Advances in Intelligent Systems Research, volume 172
Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)

Optimization Naive Bayes Algorithm Using Particle Swarm Optimization in the Classification of Breast Cancer

Vira MELINDA[1], Rifkie PRIMARTHA[1]*, Adi WIJAYA[2] and Muhammad Ihsan JAMBAK[1]

[1]*Sriwijiya University,Palembang,Indonesia*
[2]*MH Thamrin University,Jakarta,Indonesia*
**Corresponding author: rifkie77@gmail.com*

ABSTRACT
Methods of data mining classification are used in various fields of research. Naive Bayes is one of the most used algorithms of data mining classification, especially in the medical science because Naive Bayes is considered good method for the data concerned with a statistical diagnosis. Optimization of diagnosis results needs to be done in terms of various weaknesses, including data passing certain classes even though the data it is irrelevant or relevant so the need to be optimized by feature selection. Optimization was done using Particle Swarm Optimazation algorithm for feature selection in breast cancer classification using Naive Bayes. The Naive Bayes method is used for the classification of breast cancer, while the Particle Swarm Optimization Algorithm is used for the selection of irrelevant attribute features in order to obtain optimal diagnosis results. The results of the Naive Bayes method were 95.49% while after being optimized with the Particle Swarm Optimazation Algorithm the result was 98.19%.
Keywords: *classification data mining, Naive Bayes, Particle Swarm Optimization, breast cancer, feature selection*

1. Introduction

Classification is one of the data mining techniques commonly used to find models or patterns for a particular algorithm. In classification the object welding process occurs which is done by dividing objects based on groups that have been previously defined.Naïve Bayes is a classification data mining algorithm which introduced by Revered Thomas Bayes, this algorithm is usually used in various fields of research because this algorithm is good for data related to statistical diagnosis with a simple algorithm and Naïve Bayes can handle blank or missing value is be an advantage of this algorithm, but behind the strengths there is a weakness of the Naïve Bayes algorithm which has many gaps to reduce effectiveness because the Naïve Bayes algorithm has no weight in attributes so all attributes have the same value even though they are not relevant in classification process. [1]Feature selection is one of the important techniques that is often used in data preprocessing data mining to accelerate and optimization process of an algorithm. Particle Swarm Optimization (PSO) is one of the algorithms used in feature selection, Particle Swarm Optimization algorithm is also known to have a stable level of convergence with a simple and efficient concept in calculating the search algorithm with a tendency to move to a better search population after passing through the search process. The simplicity of the algorithm and its good performance have made Particle Swarm Optimization algorithm popular and become a global optimizer with most problems that can be solved properly where the variables are real numbers.[2]The thing to strengthen the purpose of this research is to optimize the Naive Bayes algorithm with Particle Swarm Optimization. There is a literature review from previous research conducted by (Bellaachia & Guven, 2006) comparing three algorithms for breast cancer prediction namely Naive Bayes, Artificial Neural Net and C4.5. Naïve Bayes classification techniques for breast cancer prediction in this study get an accuracy of 84.5% where this result is worse with other algorithms, such as Artificial Neural Net get an accuracy of 86.5% and C4.5 with the greatest accuracy of 86.7%. From the comparison of the three algorithms it can be seen that the Naïve Bayes algorithm has less performance when compared to other algorithms in this classification. As it is known that good accuracy performance is accuracy that is close to 100% so that the Naïve Bayes algorithm can be optimized by selecting features to cover the shortcomings of this algorithm because from the previous explanation it is explained that Naïve Bayes can pass data into a certain class of data which is clearly not feasible to enter in the class or not concerned.

Based on this, this study discusses the effect of using Particle Swarm Optimization in optimizing the Naïve Bayes algorithm in classification of breast cancer.

RELATED WORKS

Related research that used same dataset from UCI Machine Learning Repository were taken to compare the results that obtained from proposed method. In 2018[3] (Aslan, Celik, Sabanci, & Durdu, 2018) testing the efficacy of using machine learning techniques by comparing the ANN, ELM, k-NN and SVM methods to detect breast cancer with the aim of research is to process the results of routine blood analysis with different ML methods and to understand how effective these methods are for detection. From their study considering the input values, max and min of these values are quite different from each other. Normalization must first be applied to normalize the distribution and increase the success rate. Feature Scaling method is used for normalization. After normalization using training and test data were generated randomly from the data. 80% percent of the whole data were used in the test phase and 20% percent were used in the training phase. After separation of training and test data, results were obtained for each ML method. The results of their study showed that the highest level of accuracy and the lowest training period provided by ELM was 80%, then ANN was 79.43%, k-NN was 77.5% and SVM had the lowest accuracy which was only 73.5%. Other studies related to breast cancer research have been conducted by [4](Wiswandani, 2018) in this research using machine learning techniques for automatic diagnosis by knowing the factors that are thought to influence the cause of breast cancer. Classification analysis to classify patients as being healthy or diagnosed with breast cancer is done using machine learning methods, while the methods used are naive bayes, support vector machines with rbf kernel and linear kernel methods, kNN, random forest and decision tree with the following accuracy Naive Bayes 61.3%, SVM kernel rbf 69.7%, SVM linear kernel 53,% 4, Random Forest 64.6%, Decision Tree 46.6% in this study shows that the best method that is appropriate to be used in the Breast Cancer Coimbra data is the rbf kernel vector machine support method with an average accuracy value of 69.7%.

MATERIALS AND PROPOSED METHOD

Dataset used in this study is a secondary data type in the form of patient medical data and healthy control of female patients in Coimbra named Breast Cancer Coimbra Dataset 2018 that can be accessed on UCI Machining Learning Repository. Breast Cancer Coimbra has 9 attributes and 1 labels (infected and not infected) and has 110 data objects in excel.

To optimize Naïve Bayes algorithm by selecting features using the Particle Swarm Optimization algorithm, the research will be carried out in stages as described in the framework below:

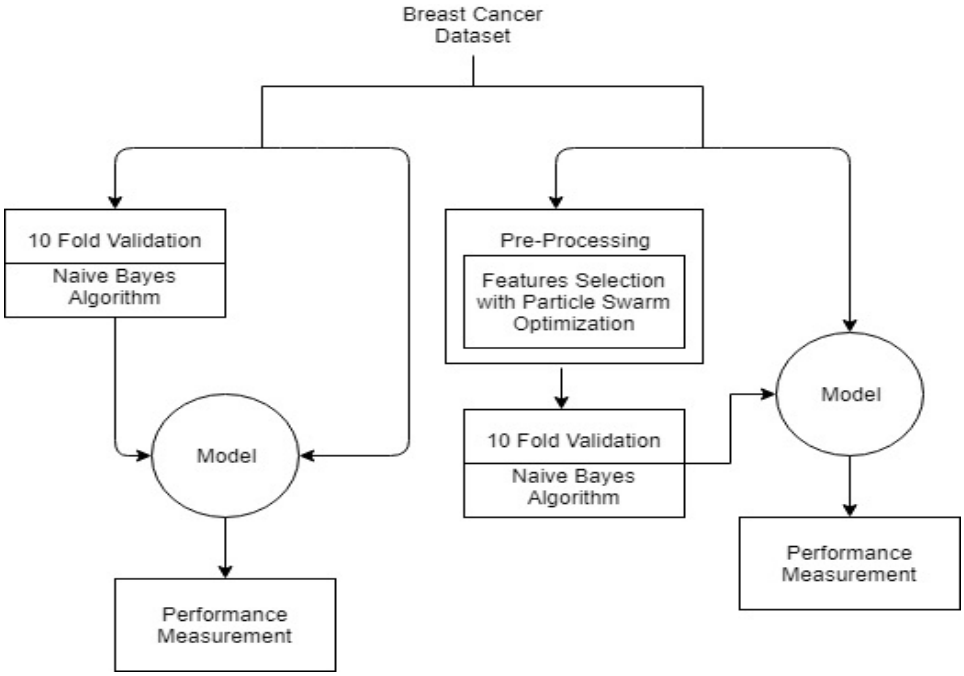

Figure 1. Flowchart of Proposed Method

In the initial testing phase of the research the data will be divided into two parts, namely training data and testing data using K-fold cross validation and with K = 10 which data is divided into 10 parts using 10 times validation where 9 parts of the data will be used as training data and 1 part of the data will be used as testing up to 10 times the test. Because in this study comparing classification with Naive Bayes Algorithm and Naive Bayes classification optimized with Particle Swarm Optimization algorithm then after the data is divided into testing data and training data the next stage for Naive Bayes classification optimized with Particle Swarm Optimization algorithm will be optimized using Particle Swarm Optimization algorithm which aims to process feature selection after obtaining the best features of the Particle Swarm Optimization algorithm data process will be processed using the Naïve Bayes algorithm to classify breast cancer while for the classification with the Naive Bayes

algorithm data that has been divided into testing data and training data will be classified using the Naive Bayes algorithm. In this research, process method will be evaluated using conflusion matrix.

EXPERIMENTAL RESULT

The study was conducted by comparing the classification of breast cancer with Algortima Naive Bayes and the classification of breast cancer with Naive Bayes after having selected features with PSO. In this study focused on comparing the accuracy of both point to see whether the classification results are better after optimization with feature selection using PSO.
The experiment was carried out using a dataset of 110 medical history records of the patient and was tried 30 times. The test was performed in 2 stages, test 1 was performed classification with Naive Bayes after optimization with Particle Swarm Optimization Algorithm, test 2 was performed classfication with Naive Bayes Algorithm.

Stage 1. Classification with Naive Bayes after optimization with Particle Swarm Optimization Algorithm

Table I. Feature Selection Results Table with Particle Swarm Optimization With the number of particles = 10 and iteration = 10

Trial no-	Partikel	Iterasion	Highest Accuracy	Selected features
1	10	10	0.9909	001001001
2	10	10	0.9819	100010111
3	10	10	0.9909	11110101
4	10	10	0.9819	110011000
5	10	10	0.9909	111101100
6	10	10	0.9909	11100010
7	10	10	0.9819	11110000
8	10	10	0.9819	000111001
9	10	10	0.9909	111111001
10	10	10	0.9909	001011000
11	10	10	0.9909	011011101
12	10	10	0.9909	0111011101
13	10	10	0.9909	1011011100
14	10	10	0.9909	11100010
15	10	10	0.9909	001001000
16	10	10	0.9819	100010011
17	10	10	0.9909	111111101
18	10	10	0.9909	111000011
19	10	10	0.9909	001111100
20	10	10	0.9909	1010010001
21	10	10	0.9909	100001000
22	10	10	0.9819	010010001
23	10	10	0.9909	111101001
24	10	10	0.9819	001010010
25	10	10	0.9909	011010000
26	10	10	0.9819	101010100
27	10	10	0.9819	010101000
28	10	10	0.9819	001010010
29	10	10	0.9909	111100000
30	10	10	0.9909	000011001
Highest Accuracy			0.9909	
Average Accuracy			0.9879	

Table II.Classification Results Table using Naïve Bayes which has passed the feature selection process using PSO With the number of particles = 10 and the number of iterations = 10

Trial no-	Partikel	iterasi	Selected features	Diagnosa									Accuracy	Classification
				F1	F2	F3	F4	F5	F6	F7	F8	F9		
1	10	10	001001001			98			1.234			876	0.9909	Infected
2	10	10	100010111	37				4.567		54	123	56	0.9819	Infected
3	10	10	11110101	20	3.56	34	1.456	89		11.45		346	0.9909	Infected
4	10	10	110011000	23	0.56			456	234				0.9819	Not Infected
5	10	10	111101100	46	54	79	17.8		5.9	54			0.9909	Not Infected
6	10	10	11100010	76	3	10					0.5		0.9909	Not Infected
7	10	10	11110000	28	0.98	80	17.54						0.9819	Not Infected
8	10	10	000111001				10	80	1.67			120	0.9819	Not Infected
9	10	10	111111001	34	12.56	2	64	10	11			900	0.9909	Not Infected
10	10	10	001011000			1		2	3				0.9909	Not Infected
11	10	10	011011101		30	2		77	15	1		567	0.9909	Not Infected
12	10	10	0111011101		60	0.986		1.097	4	9		764	0.9909	Not Infected
13	10	10	1011011100	87		0.86	10	98	17				0.9909	Not Infected
14	10	10	11100010	30	6.78	3.7					98		0.9909	Not Infected
15	10	10	001001000				0.89		1				0.9909	Not Infected
16	10	10	100010011	25				0.87			18	376	0.9819	Not Infected
17	10	10	111111101	31	0.98	12	1.98	89	65	17		765	0.9909	Infected
18	10	10	111000011	46	7.85	1.23					0.6	876	0.9909	Infected
19	10	10	001111100			98	38.9	8	34	17			0.9909	Infected
20	10	10	1010010001	52		70			7			543	0.9909	Not Infected
21	10	10	100001000	50						0.9			0.9909	Not Infected
22	10	10	010010001		29			2.87				229	0.9819	Infected
23	10	10	111101001	86	20	7.98	45		8			674	0.9909	Infected
24	10	10	001010010			67		3.976			17		0.9819	Infected
25	10	10	011010000		21	90		2.43					0.9909	Infected
26	10	10	101010100	65		60		619					0.9819	Infected
27	10	10	010101000		29.8		52		43				0.9819	Infected
28	10	10	001010010			98		3		29			0.9819	Infected
29	10	10	111100000	57	20.4	17	6						0.9909	Infected
30	10	10	000011001					0.1	5			987	0.9909	Not Infected

Stage 2. Classfication with Naive Bayes Algorithm

Table III. Table of Test Results for Naive Bayes Algoritm learning

Test Result for Naive Bayes				Status	
Data no-	Iterasion no-	Naïve Bayes prediction	Fact	Fail	Work
1	10	Not Infected	Not Infected		V
2	10	Not Infected	Not Infected		V
3	10	Not Infected	Not Infected		V
4	10	Not Infected	Not Infected		V
5	10	Not Infected	Not Infected		V
6	10	Not Infected	Not Infected		V
7	10	Not Infected	Not Infected		V
8	10	Not Infected	Not Infected		V
9	10	Not Infected	Not Infected		V
10	10	Infected	Not Infected	V	
11	10	Infected	Not Infected	V	
12	10	Not Infected	Not Infected		V
13	10	Not Infected	Not Infected		V
14	10	Infected	Infected		V
15	10	Not Infected	Infected	V	
16	10	Infected	Infected		V
17	10	Infected	Infected		V
18	10	Infected	Infected		V
19	10	Not Infected	Infected	V	
20	10	Not Infected	Infected	V	
21	10	Not Infected	Not Infected		V
22	10	Not Infected	Not Infected		V
23	10	Not Infected	Not Infected		V
24	10	Not Infected	Not Infected		V
25	10	Not Infected	Not Infected		V
26	10	Not Infected	Not Infected		V
27	10	Not Infected	Not Infected		V
28	10	Not Infected	Not Infected		V
29	10	Not Infected	Not Infected		V
30	10	Not Infected	Not Infected		V

Table IV. Classification Results Table using Naïve Bayes

Trial No-	Iterasion	Diagnosa									Acurracy	Clasification
		F1	F2	F3	F4	F5	F6	F7	F8	F9		
1	10	31	0.98	98	1.98	89	1.234	17	18	876	95.49%	Infected
2	10	37	0.98	12	1.98	4.567	65	54	123	56	95.49%	Infected
3	10	20	3.56	34	1.456	89	65	11.45	18	346	95.49%	Infected
4	10	23	0.56	12	1.98	456	234	17	18	765	95.49%	Not Infected
5	10	46	54	79	17.8	89	5.9	54	18	765	95.49%	Not Infected
6	10	76	3	10	1.98	89	65	17	0.5	765	95.49%	Not Infected
7	10	28	0.98	80	17.54	89	65	17	18	765	95.49%	Not Infected
8	10	31	0.98	12	10	80	1.67	17	18	120	95.49%	Infected
9	10	34	12.56	2	64	10	11	17	18	900	95.49%	Infected
10	10	31	0.98	1	1.98	2	3	17	18	765	95.49%	Infected
11	10	31	30	2	1.98	77	15	1	18	567	95.49%	Not Infected
12	10	31	60	0.986	1.98	1.097	4	9	18	764	95.49%	Not Infected
13	10	87	0.98	0.86	10	98	17	17	18	765	95.49%	Not Infected
14	10	30	6.78	3.7	1.98	89	65	17	98	765	95.49%	Not Infected
15	10	31	0.98	12	0.89	89	1	17	18	765	95.49%	Not Infected
16	10	25	0.98	12	1.98	0.87	65	17	18	376	95.49%	Not Infected
17	10	31	0.98	12	1.98	89	65	17	18	765	95.49%	Infected
18	10	46	7.85	1.23	1.98	89	65	17	0.6	876	95.49%	Infected
19	10	31	0.98	98	38.9	8	34	17	18	765	95.49%	Infected
20	10	52	0.98	70	1.98	89	7	17	18	543	95.49%	Not Infected
21	10	50	0.98	12	1.98	89	65	0.9	18	765	95.49%	Not Infected
22	10	31	29	12	1.98	2.87	65	17	18	229	95.49%	Infected
23	10	86	20	7.98	45	89	8	17	18	674	95.49%	Infected
24	10	31	0.98	67	1.98	3.976	65	17	17	765	95.49%	Infected
25	10	31	21	90	1.98	2.43	65	17	18	765	95.49%	Infected
26	10	65	0.98	60	1.98	619	65	17	18	765	95.49%	Infected
27	10	31	29.8	12	52	89	43	17	18	765	95.49%	Infected
28	10	31	0.98	98	1.98	3	65	29	18	765	95.49%	Infected
29	10	57	20.4	17	6	89	65	17	18	765	95.49%	Infected
30	10	31	0.98	12	1.98	0.1	5	17	18	987	95.49%	Not Infected

Based on tables II and IV, the results of Classification optimized with feature selection give classification results that are not too far from the Naive Bayes classification, although there are several different results due to the difference in diagnostic value between the Naive Bayes classification and the Naive Bayes classification that features have been selected with PSO.

Table V. Comparison table between Naive Bayes classification with Naive Bayes classification which has been optimized by particle swarm optimization

Trial No.	Particle	iterasi	*Naive Bayes*	*Naive Bayes + PSO*
			Accuracy	Accuracy
1	10	10	95.49%	99.09%
2	10	10	95.49%	98.19%
3	10	10	95.49%	99.09%
4	10	10	95.49%	98.19%
5	10	10	95.49%	99.09%
6	10	10	95.49%	99.09%
7	10	10	95.49%	98.19%
8	10	10	95.49%	98.19%
9	10	10	95.49%	99.09%
10	10	10	95.49%	99.09%
11	10	10	95.49%	99.09%
12	10	10	95.49%	99.09%
13	10	10	95.49%	99.09%
14	10	10	95.49%	99.09%
15	10	10	95.49%	99.09%
16	10	10	95.49%	98.19%
17	10	10	95.49%	99.09%
18	10	10	95.49%	99.09%
19	10	10	95.49%	99.09%
20	10	10	95.49%	99.09%
21	10	10	95.49%	99.09%
22	10	10	95.49%	98.19%
23	10	10	95.49%	99.09%
24	10	10	95.49%	98.19%
25	10	10	95.49%	99.09%
26	10	10	95.49%	98.19%
27	10	10	95.49%	98.19%
28	10	10	95.49%	98.19%
29	10	10	95.49%	99.09%
30	10	10	95.49%	99.09%
Average			95.49%	98.79%

Based on table V, for 30 times the average testing accuracy on the Classification using Naive Bayes is 95.49% while for the classification using Naive Bayes which is optimized by PSO feature selection - an average accuracy of 98.79%, so it is known that by optimizing the Naive Bayes Classification Bayes with the PSO feature selection increased accuracy by 3.3% which approached with perfect accuracy leading to 100%.

SUMMARY

Based on the experimental results explained in the previous chapter, we concluded that:
1. Classification using the Naïve Bayes method which is optimized by feature selection using the Particle Swarm Optimizitation algorithm can be applied. Proven by testing in this study shows that the accuracy of the classification using Naïve Bayes that is optimized by feature selection using Particle Swarm Optimizitation increases compared to only using Naïve Bayes classification. By optimizing classification using Naive Bayes with PSO feature selection, there was an increase in accuracy by 3.3%, from 95.49% to 98.79% which approached with perfect accuracy leading to 100%, for the dataset in this study.
2. Based on this research Naïve Bayes algorithm is an algorithm for classification with the first stage of the data set will be divided into training and testing data, then training data will be read to count the number of classes, the same number of cases with the same class and then all the calculation results will be multiplied according to the data x that the class is looking for.
3. The mechanism of the Particle Swarm Optimization algorithm in this research is as an algorithm for feature selection, where irrelevant features will be selected by giving values to the features in this study randomly using Boolean. In this algorithm there are parameters as follows: the number of particles as the location of each movement, it is the repetition of calculations for each particle, the values of C1 and C2 as a parameter of confidence in the weight of the particle (memory of the previous path) and the value of w as the weight of inertia to control the momentum of the particle by weighing the contribution from the previous speed. From this it can be seen that the way the PSO works is like a flock of birds where the best position will be chosen.

REFERENCES

[1] Akinsola Adeniyi, F., Sokunbi, M., & Okikiola, F. (2017). Data Mining For Breast Cancer Classification. *International Journal Of Engineering And Computer Science, 6*(8).

[2] Muzakkir, I., Syukur, A., & Dewi, I. N. (2014). Peningkatan Akurasi Algoritma Backpropagation dengan Seleksi Fitur Particle Swarm Optimization dalam Prediksi Pelanggan Telekomunikasi yang Hilang. *Jurnal Pseudocode, 1*(1), 1-10.

[3] Aslan, M. F., Celik, Y., Sabanci, K., & Durdu, A. (2018). Breast cancer diagnosis by different machine learning methods using blood analysis data. *International Journal of Intelligent Systems and Applications in Engineering, 6*(4), 289-293.

[4] Wiswandani, A. (2018). Analisis Klasifikasi pada Data Breast Cancer Coimbra Menggunakan Metode Machine Learning.

[5] Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age, 58*(13), 10-110.