http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.1, Desember 2021 ISSN : 2302-9706

Implementation Particle Swarm Optimization to improve the performance of Naive Bayes on Diabetes Detection Data

Handini Arga Damar Rani
Universitas IVET
Email: hani.arga@gmail.com

Abstract

Article Info
Received : 28 November 2021
Revised : 10 December 2021
Accepted : 13 December 2021

Diabetes has certainly been widely known throughout the world as a serious disease. According to WHO data in 2014 there were about 422 million adults who had diabetes. This is very interesting because the increase is very visible, almost more than doubled when compared to in 1980 people with diabetes were only around 108 million people. The more sophisticated technological developments nowadays, various types of diseases can be detected computerized using the data mining method. In this study, the researcher proposes the particle swarm optimization method to improve the quality of the data to be used in the detection of diabetes using the nave Bayes method. The resulting model was tested to obtain the accuracy and AUC (Area Under Curve) of each algorithm so that it was found that testing using nave Bayes got an accuracy value of 96.15% with an AUC value of 0.991. Meanwhile, testing using the Naïve Bayes method based on attribute selection using the Particle Swarm Optimization (PSO) method, obtained an accuracy value of 97.13% with an AUC value of 0.995.

Keywords: data mining, Diabetes, Particle Swarm Optimization, Naïve Bayes

1. Introduction

Diabetes is of course widely known and is said to be the silent killer [1]. According to WHO data in 2014 there were about 422 million adults who had diabetes. This is very interesting because the increase is very visible, almost more than doubled when compared to the 1980 population of only 108 million people with diabetes [2]. In Indonesia alone, the increase in the number of people with diabetes from 1.1% in 2007 increased to 2.1% in 2013 [3]. Therefore, diabetes needs to be detected or predicted accurately because it is used to prevent severe complications [4]. And can help medical experts to be faster in providing early analysis of the disease through monitoring data or data mining and can reduce the burden of existing financial costs so far.

Such as research conducted by Dewan Farid, et al [9], entitled Hybrid decision tree and nave Bayes classifiers for multi-class classification tasks. In this study, one of them reviews the nave Bayes problem which involves extreme and expensive computational processes in determining class conditional independence. This is due to the inability of nave Bayes in finding important parts in a subset of attributes. To overcome this problem, the research proposed a combination of the nave Bayes classification method combined with the decision tree method in the pre-processing phase of the data at the data reduction stage. Broadly speaking, the decision tree method is induced and applied to find and select the important subset of the attribute set. The path taken for the selection is by calculating the weight of the most dominant attribute in the dataset that will be used. The results obtained are the decision tree method is able to improve the performance of nave Bayes. With the comparison of the accuracy of nave Bayes before being combined

with decision tree induction as much as 76.30% then after being induced by the decision tree method it increased to 79.55%. While the data used using data from UCI, namely Pima Indians Diabetes (diabetes) which has 8 attributes with real attribute type, then 768 records and has 2 classes.

A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records was written by Tao Zheng, et al [12]. This study aims to find various genotypes of diabetes based on data taken from Electronic Health Records (EHR). The use of machine learning methods will be applied to this research. However, based on the identification, machine learning algorithms often experience low recall rates and can suffer the loss of a large number of valuable samples because they still use conservative standard filters. To overcome this problem, it is necessary to develop a framework based on machine learning methods that liberalize filter criteria to increase recall rate by keeping false positive values at a low or below normal position. In this study, the development of a framework to overcome this problem used feature selection techniques and machine learning algorithms in which there are several methods, one of which is the nave Bayes method. This framework is also used to identify which patients have diabetes and which patients do not. the amount of data taken was 300 patients with details 161 had diabetes, 60 were not and 79 were not detected. This sample data was taken randomly from a total of 23281 data originating from the Electronic Health Records (EHR) repository between 2012 and 2014. The results of this study indicate that the framework built achieves higher performance identification seen from the average value - average AUC of ~ 0.98. This is compared with the average AUC value in previous studies showing a value of 0.71 only.

Comparison of Classifiers for the Risk of Diabetes Prediction written by Nongyao Nai-arun, Rungruttikarn Moungmai [13]. This study reviews the application of algorithms to deal with classification problems on the risk of diabetes mellitus. One of the popular classification algorithms used is the Naive Bayes method. However, the performance of the Naive Bayes method needs to be improved so that it can produce better accuracy. this is very necessary because in this study the method will be applied and implemented in a website-based application. With the aim of predicting the risk of diabetes early without the need for blood tests or going to the hospital. The method to improve the performance of the naive bayes algorithm in this research is using the boosting algorithm. The procedure for this boosting algorithm begins by applying weighting for observations to all training data sets. To test the performance of the naive bayes method which is enhanced by the boosting method, data from 26 Primary Care Units (PCU) Sawanpracharak Regional Hospital was used during 2012 - 2013. The data has 11 attributes, namely BMI, age, weight, height, systolic blood pressure,diastolic blood pressure, history of diabetes in family, history of hypertension in family, alcohol drinking, smoking behaviour, smoking behavior and sex. Meanwhile, the total records were 30122, of which 19145 were normal patients and 10977 were at risk for diabetes. From this study, the results of the increase from the nave Bayes method after optimization using the bosting method. The results of this increase can be seen when using the single naive bayes method, the accuracy is 81.010% with an AUC of 0.855. while the results of the Naive Bayes optimization combined with the boosting method found an accuracy of 81,019% with an AUC of 0.864.

A Pearson's correlation coefficient based decision tree and its parallel implementation was written by Yashuang Mu, Xiaodong Liu and Lidong Wang [11]. This study discusses the decision tree method which has advantages, namely high accuracy with low parameters, and is very reliable in the field of classification through basic feature extraction of training data. Although reliable, decision trees have difficulty when faced with complex and large data. This is what makes the problem of the decision tree itself. because this can jeopardize its performance when applied in cases that have complex attributes and result in long computational times. To overcome this in this study, the Pearson correlation approach method is used which typically explores information between one attribute and another and also between attributes and labels. The approach in this research was tested using data from UCI with a total of 17 datasets. From this study, the results obtained that the Pearson correlation method can choose optimally and break down

attributes. So that the selected attribute will be used in the rules on the decision tree method. From these results we can conclude that the use of the attribute selection method approach can overcome the problem with the decision method.

As research conducted by Toni A.[14] Optimization of the Classification Method using Particle Swarm Optimization for Diabetes Retinopathy Identification. From the research conducted, it can be concluded that the Neural Network (NN) method is the best method in producing high accuracy when combined with the PSO feature selection method.

The Naive Bayes algorithm is often used to solve probability-based classification problems [5]. In addition, Naïve Bayes was chosen because of its simplicity which affects the effectiveness of computation time and good accuracy [6]. However, the nave Bayes algorithm has the assumption that all existing attributes have no relationship between one attribute and another. So that each attribute is considered to have an equally important value. This is what can cause a dramatic decrease in the performance of the nave Bayes algorithm [7]. To overcome this problem, the Pearson correlation method is used which is one of the typical models in the attribute sorting process. Because in the task Pearson correlation is used to measure information between one attribute and another and one attribute with its label.

The data preparation process is an important technique in improving data quality. Because with good data quality will be able to increase the accuracy and efficiency of the algorithm's performance [8]. Thus the constraint on the nave Bayes algorithm model lies in the quality of the data [9]. So this study uses the Particle Swarm Optimization method to improve the quality of the data to be used in the detection of diabetes using the nave Bayes method.

2. Method

The data in this study will use secondary data, namely Diabetes 130-US hospitals data obtained from the database in UCI [10]. For now, the problem that must be solved is the determination of the occurrence of diabetes mellitus by utilizing the Diabetes 130-US hospitals data, totaling 101767 data lines. There are 50 attributes in the pima data, namely encounter_id, patient_nbr, race, gender, age, weight, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, payer_code, medical_specialty, num_lab_procedures, num_procedures, num_medications, number_ingency, patient_2, number_1, patient_ingency, number_1, -diag_3, number_diagnoses, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, insulin, citoglytazone, tolbutamide metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, readmitted and diabetesMed as labels consisting of yes and no.

The total data that has been obtained is 101767. The data that has been obtained is not directly used but must go through a selection process first because not all attributes will be used. This has an effect on good data quality so that it can improve accuracy when used in the classification process later. the selection phase is commonly referred to as the data preparation phase or initial data processing with several existing techniques, namely:

1. Data validation, in this technique the data will be identified whether the data is a problem such as data outliers, data noise, data inconsistency, missing volume or incomplete data. If this is the case, then data-hungry measures will be taken
2. Data integration and transformation, used to improve the accuracy and efficiency of learning algorithms.
3. Data size reduction and discretization. To obtain a data set with fewer attributes and records with informative properties.

From the steps carried out, it can be described the method that will be proposed and carried out as follows:

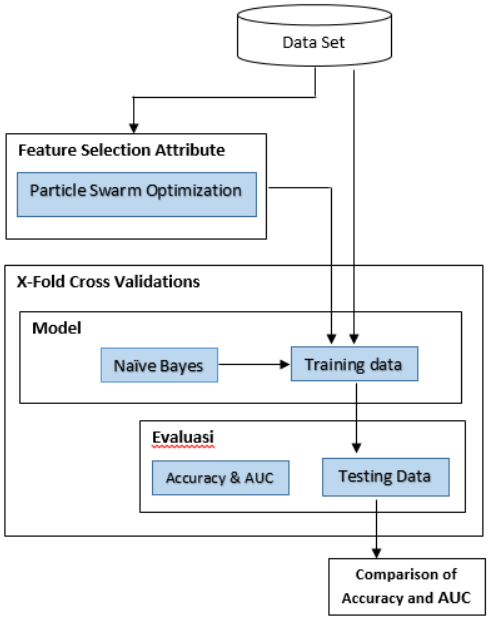

Figure 1. Proposed Method

An explanation of the description of the proposed method is as follows:
1. Diabetes dataset from US Hospital diabetes , database in UCI
2. There are two steps in this stage, the first is without using the attribute selection process and the second using a preprocessing process, namely the selection of attributes using the Particle Swarm Optimization (PSO) method.
3. The classification process uses the nave Bayes method
4. The results of the Accuracy and AUC values obtained.

Particle Swarm Optimization (PSO)

The steps for the experimental method carried out in this study are as follows:
1. Prepare the dataset
2. Calculate the weight of each attribute using Particle Swarm Optimization.
3. Determine the desired threshold. This allows attributes with a weight equal to the threshold or greater to be retained and discard attributes that are below the threshold.
4. Enter the weight value obtained in the Naive Bayes algorithm using the Naive Bayes formula
5. Calculate the classification results of the Naive Bayes algorithm using the Confusion Matrix and then measure the evaluation results.
6. Record the results of Accuracy and AUC obtained.

2.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is often used in research, because PSO has similar properties with Genetic Algorithm (GA). The advantage of PSO is that it is easy to implement and there are several parameters to adjust. The PSO system is initiated by a population of random solutions and then looks for the optimum point by updating each generation result. The approach used is more systematic mathematics to find solutions. Particle Swarm Optimization (PSO) was formulated by Edward and Kennedy in 1995. The thought process behind this algorithm is inspired by the social behavior of animals, such as flocking

birds or groups of fish [15]. Unlike GA, PSO does not have evolution operators such as crossover and mutation. Rows in the matrix are called particles (same as GA chromosomes). They contain variable values and are not binary encoded. Each particle moves around the surface of the particle at a speed. Each speed and position update is based on local and global best locations:

$$V_{i,m}^{new} = W.V_{i,m}^{old} + C_1\, x\, \left(P_{i,m}^{local\ best} - X_{i,m}^{old}\right) + C_2\, x\, R\, x\, (P_{i,m}^{global\ best} - X_{i,m}^{old})$$

Calculating the new velocity of each particle:

$$X_{i,m}^{new} = X_{i,m}^{old} - V_{i,m}^{new}$$

Description:
n : the number of particles in the group
d : dimension
V, : the velocity of the i-th particle in the i-th iteration
w : inertia weight factor
C1, C2: acceleration constant (learning rate)
R : random number (0-1)
X,:current position of the i-th particle in the i-th iteration
Pbesti : the previous best position of the i-th particle
Pgbest : the best particle among all the particles in a group or population

2.2 Naïve Bayes

Bayes is a statistical classification that can be used to predict the probability of membership of a class [15]. Bayes has very high accuracy and speed when applied to databases with large data. Here's Bayes' theorem:

$$P(H|X) = \frac{\mathrm{P(H|X)P(H)}}{\mathrm{P(X)}}$$

Description:
X = Data with unknown class
H = Hypothesis data x is a specific class
P(H|X) = Probability of hypothesis H based on condition X (posteriori probability)
P(H) = Hypothesis probability H (prior probability)
P(X|H) = Probability of X based on the conditions on the hypothesis H
P(X) = Probability of X

3. Results and Discussion

3.1 Initial Process

The dataset that has been obtained is from Diabetes 130-US hospitals obtained from the UCI database consisting of 50 attributes and 101767 data lines [10]. Of the 101767 lines of data, only 10000 lines of data will be taken so that the data processed is not too large and to facilitate the processing. The data used in the implementation of a priori algorithm are outlined in the following pattern:

Table 1. Datasate of Diabetes 130-US hospitals

No	Encounter_id	Patient_nbr	race	gender	age	… ..	Diabetes medbet
1	2278392	8222157	Caucasian	F	0-10	… .	No

2	149190	55629189	Caucasian	F	10-20	… .	Yes
3	64410	86047875	African American	F	20-30	… .	Yes
4	500364	82442376	Caucasian	M	30-40	… .	Yes
5	16680	42519267	Caucasian	M	40-50	… .	Yes

From the 10000 lines of data, we will divide it into 2 parts consisting of 7000 lines of training data. While as many as 3000 rows of data that will later be used as test data. The training data will then be transformed or converted from nominal data into numerical data. The data transformation is done by sorting the ordinal land into numeric form, for example, Female and Male data, so for Female it is given a numerical value of 0 while Male is given a numerical value of 1 so as to get the results as shown in the following table 2:

Table 2. Numerical data from Diabetes 130-US hospitals

No	Encounter_id	Patient_nbr	race	gender	age	…..	Diabetes medbet
1	2278392	8222157	2	1	0	….	1
2	149190	55629189	2	1	1	….	0
3	64410	86047875	1	1	2	….	0
4	500364	82442376	2	0	3	….	0
5	16680	42519267	2	0	4	….	0

After the data set is processed by converting the data into numeric data like the table above, then the data is ready to be used for calculating attribute weights using the Pearson correlation method which will select the 11 highest weights. After this is done, the test of the nave Bayes method and the nave Bayes method with Pearson correlation is ready to be carried out.

3.2 Selection of Relevant Attributes with Particle Swarm Optimization (PSO)

Before the dataset of diabetic patients was used in the calculation process of the nave Bayes method. The dataset with 50 attributes will be processed first in order to determine the selection of relevant attributes which will later be used in the calculation process of the nave Bayes method. The attribute selection process is carried out using the Particle Swarm Optimization method. The following are the steps for selecting the relevant attributes using the Particle Swarm Optimization method:

a. Calculating the weight of attribute values

To calculate the attribute correlation value, the equation method is used:

$$V_{i,m}^{new} = W.V_{i,m}^{old} + C_1\, x\, \left(P_{i,m}^{local\ best} - X_{i,m}^{old}\right) + C_2\, x\, R\, x\, (P_{i,m}^{global\ best} - X_{i,m}^{old})$$

b. Determine the threshold (limit) to be used

After obtaining the value of each attribute, the next step is to sort the values from the greatest value to the smallest value. Then determine the threshold or threshold level of importance (weight) of each of these

attributes. In the future, attributes that have a level of importance (weight) equal to the threshold or greater will be used or maintained, but for attributes that have a level of importance or weight value that is smaller or below the threshold value will be ignored or will not be used in the process. next calculation.
In this study, the threshold is emphasized on the numerical value of 0.0082 . The determination of the threshold value is based on several experiments carried out, namely by calculating the Naive Bayes method and Particle Swarm Optimization using various number of attributes. From the experiments carried out, it was found that the calculation with the number of attributes 34 resulted in an accuracy percentage of 97.13%,.

3.3 Evaluation and Validation

Based on the experiments that have been carried out to solve the problem of detecting diabetes, the results are as follows:

Table 3. Experiments Naive Bayes and Naïve Bayes + PSO

	Akurasi	AUC
Naïve Bayes	96.15%	0.991
Naïve Bayes + PSO	97,13%	0.995

From the table 3, it can be seen that the Naïve Bayes method has an accuracy value of 96.15% with an AUC value of 0.991. While testing the Naïve Bayes method based on Particle Swarm Optimization attribute selection, the results obtained an accuracy value of 97.13% with an AUC value of 0.995.

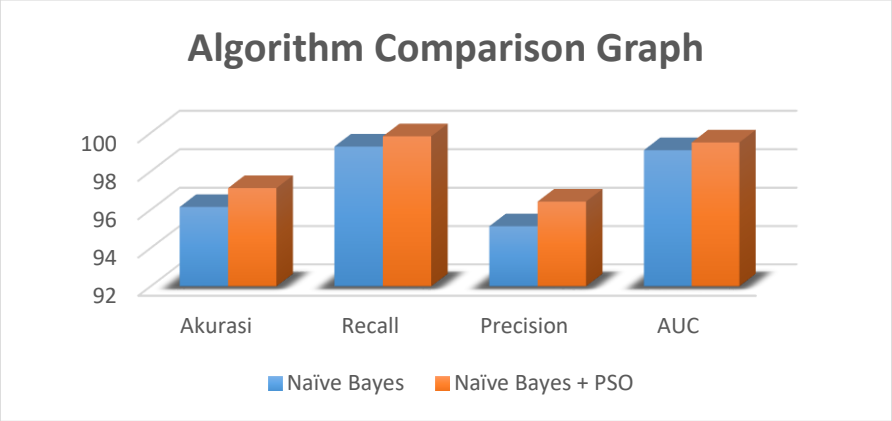

Figure 2. Algorithm Comparison Graph

Through these experiments, it can be seen that the accuracy and AUC values of the experiment using the PSO-based Naïve Bayes method have increased compared to the results obtained from the Naïve Bayes method only experiment.

This can be seen from the results of increasing the accuracy of the method as much as 0.98%. The increase in accuracy is calculated from the experimental results of the Nave Bayes method which only has an accuracy value of 96.15% and then changes to 987.13% when experimenting with the Naïve Bayes method using the Particle Swarm Optimization (PSO) method as an attribute selection.

4. Conclusions

So it can be concluded from testing diabetes data using the Naïve Bayes method based on Particle Swarm Optimization attribute selection is better than the Naïve Bayes method alone. Thus, it can be said

that the application of Particle Swarm Optimization (PSO) for attribute selection in improving the performance of nave Bayes in the detection of diabetes provides a solution to the problem seen from the increase in the value of better accuracy.

Reference

[1] N. Kumar and M. Abedin, "Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm," Comput. Methods Programs Biomed., 2017.
[2] M. Chan, "Global Report On Diabetes," World Heal. Organ., 2013.
[3] K. Kesehatan, "Situasi dan Analisis Diabetes." Kesehatan, Infodatin pusat data dan informasi kementrian RI, 2014.
[4] F. Mansourypoor and S. Asadi, "Development of a Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System for Diabetes Diagnosis Fatemeh," Comput. Biol. Med., 2017.
[5] L. Zhang, L. Jiang, C. Li, and G. Kong, "Two Feature Weighting Approaches for Naive Bayes Text Classifier," Knowledge-Based Syst., 2016.
[6] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, "Self-adaptive attribute weighting for Naive Bayes classification," Expert Syst. Appl., vol. 42, no. 3, pp. 1487–1502, 2015.
[7] Ö. F. Arar and K. Ayan, "A Feature Dependent Naive Bayes Approach and Its Application to the Software Defect Prediction Problem," Appl. Soft Comput. J., 2017.
[8] M. K. J.Han, J.Pei, Data Mining: Concepts and Techniques, vol. 3. 2012.
[9] D. Farid, L. Zhang, C. Mofizur, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," Expert Syst. Appl., vol. 41, no. 4, pp. 1937–1946, 2014.
[10] "Diabetes Hospital." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008..
[11] Y. Mu, X. Liu, and L. Wang, "A Pearson's correlation coefficient based decision tree and its parallel implementation," Inf. Sci. (Ny)., vol. 435, pp. 40–58, 2018
[12] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records," Int. J. Med. Inform., 2016.
[13] N. Nai-arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," Procedia - Procedia Comput. Sci., vol. 69, pp. 132– 142, 2015.
[14] Arifin T., Herliana A., 2018, Optimasi Metode Klasifikasi dengan menggunakan Particle Swarm Optimization untuk Identifikasi Penyakit Diabetes Retinopathy, Khazanah Informatika (Jurnal Ilmu Komputer dan Informatika), Vol 4, No 2, pp. 77-81, 2018
[15] BUDI SANTOSA, 2010, "Tutorial Particle Swarm Optimization," Institut Teknologi Surabaya, Surabaya