Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product

Ikhsan Romli1*, Toga Pardamean2, Sufajar Butsianto3, Tri Ngudi Wiyatno4, Effendi bin Mohamad5

Information Technology, Pelita Bangsa University1,2,3,4
Manufacturing Engineering, Technical University of Malaysia Melaka5

*ikhsan.romli@pelitabangsa.ac.id

Abstract. In the era of progressively more competitive industrial competition, especially in the manufacturing world, it is always required to develop the quality or quality of products and productivity. Each company is competing to win market share. One of the strategies carried out by the company is improving the quality of products and the production process conducted by the company. In the industrial world, product quality and productivity are the keys for success of the production process. Therefore, the purpose of this study is to analyze data for defective products at PT Mane Indonesia with the Particle Swarm Optimization (PSO) and Naïve Bayes Classifier method. The accuracy results using the Naïve Bayes algorithm get a value of 84.38% and an AUC value of 0.953. The results of the PSO-based Naïve Bayes algorithm get a value of 88.62% and AUC value of 0.945. Based on the research which has been performed by using Naïve Bayes based on PSO, it developed a contribution rate of 5.02% in predicting the defected products.

Keywords: Defect Product, Data Mining, Naive Bayes, PSO

1. Introduction

In the era of progressively more competitive industrial competition, especially in the manufacturing world, it is always required to develop the quality or the quality of products and productivity. Each company is competing to win market share. One of the strategies carried out by the company is to improve the quality or the quality of products and the production process carried out by the company. In the industrial world quality or product quality and productivity are the keys for success of the production process.

In manufacturing companies, production activities are very important. If the production activities produce many defective products, it will cause losses in the form of additional time, costs, and raw materials for repairing the defective products. That is why quality control is needed so that the company can produce products according to predetermined quality standards. Basically, the quality of a product illustrates the extent of the ability of a product to display its ability to carry out the functions it has. The advantages of a product are measured through the level of customer satisfaction, so it is also necessary to analyze data and computer-based information. Data and information are needed by a large, medium and small scale company.

The object of this research is PT. Mane Indonesia. This French manufacturing company engaged in the manufacturing of aromatics (flavor and fragrance) for food flavorings and fragrances is operating in the MM2100 Cibitung area. The company produces tens to hundreds of products every day, assisted by
the Quality Department whose job is to inspect the product before sending it to the customer. However, not all products can fit product standards, that is why, inspectors cannot make quick decisions in the product checking process. It causes many defective products are wasted and inspectors have difficulty in finding defective product limits or making inspection reports every day. Besides, the time required in the inspection process is quite long and it can slow down the production process. Of course, this can be a loss for the company. This makes it difficult for the companies to check their production every day. Based on these problems, this study intends to provide solutions so that these problems can be minimized. From these reasons, a classification of the defect products category is needed, which is expected to become an early warning system that is able to overcome problems that have been a problem and can facilitate checking of defective products so that this is able to provide solutions to problems faced by the company. Classification is one of the roles of data mining techniques that can produce smart, fast, and accurate decisions. Classification can be interpreted as how to study a set of data to produce rules that can classify or recognize new data that has never been studied. There are several algorithms that can be used to handle data classification problems including Decision Tree, KNN (K-Nearest Neighbor), ANN (Artificial Neural Network), Naive Bayes, SVM (Support Vector Machine) [1]. One classification algorithm that is oftenly used is Naive Bayes. This algorithm was first introduced by Thomas Bayes since 1702. Naive Bayes has several advantages, namely fast calculation, simple algorithm and high accuracy [2]. This algorithm has been used to analyze automotive market sentiment on Twitter social media where the accuracy is 93%, with positive sentiment at 90%, negative sentiment at 90% [3]. Naive Bayes Classifier has also been used for sentiment analysis of East Java governor candidates who produce a classification of positive sentiments, neutral sentiments and negative sentiments in the East Java Governor Election in 2018 [4]. Then Naive Bayes Classifier has also been implemented into a Decision Support System for the determination in the selection of web-based and mobile hijab models [5].

However, the performance of Naive Bayes algorithm is still lacking compared to C4.5. In C4.5 all attributes are selected and divided into smaller subsets, but if data sizes are large with many attributes, then the model formed becomes complicated and difficult to understand so that data trimming is needed [6]. Meanwhile, Naive Bayes is more appropriate for large data. It can overcome incomplete data (missing value) and is strong against irrelevant attributes and noise in data. But then, Naive Bayes also has a weakness that is the probability of not being able to measure the level of accuracy of a prediction besides it is weak in the selection of attributes so that it can disturb the value of accuracy. Therefore Naive Bayes needs to be optimized by giving weight to attributes so that it can be used more effectively. To overcome this problem, it is necessary to optimize the Naive Bayes algorithm. One algorithm that can optimize the Naive Bayes is PSO. The Particle Swarm Optimization (PSO) algorithm can be used to weight attributes to improve the performance of Naive Bayes. The data used in this study is the data of defects in food and fragrance products from PT. Mane Indonesia as many as 800 data records divided into 6 attributes namely Product Name, Product Weight, pH Analysis, Visual Texture Inspection, Specific Gravity and Organoleptic Sensory Analysis then has 2 classes namely OK and NOK.

There are several studies related to this research, in 2017 Muhamad et al. has conducted research using the Iris dataset to optimize Naive Bayes using PSO. Testing is done in 2 ways, namely testing the number of particles and testing a combination of parameters. In the parameter combination test, the value of the combination was generated randomly. 3 times experiment produced the highest average fitness of 97.39. Meanwhile, in testing 10 to 50 number of particles with 3 times experiment, it was obtained the highest average fitness which is the same as the previous test that is 97.39 [7].

In 2018, Jaenal et al. has conducted an analysis of the effectiveness of Particle Swarm Optimization on improving the performance of the Naive Bayes Algorithm of which experiments are carried out using 7 different datasets which are divided into 2 tests. Naive Bayes and NB-PSO. NB-PSO uses 2 parameters of inertia weights and population size with a total of 6 repetitions. From the 7 experiments used, 3 of them were able to improve the accuracy of values by an average of 0.33% [8].

PSO algorithm has also been used to optimize the C4.5 algorithm in the case of determining customer satisfaction with the Tax Service Office [9]. However, Rifai and Aulianita have proven that the PSO-based Naive Bayes algorithm has higher performance. It is done by comparing the C4.5 algorithm and
Naive Bayes based on Particle Swarm Optimization for Credit Risk Determination. The results of this comparison produce that the highest recall value obtained by the PSO-based Naive Bayes Algorithm which is 96.75% [10]. Then in 2019 a comparison of NB (Naive Bayes) and NB-PSO was made to determine the level of vehicle sales that could increase the classification accuracy of the Naive Bayes algorithm by 0.33% [11]. The purpose of this study is to analyze data mining for defective products at PT. Mane Indonesia by using one of the roles of Data Mining which is Classification using the PSO-Based Naïve Bayes algorithm and to analyze the accuracy of the Naive Bayes algorithm for defective products.

2. Research Method

As depicted on the figure 1, the first step is collecting dataset as many as 800 record data, then preprocessing which contains cleaning, replacing, reduction, and transformation. The obtained results of pre-processing are dataset which are ready to be modeled. This research compared Naïve Bayes classifier (NB) to NB-PSO optimization. After that, cross validation with k=10 folds is conducted by dividing dataset into 10 to 90 of which 10% is used as data testing and 90% is used as data training. Validation is undergone to test algorithm method used. K-folds cross validation is a method used to find out the level of success of an algorithm model by re-testing input attributes randomly. The following figure is steps of research method used in this research.

![Fig. 1. Steps of Research](image)

2.1. Dataset Collection

The dataset used in this study is data on defective and non-defective products from the Quality Department section of PT Mane Indonesia totaling which are 800 data records in total. The attributes in the dataset are 6 (six) attributes including Product Name, Product Weight, pH Analysis, Visual Texture Inspection, Specific Gravity, Organoleptic Sensory Analysis. Then the two classes or targets they have are OK and NOK (Not OK) on the product's condition. The collected data will be divided into two parts, namely training data and testing data which will be used to test the accuracy of the system in classifying defective products.
2.2. Data Preprocessing
Not all of defective product data collected from the Quality Department are completely needed. If there is incomplete data filling, then, it needs to be completed or even not be used. At this stage, the selection of data is considered useful for data mining.

2.3. Naive Bayes
Naive Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. Naive Bayes' performance is competitive in the classification process even though it assumes attribute independence (there is no link between attributes). The assumption of the independence of this attribute in the actual data is rare, but although the assumption of the independence of the attribute is violated, the performance of the Naive Bayes classification is quite high. This is evidenced in various empirical studies [12].

Bayes is a simple probability-based prediction technique based on the application of the Bayes theorem with strong (naive) independence assumptions. In other words, in Naive Bayes, the model used is an "independent feature model". Bayes' prediction is based on the Bayes Theorem with the following general formula:

\[
P(H|X) = \frac{P(H)P(X|H)}{P(X)} = \frac{P(X|H)P(H)}{P(X)}
\]

Keterangan:
X : Sample data that has an unknown class (label)
H : Hypothesis that X is class data (label)
P(H|X) : Hypothesis probability H based on condition
X P(H) : Probability of Hypothesis H
P(X|H) : Probability of sample X data based on the conditions of Hypothesis H
P(X) : Probability of X

Naive Bayesian Classifier can also be defined as a classification method based on probability theory and Bayesian theorem with the assumption that each variable or decision parameter is independent (independence), so that the existence of each variable has nothing to do with the existence of other attributes.

2.4. PSO (Particle Swarm Optimization)
Particle Swarm Optimization (PSO) is a global optimization method introduced by Kennedy and Eberhart in 1995 based on research on the behavior of flocks of birds and fish. Each particle in the Particle Swarm Optimization has the speed of the particles moving in the search space with a speed that is dynamically adjusted to their historical behavior. Therefore, particles have a tendency to move towards better search areas during the search process [7]. The advantages of Particle Swarm Optimization method are that it is easy to implement, efficient in calculations and has a simple concept when compared to mathematical algorithms and other heuristic optimization techniques [13].

3. Result and Discussion
3.1. Dataset
This stage determines the data to be analyzed using the Naïve Bayes method, so, the first step is to collect the training data of 800 data records. The training data used can be seen in the following table 1.

| Product Name | Weight | pH Checking | Visual Texture Inspection | Specific Gravity | Sensory Analysis/Organoleptic | Result |
|--------------|--------|-------------|---------------------------|------------------|-----------------------------|--------|
| Product 1    | <3 Kg  | No          | Muddy                     | >1.472 SG        | Normal                      | OK     |
| Product 2    | >25 Kg | No          | Muddy                     | >1.472 SG        | Weak                        | NOK    |
| Product 3    | >25 Kg | Yes         | Muddy                     | >1.472 SG        | Normal                      | NOK    |
The description of the attribute categories in table 1 is as follows:

a) The weight of the product for the above attributes uses the estimated amount of weight of each product with three divisions which are: less than 3kg (the real data is between 100gr to 3kg), 3kg to 25kg, and more than 25 kg.

b) pH Checking: the checks which is conducted according to customer demand to determine the acidity of a product.

c) Visual Texture: the checks of product visually whether the product is turbid, bubbly and sedimentary.

d) Specific Gravity: the mass checking of a product with national standards.

e) Sensory analysis / Organoleptic: the scent checking manually with the sense of smell whether the product has a scent that matches the specifications.

3.2. Comparison of the Results

The study was divided into twice training and twice testing. The first training used 800 data records with the Naïve Bayes algorithm, while the second training used 800 data records with the Naïve Bayes and PSO method. The first testing used 372 data records with the Naïve Bayes algorithm, while the second testing used 372 data records with the Naïve Bayes and PSO method. The results of these studies can be seen in Table 2 as follows:

Table 2. The Comparison Results of Study

| Measurement | Training | Testing |
|-------------|----------|---------|
|              | Naive Bayes | NB+PSO  | Naive Bayes | NB+PSO  |
| Accuracy    | 84.38%    | 88.62%  | 90.32%    | 92.22%  |
| Precision   | 83.76%    | 84.86%  | 88.00%    | 95.09%  |
| Recall      | 86.25%    | 94.57%  | 88.67%    | 85.33%  |
| AUC         | 0.953     | 0.945   | 0.973     | 0.945   |

Based on the obtained results, there are increases in the value of accuracy, recall, precision, and AUC. The results of Naïve Bayes algorithm using the PSO method are improving compared to algorithm which only uses one learning technique. The following are graphic pictures showing an increase in the results of the naïve bayes algorithm and the naïve bayes algorithm PSO method. Research on defective products using the PT Mane Indonesia dataset using the Naïve Bayes algorithm PSO method has increased accuracy in 800 data that is equal to 5.02% and in 372 data amounted to 2.10%. It shows that the Particle Swarm Optimization (PSO) method can result higher accuracy improvement compared to the calculation of the naïve Bayes algorithm.
4. Conclusions

The results of training data testing conducted by increasing accuracy using the RapidMiner tools produce an accuracy rate of 88.62% (800 data) and 92.22% (372 data), while the results of the evaluation using the ROC curve with an excellent classification accuracy of 0.953 (800 data) and 0.945 (372 data). And also increase accuracy by 5.02% (800 data) and 2.10% (372 data). This proves that the more data tested, the results obtained will be better.

References

[1] F. Santoso, A. Syukur and A. Z. Fanani, "Algoritma C4.5 dengan Particle Swarm Optimization untuk Klasifikasi Lama Menghafal Alquran pada Santri Ma'hadul Quran," Cyberku Jurnal, vol. 14, no. 2, pp. 92-103, 2018.
[2] A. Hamzah, "Klasifikasi Teks dengan Naive Bayes Classifier untuk Pengelompokan Teks Berita dan Abstrak Akademik," in Seminar Nasional Aplikasi Sains dan Teknologi (SNAST) III, Yogyakarta, 2012.
[3] D. Rustiana and N. Rahayu, "Analisis Sentimen Pasar Otomotif Mobil: Teewt Twitter Mengunakan Naive Bayes," Jurnal SIMETRIS, vol. 8, no. 2, pp. 603-610, 2017.
[4] A. Ghulam, "Analisis Sentimen Calon Gubernur Jawa Timur 2018 dengan Metode Naive Bayes Classifier," Jurnal JIPN, vol. 4, no. 1, pp. 27-32, 2019.
[5] N. Rahmawati, A. Nugroho and I. Romli, "Penerapan Metode Naive Bayes dalam Sistem Pendukung Keputusan untuk Menentukan Model Hijab," in Seminar Nasional Sekolah Pascasarjana UGM, Yogyakarta, 2018.
[6] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining, Boca Raton: CRC Press, 2009.
[7] H. Muhamad, C. Prasojo, N. Sugianto, L. Surtiningsih and I. Cholissodin, "Optimasi Naive Bayes Classifier dengan Menggunakan Particle Swarm Optimization pada Data Iris," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 4, no. 3, pp. 180-184, 2017.
[8] M. Jaenal, A. Nugroho and I. Romli, "Analysis of Effectiveness Particle Swarm Optimization in Improving The Performance of Naive Bayes Algorithm," in 1st ICID, Yogyakarta, 2018.
[9] I. Romli, F. Kharida and C. Naya, "Penentuan Kepuasan Pelanggan Terhadap Pelayanan Kantor Pelayanan Pajak Menggunakan C4.5 dan PSO," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 2, pp. 296-302, 2020.
[10] A. Rifai and R. Aulianita, "Komparasi Algoritma Klasifikasi C4.5 dan Naive Bayes Berbasis Particle Swarm Optimization untuk Penentuan Resiko Kredit," Journal Speed - Sentra Penelitian Engineering dan Edukasi, vol. 10, no. 2, pp. 49-55, 2018.
[11] I. Romli, E. Pusnawati and M. Jaenal, "Comparison of NB and NB-PSO to Determine Level of Vehicles Sales," in ICOSSHEET 2019, Semarang, 2019.
[12] A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," Jurnal Ilmiah Ilmu Komputer, vol. 2, pp. 1-5, 2016.
[13] T. Arifin, "Implementasi Algoritma PSO dan Teknik Bagging Untuk Klasifikasi Sel Pap Smear," Jurnal Inform, vol. 4, no. 2, pp. 155-162, 2017.