial probability of class H
$\prod_{i=1}^{a} P(E_i \mid H)$ = independent probability of class H from all features in vector E

The advantages of the Naïve Bayes algorithm include the ability to handle quantitative and discrete data, resistance to isolated noise points, sufficiency of small number of training data, ability to handle missing values by neglecting instances during the calculation of estimated probability, speed, efficiency in space, and robust against irrelevant attributes.

3. Method

a. Training Data

The use of the two algorithms in this research was administered by employing two different tests using RapidMiner application. RapidMiner is an application in the field of data mining such as machine learning, information mining, and content mining [16]. In this study, RapidMiner is used to display the performance of the two algorithms using the data of packaged snacks. Data items were taken randomly for as many as 150 data. Data have to pass a selection process in accordance with the stages of Knowledge Discovery in Database (KDD). The data were arranged based on several attributes considered to affect most on the speed of items transactions, such as packaging, size, taste, and category. Attributes description is shown in Table 1.

Table 1. Attributes of Training Data

Attribute	Description
Item Code	Code of each packaged snack
Item Type	Type of snacks such as candies, jelly, gum drop, etc.
Item Name	Name of packaged snack.
Taste	Taste of the packaged snack such as sweet, spicy, sweet and sour, etc.
Packaging	Shape and material of packaging, namely plastic, bottle, and can.
Size	Size of package such as small, medium, and large.
Category	Label regarding snack resistance, such as premature spoilage, fragile, and resistant.
Slow Moving	Label regarding transaction flow, such as Yes for slow moving and No for fast moving

c. Research Framework

To produce a result of prediction of slow moving goods, the research goes through several steps as shown in Figure 1. The first stage is data preparation, which follows the stages of the Knowledge Discovery in Database (KDD). Data selection consumes a huge amount of time in order to adjust with the classification algorithms, i.e. Random Forest and Naïve Bayes. Data have to pass the KDD stages to obtain proper quality of training data. The KDD selection produces training data as described in Table 1.

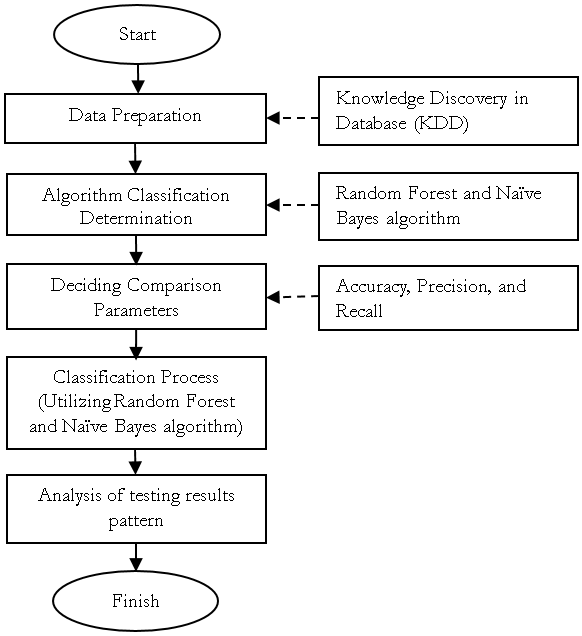

Figure 1. Research Framework

The next step is the selection of parameters as a measure to compare the performance of test results. We choose accuracy, precision, and recall, which are taken from the Gain Ratio criteria in RapidMiner. The choice of gain ratio as the comparison results parameter concerns more about its ability to calculate every data in the available sample space. The parameter selection is intended to see the comparison of the results of testing the two algorithms. In addition to the three parameters, the test results are displayed in accordance with the algorithm features. Training produces patterns that may be analyzed to obtain predictions that reflect the actual situation. This step provides the best prediction results for each of the two algorithms.

4. Results and Discussion

Research results on the two algorithms are described in the following two sections: the prediction results and the parameter results.

a. Prediction Results

Item data were tested on both algorithms by using 10 iterations. Each iteration produced a different tree structure. Confidence was displayed to indicate the level of confidence of each attribute in producing the decision whether the items belong to either slow moving or non–slow moving category. The gain ratio criterion was used as a measure to read the test results of Random Forest. There are some discrepancies in the test results, especially on the target attribute "No". This is because confidence value of the "No" is higher than the target attribute "Yes". Of 150 data, there arises 19 discrepancy data, which implies a value of 12.67% error rate of the calculation results. The rules resulting from the calculation of the Random Forest algorithm are described in Table 2.

Table 2. The Slow Moving Goods Prediction Rules of Random Forest

No	Rules
1	If item type=candy ^ taste=sweet and sour ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
2	If item type=candy ^ taste=spicy ^ packaging=can ^ size=big ^ category=resistant, then slow moving=Yes
3	If item type=candy ^ taste=spicy ^ packaging=plastic ^ size=large ^ category=premature spoilage, then slow moving=Yes
4	If item type=dried sunflower seeds ^ taste=sweet ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
5	If item type=sponge cake ^ taste=salty ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes

Table 3. The Slow Moving Goods Prediction Rules of Naïve Bayes

No	Rules
1	If item type=candy ^ taste=sweet and sour ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
2	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=fragile, the slow moving=Yes
3	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=resistant, then slow moving=Yes
4	If item type=candy ^ taste=spicy ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
5	If item type=candy ^ taste=spicy ^ packaging=can ^ size=small ^ category=resistant, then slow moving=Yes
6	If item type=candy ^ taste=spicy ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes
7	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=small ^ category=fragile, then slow moving=Yes
8	If item type=candy ^ taste=sweet ^ packaging=bottle ^ size=large ^ category=fragile, then slow moving=Yes
9	If item type=snack ^ taste=spicy ^ packaging=plastic ^ size=small ^ category=premature spoilage, then slow moving=Yes
10	If item type=snack ^ taste=spicy ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
11	If item type=sponge cake ^ taste=sweet ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes
12	If item type=sponge cake ^ taste=salty ^ packaging=can ^ size=large ^ category=resistant, then slow moving=Yes
13	If item type=biscuit ^ taste=milky ^ packaging=plastic ^ size=large ^ category=resistant, then slow moving=Yes

Based on Table 2, the calculation shows an accuracy of 87.33%. This value is the conclusion of the

accumulation of each decision tree produced through Random Forest. Calculation of gain ratio for positive class = Yes is 45.71%, while for positive class = No is 100%.

The prediction for the Naïve Bayes algorithm using the gain ratio shows a lower error rate at 8.67%. There are 13 different data. The differences between the prediction and initial data are mostly on data with attributes "No" which become "Yes" according to Naïve Bayes calculation. This means that items, which are not originally included in the slow moving category, fall into the category. The confidence value is lower here so it changes data with class attributes "No" to become "Yes". Naïve Bayes produces a model of slow moving attributes into 2 classes previously mentioned with respective value of 0.767 for the "No" class and 0.233 for the "Yes" class. The rules resulting from the calculation of the Naïve Bayes algorithm are in Table 3.

Based on Table 3, Naïve Bayes produces an accuracy of 84.67%. Calculation of gain ratio for positive class = Yes is 60% while for positive class = No is 92.17%.

b. Accuracy, Precision, and Recall Parameters

The implementation of gain ratio criteria in this training stage produces detailed calculations in the form of confusion matrix. Both algorithms reveal patterns that are hidden in the training data. Running RapidMiner with operator Performance produces results in values of metrics in 3 parameters: Accuracy, Precision and Recall, as shown in Table 4.

Table 4. Parameter Calculation Results

Parameter	Random Forest Algorithm	Naïve Bayes Algorithm
Accuracy	87.33%	84,67%
Precision	85.82%	88.33%
Recall	100%	92.17%

Entries of Table 4 show that accuracy and recall parameters of the Random Forest algorithm are higher than the Naïve Bayes algorithm. However, the precision of the Naïve Bayes algorithm is higher. Hence the Random Forest algorithm is superior in two of three metrics against Naïve Bayes algorithm. To further decide which classification algorithm is better, we need to observe the Receiver Operating Characteristic (ROC) curve and calculate the Area under the ROC Curve (AUC) [7]. An ROC curve expresses confusion matrix data, in which the horizontal line represents false positive (FP) values and the vertical line represents true positive (TP) values.

Figure 2 is an ROC curve obtained from the calculation of the Random Forest algorithm with the acquisition of AUC values of 0.975. In [8], AUC was used to measure discriminative performance by predicting the possibility of the emergence of output from random samples for positive and negative populations. The greater the AUC, the firmer the classification be recommended. AUC is part of the square unit area, AUC value will always be between 0.0 and 1.0.

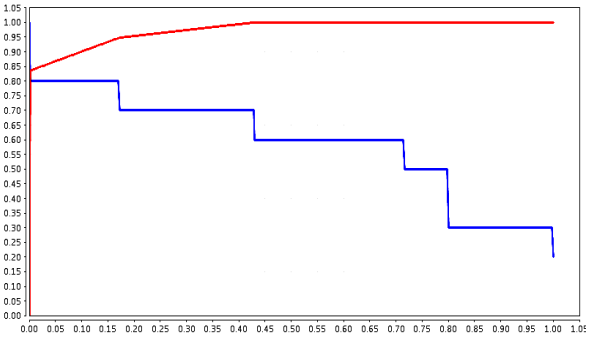

Figure 2. Random Forest Area under ROC Curve (AUC)

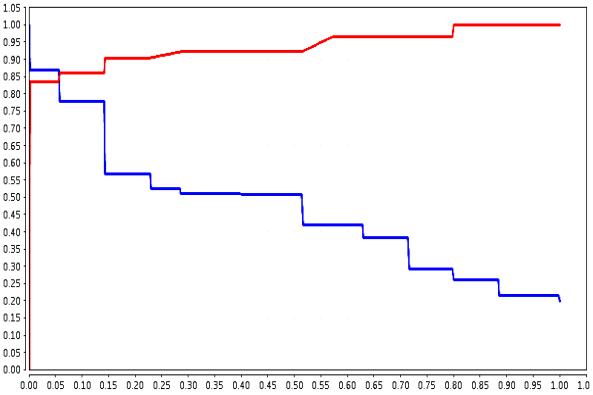

Figure 3. Naïve Bayes Area under ROC Curve (AUC)

Figure 3 is the ROC curve from the calculation of the Naïve Bayes algorithm with AUC of 0.936. The AUC for Random Forest algorithm at 0.975 is slightly higher. However, both algorithms behave as a nearly perfect classification model with AUC values close to 1.00.

c. Discussion

Both algorithms show good performance but with different results. The overall results of testing the item data with the Random Forest and Naïve Bayes are in the following Table 5.

Metrics in Table 5 show that the performance of the Random Forest algorithm is generally better. Random Forest algorithms produces a tree structure in each iteration that is easy to compare with structures in other iterations. The most results from each structure become the final result. The ability of Random Forest algorithm to analyze the results of each decision tree in 10 iteration has apparently produce higher accuracy than the Naïve Bayes algorithm. The dominantly similar rule in every iteration is one of the advantages of Random Forest, which may support its performance to achieve a high accuracy [17]. The recall value reaching 100% and the AUC value of 0.975 have brought the Random Forest as the best choice for classification of slow moving goods. Therefore, attributes that are considered responsible to cause a goods become slow moving are taken from those identified by Random Forest algorithm.

Table 5. Comparison of Random Forest and Naïve Bayes Performance

No	Indicator		Random Forest	Naïve Bayes
1	Number of Rules		5	13
2	Prediction Error	Total of Data	19	13
		Percentage	12.67%	8.67%
		Accuracy	87.33%	84.67%
3	Parameters	Precision	85.82%	88.33%
		Recall	100%	92.17%
4	Gain Ratio	Positive class Yes	45.71%	23.3%
		Positive Class No	100%	76.7%
5	AUC Value		0.975	0.936

4. Conclusion

We have observed two algorithms: the Random Forest and Naïve Bayes algorithms to classify data on packaged snacks and to identify which attributes supports the class label of slow moving. Calculation using RapidMiner on both algorithms give predictions with almost similar accuracy. The difference in the precision value of the two algorithms of 2.51% suggests that Naïve Bayes algorithm has better accuracy in slow moving goods in the training data. This is shown by the smaller prediction errors than that of Random Forest algorithm, and because the confidence values tend to be identical. However, Random Forest algorithm is more reliable to get a precise prediction because it may be obtained from several decision trees. This research shows that Random Forest algorithm provides better predictions to reflect actual conditions with a limited number of data. A total of 5 rules were produced, showing perfect compatibility with the actual situation of packaged snacks, which is 100%.

Acknowledgment

Acknowledgment is addressed to Sekolah Tinggi Ilmu Komputer Pelita Indonesia Pekanbaru which supports the completion and publication of this research.

References

[1] Rajahstan, *Reading Material Drug Store Management Rational Drug Use For Medical Officers , Nurses & Pharmacists*, no. December. 2010.

[2] D. Janari, M. M. Rahman, and A. R. Anugerah, "Analisis Pengendalian Persediaan Menggunakan Pendekatan Music 3D (Muti Unit Spares Inventory Control- Three Dimensional Approach) Pada Warehouse Di PT Semen Indonesia (PERSERO) TBK Pabrik Tuban," *Teknoin*, vol. 22, no. 4, pp. 261–268, 2016.

[3] G. Chodak, "The Nuisance of Slow Moving Products in Electronic Commerce," *MPRA Munich Pers. RePEc Arch.*, vol. 70141, no. 3, pp. 1–7, 2016.

[4] B. Lowe and A. Kulkarni, "Multispectral Image Analysis Using Random Forest," *Int. J. Soft Comput.*, vol. 6, no. 1, pp. 1–14, 2015.

[5] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," *Int. J. Adv. Comput.*, vol. 36, no. 1, pp. 1144–1156, 2013.

[6] N. Horning, "Random Forests: An algorithm for image classification and generation of continuous fields data sets," in *International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences 2010*, 2010, pp. 1–6.

[7] Susanto, E. D. S. Mulyani, and I. R. Nurhasanah, "Penerapan Data Mining Classification Untuk Prediksi Perilaku Pola Pembelian Terhadap Waktu Transaksi Menggunakan Metode Naïve Bayes," in *Konferensi Nasional Sistem dan Informatika (KNS&I)*, 2015, pp. 313–318.

[8] Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 20–28, 2018.

[9] I. Oktanisa and A. A. Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 567–576, 2018.

[10] N. H. Niloy and M. A. I. Navid, "Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients," *Am. J. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 1–12, 2018.

[11] N. Sagala and H. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 2, pp. 98–103, 2018.

[12] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," *Int. J. Adv. Electr. Comput. Eng.*, vol. 3, no. 4, pp. 5–7, 2016.

[13] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," in *Random Forest*, no. January, 2011, p. 21.

[14] E. Goel and E. Abhilasha, "Random Forest: A Review," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 7, no. 1, pp. 251–257, 2017.

[15] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787–

795, 2013.

[16] S. Dixit and S. Kr, "Collaborative Analysis of Customer Feedbacks using Rapid Miner," *Int. J. Comput. Appl.*, vol. 142, no. 2, pp. 29–36, 2016.

[17] K. . Ghose, R. Pradhan, and S. S. Ghose, "Decision Tree Classification of Remotely Sensed Satellite Data using Spectral Separability Matrix," *Int. J. Adv. Comput. Sci. Appl.*, vol. 1, no. 5, pp. 93–101, 2010.