Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Evaluation of feature selection using information gain and gain ratio on bank marketing classification using naïve bayes

To cite this article: B Prasetiyo *et al* 2021 *J. Phys.: Conf. Ser.* **1918** 042153

View the article online for updates and enhancements.

IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 182.1.83.53 on 28/06/2021 at 08:44

Evaluation of feature selection using information gain and gain ratio on bank marketing classification using naïve bayes

B Prasetiyo[1]*, Alamsyah[1], M A Muslim[1], N Baroroh[2]

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
[2]Department of Accounting, Faculty Economy, Universitas Negeri Semarang, Indonesia

*Corresponding author: bprasetiyo@mail.unnes.ac.id

Abstract. One of the efforts of banks to do marketing is by telephone to offer their products, such as deposits. There are many variables that influence whether the customer decides to subscribe or not. In this study, we present a comparison of feature selection from high features dataset. We use a bank marketing dataset which has 20 features and consists of 4,119 instances. We consider 2 ranking methods entropy-based, namely Information Gain (IG) and Gain Ratio (GR). In our experiment, we classified the various selected based on the ranking of the selected features using Naïve Bayes. We show that the selection of different features is important for classification accuracy. The different combinations of feature selection can affect the accuracy results.

1. Introduction

Strategies in facing global competition, need to be done by an organization. One of them is in the economic sector, for example banks. In today's technological advances, marketing strategies can be carried out by telemarketing. One of the telemarketing strategies is to do marketing without face to face by phone calls [1]. The success of telemarketing with data mining methods has been studied by [2]. Some of the uses of data mining in the economic sector include Credit Card Risk Prediction [3-5], to detect fraudulent in FinTech [6]; Credit card fraud detection using deep learning [7], and bankrupt prediction [8]. In bank marketing, they try to offer the product to customers in direct marketing whether to subscribe or not. However, many attributes are involved. In data mining, we can classify based on existing datasets to create a model. However, the problem that often occurs in classification is when the dataset has too many attributes. While not all tributes are used because less relevant features can reduce algorithmic performance [9]. Feature selection is one step in pre-processing. Feature selection is done by selecting relevant features that can affect the classification results [10].

Feature selection is a scientific study in the field of computer science and has continued to be researched since the 1970s in statistical pattern recognition [11-12], machine learning and data mining [13-15]. The simplest feature selection can be done using an entropy-based method, including information gain and gain ratio. Several studies that have performed entropy-based feature selection include Rathore on text classification problem using a gain ratio [16], then studies related to information gain ratio are [17-18]. In this paper, we will discuss the classification using bank marketing with a different number of features based on feature selection ranking with a different number of features.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Methods

We proposed diverse feature ranking and feature selection techniques in this study. These strategies are used to delete irrelevant or redundant features from a function vector. In our experiment, first we do the pre-processing step. This step takes the input dataset to perform cleaning tasks. Then we normalize the data using standardization. Thus, the data used is Bank Marketing dataset that contains 4,119 examples and 20 attributes [2].

In this paper, we consider evaluation of the practical usefulness commonly used: entropy-based with two domain i.e: Information Gain (IG) attribute evaluation and Gain Ratio (GR) attribute evaluation. We use entropy that is the commonly used in the information theory measure [19], which characterizes the purity of an arbitrary set of examples, usually uses entropy. The formula of IG given by equation 1.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{1}$$

The Gain Ratio (GR) is given by equation 2.

$$GR = \frac{IG}{H(X)} \tag{2}$$

After we implemented the feature selection, then we use diverse feature ranking to classify using the Naïve Bayes algorithm [20]. To matrices the performances, we use 10-cross validation. Then we calculate the accuracy metrics then analyze.

3. Results and Discussion

The Bank Marketing data related to a Portuguese banking institution's direct marketing strategies using on phone calls in deposit subscription [2]. The number of examples is 4,119 instances, the number of feature 20, and there are no missing values. The result implementation of IG and GR is shown in Table 1.

Table 1. Results of ranking feature Bank Marketing dataset

Features	IG	GR	IG rank	GR Rank
duration	0.083	0.066	1	3
nr.employed	0.054	0.044	2	4
euribor3m	0.05	0.042	3	5
poutcome	0.047	0.03	4	1
month	0.036	0.025	5	8
previous	0.032	0.017	6	2
emp.var.rate	0.026	0.016	7	6
contact	0.015	0.013	8	7
pdays	0.013	0.007	9	9
cons.price.idx	0.013	0.007	10	11
job	0.011	0.007	11	13
cons.conf.idx	0.01	0.005	12	12
default	0.005	0.004	13	10
age	0.005	0.002	14	14
education	0.004	0.002	15	16
campaign	0.004	0.002	16	15
marital	0.002	0.001	17	17
loan	0,000	0,000	18	18
housing	0,000	0,000	19	19
day_of_week	0,000	0,000	20	20

The table provides three columns – in which the first column provides the name of features; the second column presents the rank feature of IG and the last column lists the rank features of GR. We can show that the first best rank in IG is “duration”, while the first rank in GR is “poutcome”. The last three columns has yero value of IG and GR, so we not use this feature. Then we experiment using the diverse feature rank (based on Table 1) and the diverse number of features. We use the Naïve Bayes for classification [20] and 10-cross validation to matrices performa accuracy. The result of our study is shown in Figure 1 and Table 2.

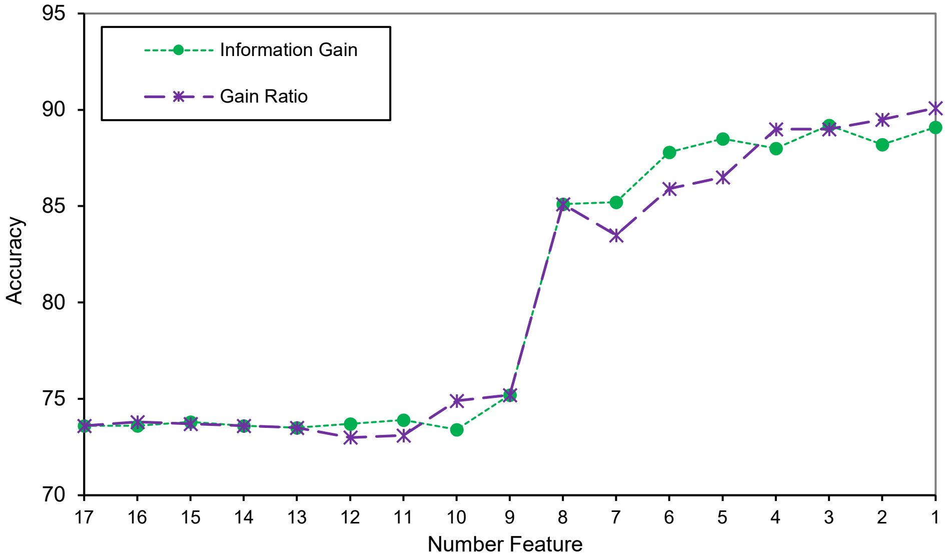

Figure 1. Ranking methods and classification accuracy for Bank Marketing dataset using Naïve Bayes

Table 2. Detail of accuracy performance

Selected feature		17	16	15	14	13	12	11	10	…
Accuracy	IG	73.6	73.6	73.8	73.6	73.5	73.7	73.9	73.4	…
(%)	GR	73.6	73.8	73.7	73.6	73.5	73.0	73.1	74.9	…

Selected feature		…	9	8	7	6	5	4	3	2	1
Accuracy	IG	…	75.2	85.1	85.2	87.8	88.5	88.0	89.2	88.2	89.1
(%)	GR	….	75.2	85.1	83.5	85.9	86.5	89.0	89.0	89.5	90.1

We remove the feature that contain zero value of IG and GR (Table 1). Then we classify strat from 17 selected features untul 1 selected features. We can show, that the diverse of selected features from features rank affect the difference result of perfromances. The Figure 1 provides comparison of features selection using IG, GR and the accuracy. The detailed performance of our study shown in Table 1. We can show that the low of IG or GR index result the low performance (between 70-75%), but the performance raise significance start from 8 feature selected (between 80-90%).

4. Conclusion

Feature selection has important step in classification accuracy performance. We presented the two ranking method features selections, using IG and GR. Ranking methods can be used to reduce dimensionality of the feature space. The diverse rank feature and the selected feature can affect the accuracy performance. We can show that the performance raises significance start from 8 selected feature according.

References

[1] Kotler P, Keller KL 2012 *Framework for Marketing Management, 5th edition* (New York: Pearson)
[2] Moro S, Cortez P and Rita P 2014 *Decis. Support Syst.* **62** 22
[3] Muslim M A, Nurzahputra A and Prasetiyo B 2018 *Int. Conf. on Inf. and Comm. Tech (ICOIACT)* p 141
[4] Kiruthika U, Raja S K S, Raman C J and Balaji, V 2020 *IEEE 2nd Int. Conf. Power Energy Control Transm. Syst. Proc.* p 1
[5] Sahu A, Gm H and Gourisaria M K 2020 *2020 IEEE 17th India Counc. Int. Conf. INDICON*p 1
[6] Abdulsattar K and Hammad M 2020 *Int. Conf. Innov. Intell. Inform. Comput. Technol. 2020* p 6
[7] Babu, A M and Pratap A 2020 *2020 IEEE Recent Adv. Intell. Comput. Syst. RAICS* p 32
[8] Prasetiyo B, Muslim M A and Baroroh N 2020 *J. Phys.: Conf. Ser.* **1567** 1
[9] Xue B, Zhang M, and Browne, W N 2012 *IEEE Trans. Cybern.* **43** 1656
[10] Gheyas I A and Smith L S 2010 *Pattern Recognit.* **43** 5
[11] Ben-Bassat M 1982 *Pattern recognition and reduction of dimensionality. Handbook of Statistics* 2 p 773
[12] Siedlecki, W and Sklansky J *Int. J. Pattern Recognit. Artif. Intell.* **2** 197
[14] Haghighi, M S and Hoseini M J M 2020 *6th Iran. Conf. Signal Process. Intell. Syst. (ICSPIS) 2020* p 1
[15] Nafis N S M and Awang S 2020 *Iraqi Journal of Science* **61** 3397
[16] Rathore M S, Saurabh P, Prasad R AND Mewada P 2020 *Progress in Computing, Analytics and Networking* (Singapore: Springer) p 23
[17] Ren W, Qiu Y and Li X 2018 *Proc. 2018 Int. Conf. Algorithms Comput. Artif. Intell.* p 1
[18] Singer G, Anuar R and Ben-Gal, I 2020 *Expert Syst. Appl.* **152** 113375
[19] Abe N and Kudo M 2005 *Lect. Notes Comput. Sci.* **3684** 689
[20] Gorunescu F 2011 *Data Mining Concept Model and Techniques* (Berlin: Springer)