International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 2, No. 2, September 2019, pp. 56-60

Naive Bayes Algorithm Using Selection of Correlation Based Featured Selections Features for Chronic Diagnosis Disease

Irfan Santiko [a,1,*], Ikhsan Honggo [b,2]

[a] Informatic Departement Universitas AMIKOM Purwokerto, Let.Jend. Pol Sumarto, Purwokerto, 53113, Indonesia
[b] Informatic Departement Universitas AMIKOM Purwokerto, Let.Jend. Pol Sumarto, Purwokerto, 53113, Indonesia
[1] irfan.santiko@amikompurwokerto.ac.id *; [2] iksanhonggo02@gmail.com
* corresponding author

Abstract

Chronic kidney disease is a disease that can cause death, because the pathophysiological aetiology resulting in a progressive decline in renal function, and ends in kidney failure. Chronic Kidney Disease (CKD) has now become a serious problem in the world. Kidney and urinary tract diseases have caused the death of 850,000 people each year. This suggests that the disease was ranked the 12th highest mortality rate. Some studies in the field of health including one with chronic kidney disease have been carried out to detect the disease early, In this study, testing the Naive Bayes algorithm to detect the disease on patients who tested positive for negative CKD and CKD. From the results of the test algorithm accuracy value will be compared against the results of the algorithm accuracy before use. After feature selection using feature selection Featured Correlation Based Selection (CFS), it is known that Naive Bayes algorithm after feature selection that is 93.58%, while the naive Bayes without feature selection the result is 93.54% accuracy. Seeing the value of a second accuracy testing Naive Bayes algorithm without using the feature selection and feature selection, testing both these algorithms including the classification is very good, because of the accuracy value above 0.90 to 1.00. It was included in the excellent classification—higher accuracy results.

Keywords:Chronic Kidney Diseas; Naive Bayes; CFS;

1. Introduction

Chronic Kidney Disease (CKD) has now become a serious health problem in the world. According to [1] and Burden of Disease, kidney and urinary tract diseases have caused the death of 850,000 people each year. This suggests that the disease was ranked the 12th highest mortality rate. Chronic Kidney Disease is a process pathophysiological etiology diverse, leading to decreased kidney function progressive, and usually ends with a clinical condition characterized by decreased renal function irreversibly, to a degree requiring renal replacement therapy were fixed, in the form of dialysis or transplantation kidney [6].

Based on interviews conducted research on chronic kidney disease can be concluded that chronic kidney disease is very dangerous, because if it is not dealt with immediately, the kidneys may stop functioning if the kidney stops functioning, the result is fatal, the risk of death. Chronic kidney disease can occur at any age, there is no age limit, but are more common in adults, and is more common in men than in women.

Parkinson's is a disease that attacks the brain and can cause a gradual loss of motor control due to a lack of dopamine in the brain. This disease results in movement become slow and muscles become stiff. The disease is difficult to prevent and cure because the exact cause is unknown [1]. With the feature selection is expected to improve the performance of data analysis in the classification more accurate, but the implementation is a feature reduction can significantly affect the results of the classification. In his research used four methods of feature selection Featured namely Correlation Based Selection, Wrapper, Gain Ratio and RELIEF were then classified using two methods of classification Bagging J48 and SMO. Bagging algorithm J48 on the results of feature selection methods Featured Selection Based Correlation can improve the accuracy of the results. Values obtained accuracy becomes better, namely 91% compared to prior to the feature selection of 90% [1].

Tuberculosis is an infectious disease caused by a bacterium called Mycobacterium tuberculosis and is the highest cause of death occurring in the productive age of 15-50 years, the weak economy, and less educated. In research to be done is to compare several methods of classification data mining, among which the algorithm C4.5, Naive Bayes, Neural Network and Logistic Regression, four methods used in predicting the diagnosis of tuberculosis in order for the selected algorithm is the most accurate algorithm so can make early diagnosis of tuberculosis disease. note that the naïve Bayes algorithm has the highest accuracy value, the result of calculation accuracy naïve Bayes with 91.61%, followed by C4 algorithm [4]

2. Method

The following description of the flowchart of the study:

2.1 Identification of problems

Problem identification is performed to determine the problems and appropriate methods so that it can do the proper steps to diagnose chronic kidney disease.

2.2 Data collection

Preliminary data obtained from the UCI Machine Learning Repository namely dataset Cronic disease Kidney Disease.

2.3 Feature Selection

Selection feature works find distinguishing characteristics that represent the main characteristics of the signal and reducing the dimension of the signal into a set of numbers that are slightly but representative. In this study using feature selection CFS (Correlation Based Features Selection) using Weka application with the aim to obtain a dataset with selected features. 155 instances with 24 attributes that have been eliminated missing values her with the intention to obtain a dataset with the features selected, and produced 12 attributes are selected, namely bp, sg, al, su, RBC, ba, bgr, sc, hemo, PCV, wbcc, rbcc , HTN and the class attribute decision CKD and notckd which describes a positive decision with chronic kidney disease and notckd explain the negative decision with chronic kidney disease.

2.4 Results Analysis

Measurement of the performance of a data mining algorithm can be based on several aspects among others of accuracy, computing speed, robustness, scalability and interpretabilitas. This study only measuring the performance of a data mining algorithm based on aspects of accuracy.

What is the result of Accuracy in getting to diagnose diabetes mellitus based on the value of accuracy. with the level of diagnosis, namely: excellent classification = 0.90 - 1:00, good classification = 0.80 - 0.90, poor classification = 0.60 - 0.70 (Gorunescu, 2011). Conclusions are to answer the research done on purpose by the author.

3. Results and Discussion

3.1 Identification of problems

With the feature selection is expected to improve the performance of data analysis in the classification more accurate, but the implementation is a feature reduction can significantly affect the results of the classification. CFS feature selection can improve the accuracy becomes better results (Vita, 2014). CFS algorithm is the most stable algorithm, the duration of the classification of the fastest and highest accuracy.

dataset used a dataset of chronic kidney disease are taken from the database on UCI machine learning repository. This dataset contains 400 instances with 24 regular and first class attributes.

3.2 Preprocessing Data

a. Dataset Early

Table 1 Dataset early chronic kidney disease

No.	age	blood preasure	the specific gravity	albumin	...	appetite	pedal edema	anemia	class
1	48.0	80.0	1,020	1	...	good	No.	No.	ckd
2	7.0	50.0	1,020	4	...	good	No.	No.	ckd
3	62.0	80.0	1,010	2	...	poor	No.	yes	ckd
4	48.0	70.0	1,005	4	...	poor	yes	yes	ckd
5	51.0	80.0	1,010	2	...	good	No.	No.	ckd
6	60.0	90.0	1,015	3	...	good	yes	No.	ckd
7	68.0	70.0	1,010	0	...	good	No.	No.	ckd
8	24.0	???	1,015	2	...	good	yes	No.	ckd
9	52.0	100.0	1,015	3	...	good	No.	yes	ckd
10	53.0	90.0	1,020	2	...	poor	No.	yes	ckd
...	...	...	...	...	...	...	...	...	...
400	58.0	80.0	1,025	0	...	good	No.	No.	notckd

Table 1 An initial dataset chronic kidney disease, which has 24 regular attributes and one attribute class, 400 instances.

b. Handling Missing Value
On the dataset chronic kidney disease has 24 attributes that there are many missing values on each attribute. Of total instances as much as.

3.3 Feature Selection

The next stage after going through the process of missing values is to do the feature selection. In this study using feature selection Correlation Based Featured Selection (CFS), a software application Weka, to select attributes that are important and influential in the calculation of the final results of accuracy, and has a major contribution to the process of classification and reduce the computational load so much faster.

3.4 Classification Algorithm Implementation

In this research will be done some testing, the testing by processing datasets ready for use after the handling of missing values, will be tested into Naive Bayes algorithm without any selection first feature, the first test will be processed using the Rapid Miner software. The steps are as follows:

a. Rapid Miner open the software, enter the ready-made datasets that have gone through the process of handling missing values into the software by adding datasets format, in this study using .ARFF format. then select datasets to be processed.
b. Enter Role Set, the attribute name is filled class, and the role filled label targets, as set this role describes the attributes that will be processed.
c. Enter xvalidation, is used to determine the results of the confusion matrix. Then double click for further validation process.
d. Enter the algorithm to be processed in the training room, in this study using a Naive Bayes algorithm, then the testing chamber insert apply the model, then enter the performance, is used to calculate the results of the accuracy of the algorithm is processed.
e. After being processed in software running RapidMiner, makadi Naive Bayes algorithm results obtained before using the feature selection shows the results of the accuracy of 93.54% with 0 seconds of computing time. Can be calculated also the result of the accuracy through the confusion matrix as in table 1.3 below.

Table 2 Results of calculation accuracy

Class	ckd	Notckd
ckd	41	1
notckd	9	104

Table 1.3 shows the true positive (tp) 41 records, false negative (fn) 1 record, false positive (fp) 9 record, and true negative (tn) 104 records.
Calculation confusion matrix:

accuracy = $\frac{TP+TN}{TP+FN+FP+TN} x\ 100\%$

Accuracy = x 100%$\frac{41+104}{41+1+9+104}$

Accuracy = x 100%$\frac{145}{155}$

Accuracy = 93.54%

After the calculation process the data mining algorithm without using feature selection, the author will also test using Naive Bayes algorithm using feature selection Featured Correlation Based Selection. The steps are as follows.

1) Rapid Miner open the software, enter the dataset that has been done previously on the feature selection Weka software using the Featured Selection Based Correlation and already there are no missing data

values. Enter the Read ARFF in the process chamber, as this study dataset using .ARFF format. then select datasets to be processed.

2) Enter Role Set, the attribute name is filled class, and the role filled label targets, as set this role describes the attributes that will be processed.
3) Enter xvalidation, is used to determine the results of the confusion matrix. Then double click for further validation process.
4) Enter the algorithm to be processed in the training room, in this study using a Naive Bayes algorithm, then the testing chamber insert apply the model, then enter the performance, is used to calculate the results of the accuracy of the algorithm is processed.
5) After being processed in software running RapidMiner, will gain from the process of Naive Bayes algorithm using feature selection Featured Correlation Based Selection, which made the process of running repeated 10 times, the value of its accuracy is 93.58% with 0 seconds of computing time. The following is a calculation of the value of accuracy of Naive Bayes algorithm using the Correlation feature selection Selection Based on the confusion matrix Featured:

Table 1.4 Confusion matrix Naive Bayes using feature selection

Class	**ckd**	**Notckd**	**Class precission**
ckd	41	1	97.62%
notckd	9	104	92.04%
recall	82.00%	99.05%	

Table 1.4 shows the true positive (tp) 41 records, false negative (fn) 1 record, false positive (fp) 9 record, and true negative (tn) 104 records.

Calculation confusion matrix:

accuracy = $\frac{TP+TN}{TP+FN+FP+TN} x$ 100%

Accuracy = $\frac{41+104}{41+1+9+104}$x 100%

Accuracy = $\frac{145}{155}$x 100%

Accuracy = 93.58%

3.5 Comparative Analysis of Data Mining Algorithms

Based on calculations that have been done, the results of the accuracy of the value of Naive Bayes algorithm without using feature selection and the Naive Bayes algorithm using feature selection Featured Correlation Based Selection, will be compared to determine how the results of each calculation of the data mining algorithms. Here as in table 1.5 of this.

Table 1.5 The result of the calculation algorithm

No.	**Naive Bayes algorithm**	**Results Accuracy (%)**	**time Computing**
1	Without feature selection CFS	93.54%	0 sec

2	Using the feature selection CFS	93.58%	0 sec

Based on the above table produced some results accuracy and computing time is processed from Rapid Miner software, with Naive Bayes algorithm without feature selection, and Naive Bayes algorithm using feature selection Featured Correlation Based Selection.

3.6 Result & Analysis

Based on the results of the analysis show that the accuracy of some tests, using a Naive Bayes algorithm without generating feature selection accuracy value 93.54%, and Naive Bayes algorithm using the Correlation feature selection Selection Based Featured generate value 93.58% accuracy. Based on the outcome of some tests that have been carried, Naive Bayes algorithm using the Correlation feature selection Selection Based Featured proven to increase the accuracy of the results by a margin of 0:04% of Naive Bayes algorithm without using feature selection. For the results of the same computing time, which is 0 seconds.

Based on the results of several test accuracy that has been done on Naive Bayes algorithm, the accuracy of the results obtained without using feature selection is 93.54%, and 93.58% accuracy results using feature selection. The results of the accuracy of the diagnosis is the excellent level of classification.

4. Conclusion

From some of the testing that was done, the accuracy of the results obtained using naïve Bayes algorithm without feature selection is 93.54%, and the Naive Bayes algorithm using the Correlation feature selection Selection Based Featured accuracy result is 93.58%. For the computation time of testing the Naive Bayes algorithm without feature selection, and as well as using a correlation-based feature selection selection featured the same result, ie 0 seconds. It can therefore be inferred using Naive Bayes algorithm using feature selection can improve classification accuracy results for the diagnosis of chronic kidney disease. Value 93.58% accuracy including into excellent classification

Acknowledgment

In order for this research to continue to grow, here are suggestions proposed:

a. Testing performed using the entire dataset bank chronic kidney disease, which amounts to 15 157.
b. Tests using different algorithms on feature selection algorithm Featured Selection Based Correlation or using any other feature selection.

References

[1] Dewi, Sarini Vita. 2014. Analysis of Performance Classification For the diagnosis of Parkinson's Disease. Yogyakarta. Gadjah Mada University.
[2] Gorunescu, Florin. 2011. Data Mining Concepts, Models and Techniques. Romania.
[3] Bradley, P., Andrew, P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159.
[4] Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
[5] Dash, D., Cooper, G.F., 2002. Exact model averaging with naive Bayesian classifiers. In: Proceedings of the 19th European Conference on Machine Learning, Morgan Kaufmann, Sydney, Australia, pp. 91–98.
[6] Dash, D., Cooper, G.F., 2004. Model averaging for prediction with discrete Bayesian networks. J. Mach. Learn. Res. 5, 1177–1203.
[7] Feng, G., Guo, J., Jing, B.Y., Sun, T., 2015. Feature subset selection using naive Bayes for text classification. Pattern Recognit. Lett. 65, 109–115.
[8] Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.
[9] Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA,⟨http://archive.ics.uci.edu/ml⟩.
[10] Hall, M., 2000. Correlation-based feature selection for discrete and numeric classmachine learning. In: Proceedings of the 17th International Conference on Machine Learning, pp. 359–366.
[11] Hall, M., 2007. A decision tree-based attribute weightingfilter for naive Bayes. Knowl.-Based Syst. 20 (2), 120–126.
[12] Hall, M., Holmes, G., 2003. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans. Knowl. Data Eng. 15 (6), 1437–1447.
[13] Javed, K., Maruf, S., Babri, A., Haroon, 2015. A two-stage Markov blanket based feature selection algorithm for text classification. Neurocomputing 157, 91–104.
[14] Jiang, L., Cai, Z., Wang, D., 2010. Improving naive Bayes for classification. Int. J. Comput. Appl. 32 (3), 328–332.
[15] Khalil, E.H., 2014. A noise tolerantfine tuning algorithm for the naive Bayesian learning algorithm. J. King Saud Univ. – Comput. Inf. Sci. 26 (2), 237–246.