ANALYSIS OF EFFECTIVENESS PARTICLE SWARM OPTIMIZATION IN IMPROVING THE PERFORMANCE OF NAÏVE BAYES ALGORITHM

Muhammad Jaenal[1)]
Departement of Informatics Engineering
STT Pelita Bangsa
Cikarang Bekasi, Indonesia
muhammadjaenal@mhs.pelitabangsa.ac.id

Agung Nugroho[2)]
Departement of Informatics Engineering
STT Pelita Bangsa
Cikarang Bekasi, Indonesia
agung@pelitabangsa.ac.id

Ikhsan Romli[3)]
Departement of Informatics Engineering
STT Pelita Bangsa
Cikarang Bekasi, Indonesia
ikhsan.romli@pelitabangsa.ac.id

Abstract*—Technological development makes the need to exchange accurate information based on data stored in a database is needed. Data mining itself is one of the techniques to find information and hidden knowledge from a data. One algorithm that is quite widely used in the data mining classification process is naïve bayes, chosen because it has a relatively short time, simple and has high accuracy. Naïve Bayes algorithm has a weakness which is the probability cannot measure the accuracy of a prediction. Therefore, a particle swarm optimization (PSO) method is needed. This study applies PSO to improve the performance of the naïve bayes algorithm where the PSO measurement parameter lies in its performance effectiveness against naïve bayes. The method used is experiment by using 7 cases of different datasets which are divided into 2 tests. Naïve Bayes and naïve Bayes tests are based on PSO (NB-PSO) wherein the NB-PSO test uses 2 parameters of inertia weight and population size with a total of 6 repetitions. The results of the experiments carried out showed that the effectiveness of the use of the optimization method. From 7 datasets used, 3 of them were able to increase the accuracy value with an average of 0.33% with the PSO success rate of all data reached only 42.68%.

Keywords : naïve bayes; particle swarm optimization; performance; effectiveness; swarm intelligence

I. INTRODUCTION

Data mining is one part of computer science that is used widely in the world of information and technology, especially today when big data has been used. Every second there are millions of data obtained so that the role of data mining is very important to find hidden information from a data.

To find a hidden information or knowledge from a data, a computational process is needed. Several techniques are used, including classification, statistics, clustering, prediction, estimation and association; according to the needs and characteristics of a case. those were done in order to determine the latest knowledge patterns of data.

There are several algorithms that can be used to handle computing with classification techniques, including C4.5 algorithm, KNN (K-Nearst Neighbor), and SVM (Support Vector Machine) but the algorithm that is often used in handling data classification cases is Naïve Bayes. [2] the naïve bayes algorithm has several advantages which are simple, relatively short time in calculation, and high in accuracy.

However, The Naive Bayes algorithm has a weakness which is the probability cannot measure the accuracy of a prediction. So That is why the Naive Bayes algorithm needs to be optimized using the method of giving weight. [2]

To overcome this problem, the particle swarm optimization (PSO) method is used as an optimization method. This method is able to improve the performance of the Naïve Bayes algorithm. PSO is an optimization method which is based on the behavior of a flock of birds. This method is often called as behaviourally inspired method, which is an alternative of genetic algorithm.

Several previous relevant studies have been carried out by [3]. This study proved that the optimization method managed to improve the accuracy of the naïve bayes algorithm with an accuracy of 92.86%, its precision value was 80% and the recall was 40% and the AUC value (area under cover) was 0839 results differed considerably before the Naïve Bayes algorithm

ISBN 978-602-53524-0-9

in optimization which only produced 82.14% accuracy value, precision value 23.91%, and recall value of 36% and AUC value of 0.686.

Based on research conducted by [3], it is proved that the optimization method influences the results of the performance of the naïve Bayes algorithm. However, this cannot represent that the PSO optimization method is very effective because based on research [4] there are several case studies where in certain data patterns the PSO has no effect on naïve Bayes.

Problem solving using naïve bayes algorithm based on PSO is not so significant in influencing naïve Bayes performance improvement. So it requires a study by applying the parameters of the same treatment in different datasets to measure the effectiveness of particle swarm optimization (PSO) in improving the performance of naïve Bayes algorithm performance.

To find out the exact measurement results, we need several case studies of data that are different from different sizes or characteristics. In this study, 7 cases used were different, including classification of mushroom, iris, breast cancer coimbra, chronic kidney disease, crypotherapy, credit cards client, and cardiotocography which all data was taken from the site of the UCI Repository.

Applying the same treatment to each dataset makes the data more measurable and easier to analyze and it is also easier to see how much effectiveness PSO is in improving the performance of the naïve Bayes algorithm in handling a particular data.

II. METHOD

A. *Working Principles of Naïve Bayes PSO*

Particle swarm optimization, abbreviated as PSO, is based on the behavior of a flock of insects, such as ants, termites, bees or birds. The PSO algorithm mimics the social behavior of this organism [2]. Social behavior consists of individual actions and the influence of other individuals in a group [2]. The word "particle" shows an individual.

Each individual or particle behaves mutually connected by using his own intelligence and is also influenced by the behavior of his collective group [2]. Thus, if one particle or a bird finds the right or short path to the food source, the rest of the group will also be able to immediately follow the path even though their location is far away in the group [2]

TABLE 1. HISTORY OF PSO

History of PSO	Scientist	Year
PSO Standard	James Kennedy and Russ Eberhart	1995
Modification Inertia Weight	Y. Shi and Russ Eberhart	1998
Modification Koefisien K	Maurice Clerc, Y.Shi and Eberhart	1999
PSO Complexity	Carlisle and Dozier	2001

a. Sample of a Table footnote. *(Table footnote)*

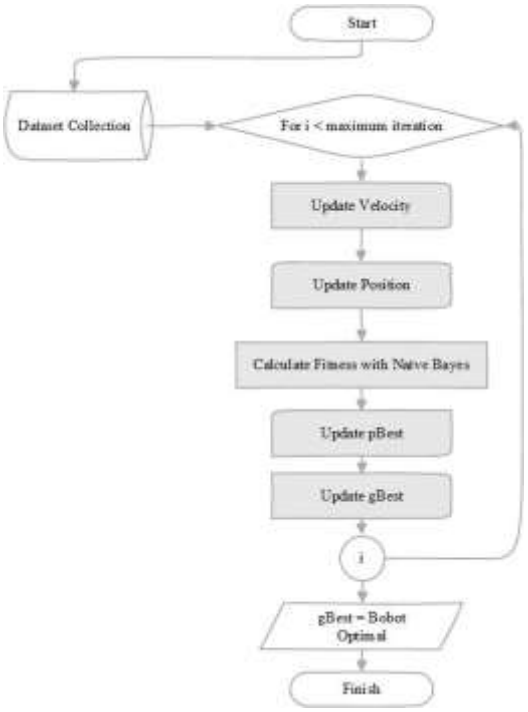

Figure 1. Working Principles of Naive Bayes PSO

B. *The Proposed method*

The method proposed in this study is illustrated in Figure 3.2 below:

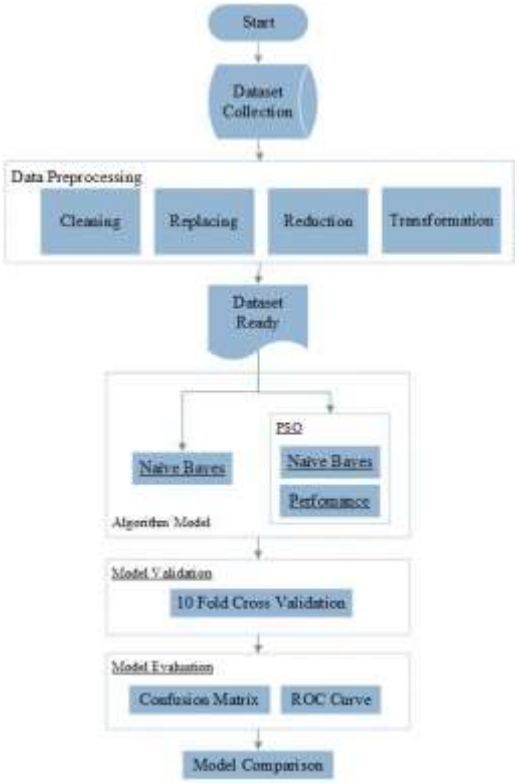

Figure 2. The Proposed Method

ISBN 978-602-53524-0-9

The dataset used in this study uses public data taken from the UCI Repository site as many as 7 different types of datasets

TABLE 2. INFORMATIONS DATASETS

No	Datasets Name	Attribute	Year
1	Cryotherapy	Integer/Real	2018
2	Iris Data	Real	1988
3	Mushroom Data	Categorical	1987
4	Breast Cancer Coimbra	Integer	2018
5	Credits Cards Client	Integer, Real	2016
6	Chronic Kidney Disease	Real	2015
7	Cardiotocography	Real	2010

a. Sample of a Table footnote. *(Table footnote)*

III. RESULT

The results of this study were carried out using the Rapidminer application with the number of datasets that have been determined and explained in advance, namely there are 7 different datasets with each attribute and different amounts of data. The tests conducted in this study were divided into two tests using the naïve Bayes method and the naïve bayes based on PSO.

TABLE 3. FEATURE SELECTION

Population Size	Inertia Weight	
	0.8	1.6
10	Result	Result
30	Result	Result
50	Result	Result

a. Sample of a Table footnote. *(Table footnote)*

A. *Experiment 1 (Naïve Bayes)*

The first test uses an algorithm that is naïve bayes without using optimization methods in classifying data as many as 7 different datasets. The following is the first test model used can be seen in Figure 4.1:

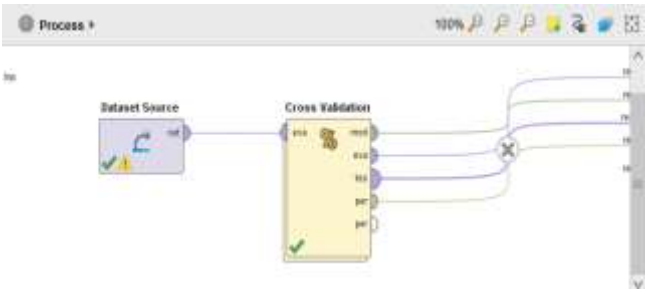

Figure 3. Model Experiment 1

Figure 4.1 describes a series of test models 1 globally where the source datasets will be adjusted according to the available datasets. Then to validate the model of the naïve bayes algorithm, cross validation method is used. Wherein there is a performance using confusion matrix and roc curve as an evaluation model of the naïve Bayes algorithm performance. For more details, see the following 4.2:

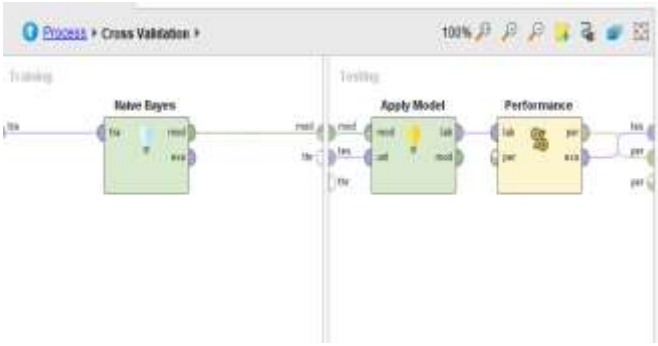

Figure 4. Cross Validation Models

The results of research conducted globally can be seen as a whole the average value of all measurements in table 4.1 below:

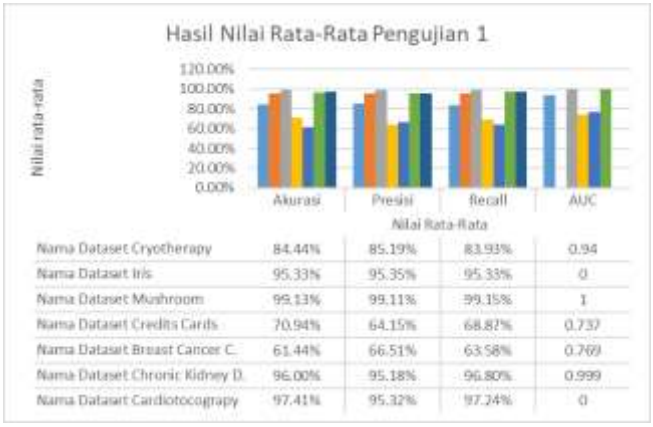

Figure 5. Result of Experiment 1

B. *Experiment 2 (Naïve Bayes PSO)*

This second test will use the naïve Bayes algorithm with an optimization method in classifying data. Following this, the second test model used can be seen in Figure 4.4:

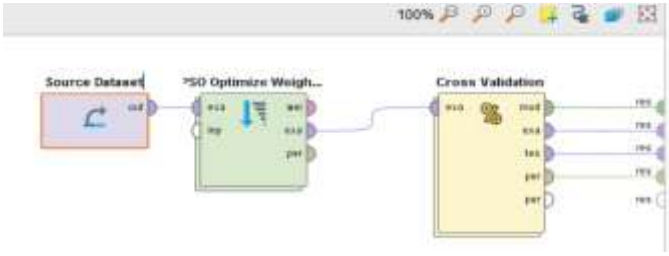

Figure 6. Model Experiment 2

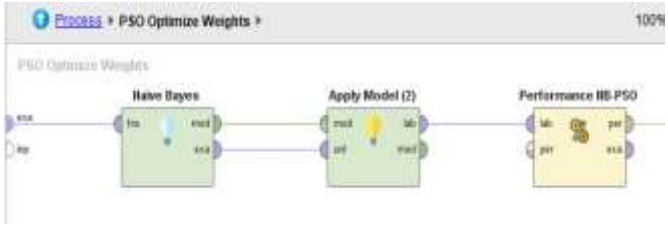

Figure 7. Model Experiment 2

ISBN 978-602-53524-0-9

Figure 4.4 illustrates a series of testing models 2 globally. Then to distinguish from the first test is the application of the PSO optimization method to give the following weighting

To make it easier to do research by applying parameters, each treatment or test is given a name that can be seen in the following table 4.2:

TABLE 4. INFORMATION OF PARAMETER PSO

Experiment	Population Size	Inertia Weight
Uji A	10	0.8
Uji B	10	1.6
Uji C	30	0.8
Uji D	30	1.6
Uji E	50	0.8
Uji F	50	1.6

a. Sample of a Table footnote. *(Table footnote)*

Based on the results that have been obtained as a whole in test 2, it can be seen the results of overall accuracy in table 4.3 below:

TABLE 5. RESULT EXPERIMENT 2

Nama Dataset	Algoritma yang Digunakan						
	Naïve Bayes	Naïve Bayes PSO					
		Uji A	Uji B	Uji C	Uji D	Uji E	Uji F
Cryotherapy	84,44%	85,56%	85,56%	85,56%	85,56%	85,56%	85,56%
Iris Data	95,33%	94,67%	94,67%	95,33%	95,33%	95,33%	95,33%
Mushroom Data	99,13%	99,11%	99,11%	99,13%	99,13%	99,13%	99,13%
Credits Cards Client	70,94%	71,14%	71,14%	71,14%	71,14%	70,72%	70,72%
Breast Cancer Coimbra	61,44%	62,20%	62,20%	63,79%	63,79%	63,03%	63,03%
Chronic Kidney	96 %	95 %	95 %	96 %	96 %	96,25%	96,25%
Cardiotocography	97,41%	97,41%	97,41%	97,32%	97,32%	97,32%	97,32%

Based on table 4.10 it can be concluded that out of 7 different data cases 4 of which the accuracy of the data was affected by PSO. The following is a visualization of the results of the accuracy of all tests.

The case of crypotherapy data can be seen that the PSO method is effective in improving the accuracy performance of naïve bayes by 1.12% with the parameters of Tests A through F getting the same results. To make it easier the following are the results of visualization through a diagram that can be seen in figure 4.7:

The results of the average PSO effectiveness on naïve bayes accuracy performance are as follows:

Figure 8. Result Performance PSO

TABLE 6. AVERANGE ACCURACY

Nama Dataset	Averange PSO
Cryotherapy	1.20%
Iris Data	-0.22%
Mushroom Data	-0.01%
Credits Cards Client	0.06%
Breast Cancer Coimbra	1.57%
Chronic Kidney Disease	-0.25%
Cardiotocography	-0.05%
Total Averange	**0.33%**

TABLE 7. AVERANGE OF MEASUREMENT

Nama Dataset	Avg. PSO (Accuracy)	Avg. PSO (Precision)	Avg. PSO (Recall)
Cryotherapy	1.20%	1.44%	1.04%
Iris Data	-0.22%	-0.23%	-0.22%
Mushroom Data	-0.01%	0.02%	0.00%
Credits Cards Client	0.06%	0.01%	-0.01%
Breast Cancer Coimbra	1.57%	1.75%	1.69%
Chronic Kidney Disease	-0.25%	-0.57%	-0.33%
Cardiotocography	-0.05%	-0.18%	0.13%
Averange	**0.33%**	**0.32%**	**0.33%**

Based on the results of the research produced, the Author's hypothesis was rejected, because it turns out that the hypothesis

ISBN 978-602-53524-0-9

made that is 5% in fact in this study only reached the highest average of 0.33%.

Then about the success rate of PSO influence on all data only reached 42.86% of the entire dataset used. In other words, from the 7 datasets used only 3 that are affected by the performance of the PSO optimization method.

This happens because not all datasets are significantly affected by PSO, other factors that influence the failure of the research hypothesis are due to the selection of datasets which are divided into two characteristics of datasets namely polynominal and binominal whereas in polynominal data cannot use ROC Curve measurements to determine the value of AUC

IV. CONCLUSION

Based on the results of the research that has been done, it can be concluded as follows:

- Other factors that influence the level of accuracy of the Naïve Bayes algorithm in addition to applying the PSO optimization method is at the preprocessing stage which at this stage it is necessary to focus on preparing a data to be clean of all existing data noise so that the classification process can run well and accurately.
- The thing that most influences the level of a naïve bayes algorithm performance is population size exactly as said by (charlisle and dozier, 2001) in the book (Suyanto, 2017) states that PSO with a smaller number of particles will work very fast, but often stuck at the local optimum while the particles in large quantities are rarely trapped at the local optimum but require a longer time.
- Based on research that has been carried out the effectiveness level of the use of optimization methods from 7 datasets used is 3 of them are able to increase the accuracy value with an average of 0.33% with the PSO success rate of all data only reaches 42.68%. That way it can be concluded that the method of running optimization is less effective in improving the performance of naïve bayes algorithms.

Based on the results of the research that has been done the author has several suggestions that can be used in the future research as follows:

- Using other measurement parameters in using the PSO method such as the use of K coefficient limits to limit and control the velocity movement.
- Separating data between polynominal and binominal because it greatly affects what will be observed later. This is done to provide more accurate and measurable research.
- Using other optimization methods such as genetic algorithms, or other swarm intelligence in an effort to find the best optimization methods that can improve the performance of the naïve Bayes algorithm to the fullest.
- Trying to use other tools like matlab or weka studio in carrying out data mining processes other than rapidminer.

ACKNOWLEDGMENT

His research was supported/partially supported by Pelita Bangsa Go Open Source. We thank our colleagues from STT Pelita Bangsa who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

We would also like to show our gratitude to the Dr.Ir. Supriyanto, M.M STT Pelita Bangsa for sharing their pearls of wisdom with us during the course of this research, and we thank 3 “anonymous” reviewers for their so-called insights.

REFERENCES

[1] H. Muhamad *et al.*, “Optimasi Naive Bayes Classifier dengan Menggunakan Particle Swarm Optimization pada Data Iris,” *Teknol. Inf. Dan Pendidik.*, vol. 4, no. 3, pp. 180–184, 2017.

[2] I. Cholissodin and E. Riyandani, *Swarm Intelligence (Teori & Case Study)*. Universitas Brawijaya, 2016.

[3] N. A. Widiastuti, S. Santosa, and C. Supriyanto, “Algoritma Klasifikasi Data Mining Naïve Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung,” *J. Pseudocode*, vol. 1, no. 1, pp. 2355–5920, 2014.

[4] Y. Li, Z. Chen, Y. Wang, L. Jiao, and Y. Xue, “A Novel Distributed Quantum-Behaved Particle Swarm Optimization,” *J. Optim.*, vol. 2017, pp. 1–9, 2017.

[5] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques,” *Elsevier, San Fr.*, vol. 3 edition, p. 629, 2011.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, pp. 37–54, 1996.

[7] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[8] Suyanto, *SWARM INTELLIGENCE KOMPUTASI MODERN UNTUK OPTIMASI DAN BIG DATA MINING*, Pertama. Bandung: Informatika Bandung, 2017.

AUTHORS PROFILE

Muhammad Jaenal was born in 1996 in Bekasi Jawa Barat, Indonesia. His Bachelor degree. Was Informatics Engineering (Software Engineer) from STT Pelita Bangsa in 2018. His research interest include data mining, python and web development. His active writing article and study literature.

Agung Nugroho His Bachelor degree, was Computer Science from university Ahmad Dahlan Yogyakarta in 2004. He is persuing the M.Kom. was in Informatics Engineering from STMIK Eresha 2016. He is Lecturer in STT Pelita Bangsa from 2012 – present.

Ikhsan Romli His Bachelor of Science (Mathematics) from Sepuluh November Institute of Technology. He is persuing the Master in Manufacturing Engineering (Manufacturing Management) from Universiti Teknikal Malaysia Melaka. He is Lecturer Secretary in STT Pelita Bangsa 2015 – Present.

ISBN 978-602-53524-0-9