Scientific Journal of Informatics
Vol. 3, No. 2, November 2016
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naïve Bayes Classifier for The Classification of the Ratio of Inpatients

M Rizky Wijaya[1], Ristu Saptono[2], Afrizal Doewes[3]
[1,2,3]Informatics, FMIPA, Sebelas Maret University
Email: [1]muhamad.rw@student.uns.ac.id, [2]ristu.saptono@staff.uns.ac.id, [3]afrizal.doewes@staff.uns.ac.id

Abstract

Diabetes can lead to mortality and disability, so patients should be inpatient again to undergo treatment again to be saved. On previous research about feature selection with greedy stepwise forward fail to predict classification ratio inpatient of patient with the result of recall and precision 0 on data training 60%, 75%, 80%, and 90% and there is suggestion to handle unbalanced class data problem by comparison of data readmitted 6293 and the otherwise 64141. The research purposed to know the effect of choosing the best model using best first instead of greedy stepwise forward and data sampling with spreadsubsample to resolve unbalanced class data problem. The data used was patient data from 130 American Hospital in 1999 until 2008 with 70434 data. The method that used was best first search and spreadsubsample. The result of this research are precision found 0.4 and 0.333 on training dataset 75% and 90% with best first method, while spreadsubsample method found that value of precision and recall is more significantly increased. Spreadsubsample has more effect with the result of precision and recall rather than using best first method.

Keyword: best first, unbalanced dataset, spreadsubsample, classification, feature selection

1. INTRODUCTION

Hyperglycemia chronic conditions in diabetic patients known associated with increased mortality and morbidity, so the hospital need to make a policies on hyperglycemia treatment [1]. Early research has done wrapper features selection greedy stepwise with naive classifier bayes for the classification of the patients diabetes hospitalized, found there are irregularities on a recall and precision at confusion matrix from model greedy stepwise forward with data training 66%, 75%, 80% and 90% got 0 (zero result) on readmitted condition. There are suggestions to solve the problems unbalanced class in the early research, 6293 class the return of patients(readmitted) and 64141 did not return(otherwise) [2].

Features selection is the selection of a subset features have little dimension that contribute much on accuracy, otherwise it would be eliminated features that unnecessary [3]. Features selection applied to reduce some features in many application where data has hundreds or thousands of features. The selection of focusing on find features relevant [4].

The two approach in features selection, the filter approach and approach wrapper.Filter approach, any feature evaluated independently with respect to label class in training sets and rank of all features, which features with the top selected. Approach wrapper uses the search artificial intelligence classics like greedy to find a subset best of features, and recursively evaluate a subset features a different induction

certain algorithm. Features vote with wrapper smaller than a subset of the original, that model is more understandable [5].

Research that will be implemented using best first search and sampling data by spreadsubsample.Spreadsubsample is one of technique undersampling (reducing class data major) where produce random subsample from a dataset [6]. Best first will be used in the search for combining every features. Best first is greedy hillclimbing method with backtracking facilities, methods to generate nodes from nodes(currently is best node according to him)[7]. The use of best first to this research for testing whether greedy stepwise forward that makes value 0 appear. Spreadsubsample used to test whether the 0 appear because of unbalanced class (status readmitted/otherwise).

The purpose of this research is to find the best first and spreadsubsample on selection wrapper features and the classification methods naive bayes for the classification of inpatient, that would give a prediction patients who will inhopitalized again for treatment so patients can avoid morbidity and mortality.

2. METHOD

Data on this research using data patients diabetic on 130 hospital in america in 1999-2008, dataset was taken from the university of california repository irvine calif repositories (uci). There are 50 feature on this dataset: encounter id, patient number, admission type, discharge disposition, admission source, time in hospital, payer code, medical specialty, race, gender, age, weight, number of outpatient visits, number of emergency visits, number of inpatient visits, readmitted, numbers of lab procedures, number of procedures, number of medications, diagnosis 1, diagnosis 2, diagnosis 3, number of diagnoses, glucose serum test result, a1c test result, change of medications, diabetic medications, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone.

The workflow of the program on this research can be seen in Figure 1. Dataset processed with spreadsubsample with distribution 6, 7, 8, 9, 10 and full data without distribution. The data is divided into training data and test data. The data training processed with wrapper feature selection along with the search method and his naïve bayes. Data training become material by naïve bayes then will be tested for classifying data, so resulting confusion matrix.

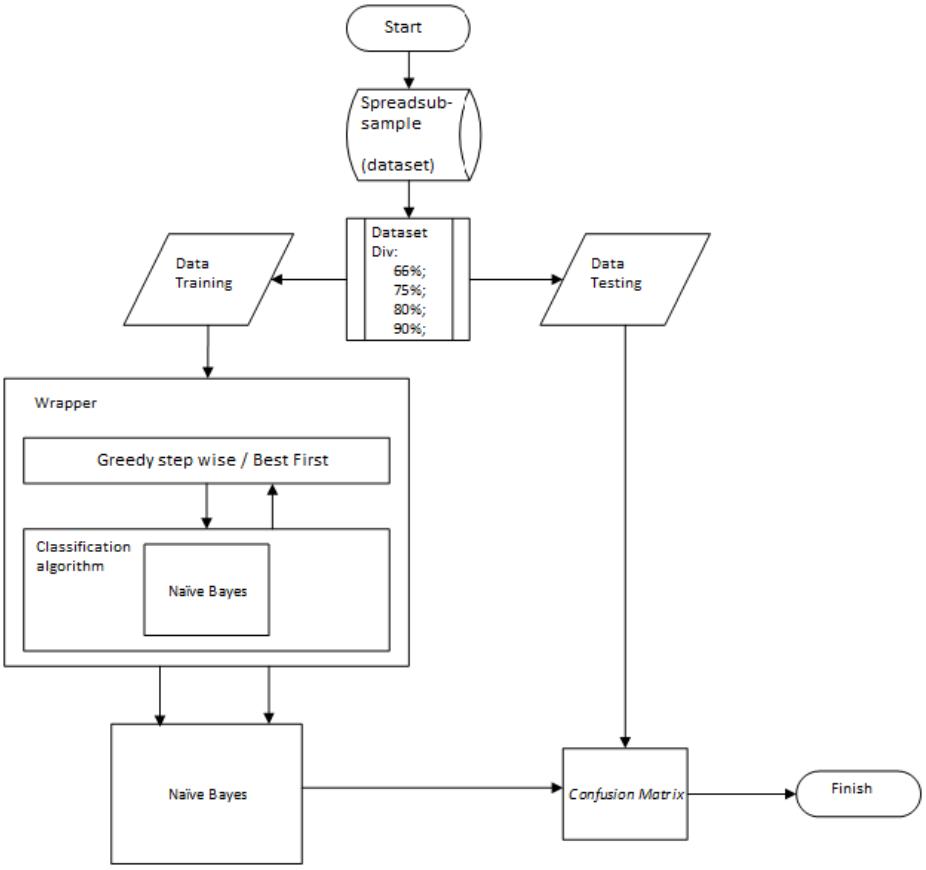

Figure 1. Method implementation

2.1 Forward Feature Selection

This method works with features of inserted one after another according to large order their influence on the model and stop if everything that qualifies have come. Started by examining matrix and then took features free that produces a subset features with evaluasinya value of accuracy. All a subset features taken. Given hypothesis limit to a subset received or rejected. Until they reached a subset features with best value accuracy. [8].

2.2 Wrapper

Wrapper feature selection: Features selection this type wrapper selecting features by conducting an election simultaneously with the conduct of modeling. The election of this type use of a criteria use the classification rate of the pengklasifikasian / modeling used. To reduce computational cost, the selection of generally conducted by using classification rate of the pengklasifikasian / modeling for modeling with the lowest.

For type wrapper, need to first made features a subset selection before determine a subset which is a subset with the upper best. Features subset selection can be done by using method sequentialforward selection(from one be many features), sequential backward selection(of numerous into one), sequential floating selection(can from anywhere), ga, greedy search, hill climbing, simulated annealing [9]. Wrapper process can be seen at figure 2.

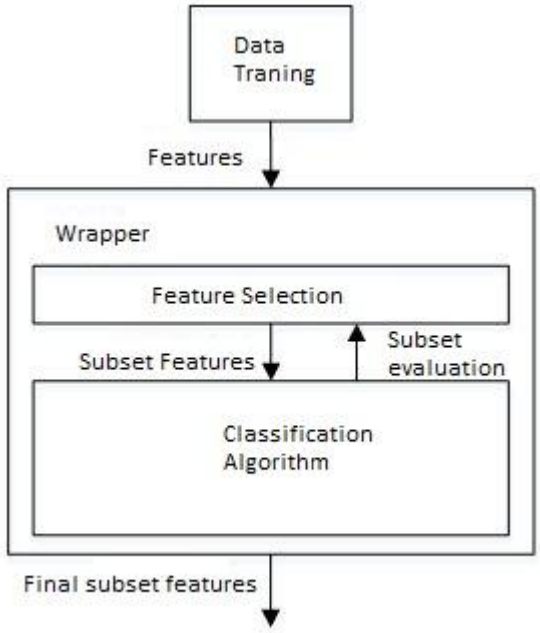

Figure 2. Wrapper feature selection

2.3 Greedy Stepwise Forward

Selection features is a process in the data mining in which this process select m a subset features of n features early. The purpose of the implementation of features selection is to find features contribute of the results of classifications by means of ignoring features not relevant and features have the role of same(redundant). The selection of begins with an empty of feature as a subset, features best of the original features determined and added to the set of a subset. On each iteration next will be added features into a subset [10]. Feature selection greedy stepwise forward can seen on Table 1.

Table 1. Example of feature selection process

Greedy Stepwise Forward Selection	Induction of Naïve Bayes Classifier

Greedy Stepwise Forward Selection	**Induction of Naïve Bayes Classifier**
OriginalFeature: {F_1, F_2, F_3, F_4, F_5, F_6, F_C}	
Feature Selection: {}	
First iteration: {F_1: 88}, {F_2: 88.5}	
Second iteration: {F_2,F_3: 89}, {F_2,F_4: 89}, {F_2,F_5: 91}	
Third iteration: {F_2,F_5,F_6: 91.5}	{F_1, F_2, F_3, F_4, F_5, F_6, $F_{C:88.8}$}
Feature Selection final: {F_2,F_5,F_6,F_C}	

2.4 Best First Search

Best first searching nodes expanded one by one based on best definition. An algorithm search simple all node order to a single criterion, a kind of using cost estimates through node solution [6]. An algorithm best first search is a division of type informed search. An algorithm it uses a heuristic values on every opened node. A node with a heuristic best value will open first. If goal state has not been found, will be examined at the nodes the next with a heuristic best value at the same depth. A node was opened and then examined whether there is goal state on branch. If goal state has not been found, will occur similar process at the next nodes.

2.5 Naïve Bayes Classifier

Naïve bayes classifier is the classification methods based on bayes theorem.The main characteristic of naïve bayesclassifier is assumption powerful independence from their conditions. Naïve bayes classifier is classifications with a model statistics to calculate possibility of a class each group attributes there, and determine class which the optimal. All the attributes will contribute in decision making, with weights attributes same important and any attribute mutual independence each other [11]. The basis of theorem naïve bayes classifier used in programming is bayes formula shown on formula (1).

$$P(H|X) = \frac{\boldsymbol{P(X|H).P(H)}}{\boldsymbol{P(X)}} \quad (1)$$

Note:

P(H|X) = probability of posterior H inside X

P(X|H) = probability of posterior X inside H
P(H) = probability of prior from H
P(X) = probability of prior from X

2.6 Spreadsubsample

Deprive sample in class the majority until the sample the majority of the minority more equal to the ratio particular. How to work spreadsubsample is to produce subsample at random from a dataset [6].

2.7 Confusion Matrix

Confusion matrix is a method typically used to perform calculation accuracy on the mining data.This equation performing calculations with four output the recall, precision, accuracy, and error rate [12].

- recall is the proportion of positive cases identified correctly.The calculations are see formula (2).
- precision is the abilities of a system to match information an answer with demand.This research states that precision are defined the ability system to know patients will undergo in-patient back. The calculation see formula (3).
- accuracy is the ratio cases identified true by the sum of all cases. The calculation see formula (4).
- error rate are comparisons cases identified wrong with all cases.

Table 2. Confusion matrix

Test Outcome	**Condition**	
	Condition Positive	**Condition Negative**
Test Outcome Positive	True Positif	False Positif
Test Outcome Negative	False Negatif	True Negatif

Formula of recall, precision dan accuracy:

$$\text{Recall} = \frac{\text{TruePositif}}{\text{TruePositif+FalseNegatif}} \quad (2)$$

$$\text{Precision} = \frac{\text{TruePositif}}{\text{Truepositif+FalsePositif}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TruePositif+TrueNegatif}}{\text{TruePositif+TrueNegatif+Falsepositif+FalseNegatif}} \quad (4)$$

True positive and negative is true situation where outcomes matches the really happened false false positive and negative is situation where outcomes not in accordance with the condition of being actually happened.

3. RESULT AND ANALYSIS

The results showed that spreadsubsample affect the recall. The less distribution or comparative data major with minor so recall getting higher. To see the result of the recall replacement greedy stepwise forward with best first can be seen in table 3 on distribution full or no distribution. The recall seen in Table 3.

Table 3. Recall distribution of spreadsubsample

Method	6	7	8	9	10	full
Greedy + NBC data training 66%	0.005	0.005	0.002	0.005	0.001	0
Greedy + NBC data training 75%	0.008	0.003	0.003	0.004	0.001	0
Greedy + NBC data training 80%	0.006	0.003	0.003	0.005	0.002	0
Greedy + NBC data training 90%	0.003	0.002	0.002	0.002	0.003	0
Bestfirst + NBC data training 66%	0.006	0.002	0.002	0.004	0.001	0
Bestfirst + NBC data training 75%	0.001	0.003	0.003	0.003	0.002	0.001
Bestfirst + NBC data training 80%	0.006	0.004	0.001	0.004	0.002	0
Bestfirst + NBC data training 90%	0.003	0.003	0	0.003	0.003	0.002
NBC data training 66%	0.125	0.096	0.093	0.09	0.097	0.08
NBC data training 75%	0.118	0.091	0.087	0.09	0.095	0.078
NBC data training 80%	0.112	0.087	0.089	0.087	0.094	0.074
NBC data training 90%	0.111	0.092	0.086	0.077	0.083	0.071

Following the results of research show that the spreadsubsample affect the value of precision.The less the distribution or comparison data major with minor hence precision getting high. To see the result of the replacement of precision method stepwise greedy forward with the best first can be seen in table 4 on distribution full or no distribution. Precision results can be seen in Table 4.

Table 4. Precision distribution of spreadsubsample

Method	6	7	8	9	10	full
Greedy + NBC data training 66%	0.294	0.435	0.455	0.333	0.25	0
Greedy + NBC data training 75%	0.5	0.385	0.5	0.25	0.5	0
Greedy + NBC data training 80%	0.467	0.286	0.8	0.375	0.286	0
Greedy + NBC data training 90%	0.333	0.143	1	0.333	0.5	0

Bestfirst + NBC data training 66%	0.31	0.417	0.625	0.286	0.25	0
Bestfirst + NBC data training 75%	0.455	0.5	0.8	0.238	0.273	0.4
Bestfirst + NBC data training 80%	0.438	0.357	0.5	0.417	0.273	0
Bestfirst + NBC data training 90%	0.333	0.222	0	0.333	0.667	0.333
NBC data training 66%	0.31	0.272	0.243	0.239	0.227	0.21
NBC data training 75%	0.32	0.263	0.237	0.248	0.224	0.208
NBC data training 80%	0.302	0.261	0.241	0.232	0.228	0.2
NBC data training 90%	0.308	0.26	0.239	0.203	0.195	0.204

The results showed that spreadsubsample affect the accuracy. The less distribution or comparative data major with minor its accuracy become lower. To see the result of the accuracy replacement greedy stepwise forward with best first can be seen in Table 5 on distribution full or no distribution. The accuracy seen in Table 5.

Table 5. Accuracy distribution of spreadsubsample

Method	6	7	8	9	10	full
Greedy + NBC data training 66%	85.7448	87.7257	88.9495	89.9701	90.7418	90.993
Greedy + NBC data training 75%	85.7078	87.6212	88.9611	90.1793	90.9049	90.7996
Greedy + NBC data training 80%	85.6413	87.2778	88.8673	90.2829	90.8487	90.6864
Greedy + NBC data training 90%	85.2667	87.5844	88.7359	90.704	91.1008	90.629
Bestfirst + NBC data training 66%	85.7315	87.7315	88.9651	89.9607	90.7418	91.0139
Bestfirst + NBC data training 75%	85.6806	87.645	88.9823	90.1856	90.876	90.811
Bestfirst + NBC data training 80%	85.63	87.2976	88.8408	90.2987	90.8342	90.6793
Bestfirst + NBC data training 90%	85.2667	87.5844	88.7182	90.6881	91.1153	90.6148
NBC data training 66%	83.6616	85.7744	86.7892	88.0632	88.6047	89.0304
NBC data training 75%	83.8191	85.6189	86.8211	88.4757	88.9518	88.7949
NBC data training 80%	83.5528	85.3114	86.6955	88.368	88.8118	88.6278
NBC data training 90%	83.269	85.5979	86.5996	88.6064	88.8038	88.698

To see the outcome of replacement method greedy stepwise forward with best first can be seen in table 6 on distribution full or no distribution. Features produced by best first less than that of the greedy stepwise forward. Results of an election features seen in Table 6.

Table 6. Sum of selected features distribution of spreadsubsample

Method	6	7	8	9	10	full
Greedy + NBC data training 66%	23	24	26	25	26	32
Greedy + NBC data training 75%	23	26	21	30	35	38
Greedy + NBC data training 80%	25	24	26	31	27	40
Greedy + NBC data training 90%	22	27	26	24	30	37
Bestfirst + NBC data training 66%	7	9	12	12	4	1
Bestfirst + NBC data training 75%	15	8	11	10	13	7
Bestfirst + NBC data training 80%	13	19	10	16	12	7
Bestfirst + NBC data training 90%	5	17	8	8	7	8

4. CONCLUSION

Based on the results of this research can be taken conclusion that methods replacement of greedy stepwise forward with best first affects the recall and precision. On the model of by features selection best first can choose fewer a feature and its accuracy increased. The result of spreadsubsample more significant from being with replacement method greedy stepwise forward with best first, precision and recall increased but the less distribution under its accuracy declining.

5. REFERENCE

[1] Zuanetti, G., Latini, R. & Maggioni, A., 1993. Influence of diabetes on mortality in acute: data from the GISSI-2 study. *J Am Coll Cardiol*. Vol 22(7):1788-1794.

[2] Alim, M. S. 2016. Seminar Hasil m0508125 Syahirul : Penerapan Method Seleksi Feature Wrapper Greedy Stepwise Dengan Naïve Bayes Classifier Untuk Klasifikasi Rasio Pasien Rawat Inap. https://www.scribd.com/doc/310822916/Seminar-Hasil-m0508125-Syahirul, diakses 07 September 2016.

[3] Deepa, T. & Ladha, L., 2011. Feature Selection Methods and Algorithms. *International Journal on Computer Science and Engineering (IJCSE).* Vol 3(5):1787.

[4] Yu, L. & Liu, H., 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *The Journal of Machine Learning Research*. Vol 5:1205-1224.
[5] Kohavi, R. & John, G. H., 1998. The Wrapper Approach. Feature extraction, construction and selection. Springer US.
[6] Hall, M., Frank, E, Holmes, G., Pfahringer. B., Reutemann, P. and Witten, I. H. 2009. The WEKA data mining software: An update. *SIGKDD*. Vol 11(1):10–18.
[7] Christopher, W., Thayer, J. & Ruml, W., 2010. A Comparison of Greedy Search Algorithms. Department of Computer Science University of New Hampshire, Durham.
[8] Sembiring, R. K., 1995. Analisa Regresi. Penerbit ITB, Bandung.
[9] Kohavi, R. & John, G. H., 1997. Wrappers for feature subset selection. *Artificial Intelligence*. Vol 97(1-2): 273-324.
[10] Han, J. & Kamber, M., 2006. Data Mining Concepts and Techniques 2nd ed. Elsevier, San Francisco.
[11] Kusumadewi, S., 2009. Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification. *CommIT*. Vol 3(1):6-11.
[12] Doreswamy & Hemanth, K. S., 2011. Performance Evaluation of Predictive Classifiers For Knowledge Discovery From Engineering Materials Data Sets. *CoRR*. Vol 3(3):1209-2501.