Optimizing Naive Bayes Probability Estimation in Customer Analysis Using Hybrid Variable Selection

R. Siva Subramanian and D. Prabha

Abstract Customer study is considered as an important business plan to improve the enterprise's goal. The purpose of customer analysis is to understand the potential customer within the enterprises and their organizational needs and how well the customers are pleased with the company service. To perform better customer analysis, the need for CRM is studied. But the customer data generated are in large dimensional which possibly holds correlated and uncertainties variables in the dataset. To perform better analyzes with these customer data, NB an ML model is applied. But the violation of NB assumption proposed toward variables causes NB to work shoddily. To improve customer analysis using the NB, the variable selection mechanism is proposed. The proposed hybrid mechanism is based upon the filter and the wrapper mechanism. The hybrid mechanism comprises of two phases—first using the ReliefF filter approach, the customer data are processed and ranked attribute subset is generated. Then using threshold value, best attribute set is obtained from the scored attribute subset. Then the preselected variable set is processed using SFS and genetic wrapper approaches individually to get the best optimal variable subset. Further, the variable set acquired using the proposed technique is analyzed with the NB model and performance is computed. The performance hybrid-NB is compared using the filter-NB, wrapper-NB and NB without using any variable selection mechanism. The results present proposed hybrid work better to get the best variable subset and also increase the performance of the NB classifier. Compare to the wrapper approach, the proposed hybrid approach exits less computational time.

Keywords Naive Bayes (NB) · Variable selection · Hybrid approach · Customer analysis · CRM · Prediction

R. Siva Subramanian (✉)
Anna University, Research Scholar, Chennai, India

D. Prabha
Associate Professor, Dept of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore-641008, Tamil Nadu, India
e-mail: prabha@skcet.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 58,
https://doi.org/10.1007/978-981-15-9647-6_47

1 Introduction

In the fast commercial business world, the need for holding back the existing customers and developing the business strategies to satisfy consumers is considered paramount. Most enterprises give importance to the acquisition of new consumers, but the priority to existing customers is not considered. This leads to the abandonment of existing consumers with the enterprises and also makes the enterprises business degrades. To analyze the value of the existing consumers and to enhance the new customers with the enterprises and to improve the business process, the impact of CRM is analyzed. Customer relationship management (CRM) involves people, technology and the process to make better efficient analyzes of a customer with the enterprises. CRM, an integrated procedure, intelligently captures the customer interaction and makes use of consumers instances to obtain better insight into customer patterns [1]. CRM helps to build strong customer relationships by understanding the consumer's needs and create a path to develop new customers and developing consumers services. The need for CRM is motivated due to 1.Increase profitability, 2. Increase productivity, 3. Enhanced customer service, 3. Quick access to consumer data, 4. Identify potential customers, 5. Easily to monitor the customers, 6. The better reporting process, 7. Increase referrals, 8. Enhance services and products, 9. Minimize the costs, 10. Create new customers. The analytics process in the CRM (called CRM analytics) uses different programming modules to analyze the customer data and helps to develop business decisions [2]. CRM analytics is applied for different purposes like 1. Analysis of profitability, 2. Customer segmentation, 3. Analysis of customer value, 4. Customer personalization and 5. Predictive modeling is performed accordingly to the need of the enterprises. But in the CRM, customer interaction collected is from different department of the enterprises and stored in one roof. This has the chance of generating large-dimensional data about the customer and also has the chance of storing redundant, missing, noisy and irrelevant variables in data. To perform better customer analyzes the use of NB, an ML model is considered in this research. NB a simple and efficient model in performing better analyzes of customer data [3]. But the availability of correlated and uncertainties variables in the dataset causes the NB to execute shoddily. This happens due to the NB assumption imposed on the dataset. NB proposes two important assumptions on the dataset. One is conditional independence among the input predictors—that is input predictors present in the dataset should be independent of each other; the other one is all input predictors present in learning set should be considered as equal. But in the case of real-time customer dataset, the dataset generated contains the correlated variables, which is a violation of NB assumption. Further, using these customer datasets without eliminating the correlated variables makes the NB model perform worsely. By removing uncertainties present in a dataset like missing, noisy and irrelevant variables should be processed; otherwise, poor performance prediction can be witnessed [4]. To overcome the above problem, use of variable selection approach is suggested. The variable selection approach selects the best variable subset by eliminating the correlated and uncertainties variables in the dataset and to maximize the classifier performance. In

general, filter and wrapper, variable mechanism are widely applied. But in the two approaches, there exit some disadvantages linked with the variable selection mechanism. That is in filter approach, the results generated are not satisfied, but the approach works fast in selecting the variable subset [5]. But in the filter technique, need of getting the effective threshold value is an important one to choose the best variable subset. In the wrapper mechanism, the results generated are best compare with filter, but the approach exhibits high computational time [6]. This brings backlog to the wrapper methodology. Considering the drawbacks of filter and wrapper approach, a new hybrid variable selection mechanism is proposed. The hybrid procedure is developed by considering the advantage of filter and wrapper approach. The hybrid approach comprises of two phases; first using the ReliefF method to reduce the variable set; second, using two wrapper approaches genetic and SFS to achieve best variable subset. The suggested technique shows the best variable subset is generated and also using the attribute subset obtained are further applied to enhance the NB prediction. The experiment is performed in three various perspectives; one is performing NB using a variable subset obtained using the filter method; second is performing NB using variable subset obtained using wrapper the method; third is performing NB using variable subset obtained from the proposed hybrid method. The empirical results obtained are compared using different validity scores between filter-NB, wrapper-NB and hybrid-NB. The experimental outcome proves suggested hybrid technique selects the best variable to improve the NB model compare to other filter and wrapper-NB approach. The research is followed by a literature survey, Naive Bayes, feature selection, methodology, experiment, results, conclusion.

2 Literature Survey

Aliezanejad et al. [5] The research aims to identify the importance of the optimal variable subset in gene selection. In gene selection datasets, number of instances is small and number of attributes is high. Using these datasets, optimal performance is not obtained and also the complexity of the model increases. To sort out an effective attribute subset, two filter approaches are proposed in this research. One is Xvariance and the second one is mutual congestion. The proposed mechanism is applied to eight medical datasets. The empirical research output shows Xvariance compute better with standard datasets and the mutual congestion gets better prediction in large-dimensional datasets [7]. The research intends to study the importance of obtaining the right prediction to carry out better decision making in fields like medical, engineering, finance, environmental studies and emerging technologies. To perform better prediction using the classifier, the need of selecting the best attribute subset is important and also removing the uncertainties in the dataset should be performed. For that, the research uses a variable selection mechanism. The research proposes a variable mechanism called FSULR. The suggested approach works better to get the optimal variable subset and significantly improves model accuracy. The FSULR approach is compared using the seven exiting variable selection mechanism.

The variable subset obtained is examined using NB, IB1 and J48 classifiers. The results present the proposed mechanism FSULR works better with the classifiers to get better prediction [8]. The author analyzes the concern of increasing documents on the Internet and the problem of how to effectively retrieval the text information from the large documents. To carry out better analyzes, ML models are applied to the text documents. But with the large features in the dataset, to perform better text documents, the need for variable selection mechanism is important. For that author suggests a variable selection mechanism called best terms. The BT approach is evaluated using the Reuters-21578 and 20 newgroup document datasets. The variable subset generated from the BT methodology is evaluated using NB and SVM models. The results present BT improves the prediction in both NB and SVM when compare to other variable selection methods [9]. The author performs the study on selecting the best variable subset from the cancer datasets to predicate which type of cancer belongs. For that study, the author implements an enhanced JNMI variable selection based upon the filter approach. The proposed approach is tested using seven cancer datasets. The attribute subset generated from the JNMI approach is further examined using five ML models. The results present the proposed JNMI approach get better prediction to compare to IG, GR and SU variable selection mechanism [10]. The author analyzes the use of redundant and irrelevant variables in the learning data make the NB model achieve shoddily in the prediction. To solve the issue, the author suggests an approach called BHFS. The BHFS approach chooses the best variable set in the leaning data by expelling the redundant and insignificant attributes in the datasets. The approach uses an ensemble method to get the optimal stable attribute set. In the BHFS approach, variable selection mechanisms like chi-square, GR, ReliefF and SU are applied. The variable subset obtained is further examined using NB. The results present the BHFS approach work better to generate better a variable subset and also the approach requires only minimum running time [11]. The author addresses the problem with large-dimensional variables in the bioinformatics datasets. To remove the irrelevant and redundant features in the datasets, the author applies the variable selection mechanism. The author uses a wrapper learner mechanism to get the variable subset and which uses three wrapper learner (5-NN, LR and NB) models to figure the best one. The procedure is tested using nine bioinformatics datasets. The variable subset obtained from wrapper method is further examined using classification learner like 5-NN, LR and NB to compare the performance prediction [12]. The author identity that due to correlations among the features in dataset makes the classifier to perform poor prediction. To maximize the classifier, variable selection technique is applied to remove the variables which hold dependencies among other variables. In this study, two filters and two wrappers approaches are considered. The empirical procedure is conducted using Relief, CFS, NB-GA and NB-BOA. The results present Relief witness best prediction compared with other variable selection mechanisms [6]. In large variables datasets identification of best variable set using a variable selection, the mechanism is considered as an important one. For this, the research applies a variable selection mechanism in which three approaches are considered. The variable selection mechanisms used are chi-square, IG and BA. Further, the variable subset generated is examined using three

ML models like KNN, NB and DT. The procedure is experimented using fourteen datasets and results obtained show that BA approach trends to improve the classifier performance more [13]. The author addresses the issues concerned with the phishing detection. With the existence of large variables in phishing datasets, poor performance is witnessed with the classifier. The author explores the use of an attribute selection mechanism to get a good feature set from the phishing datasets. For that purpose, the CFS and wrapper mechanism is applied. In the wrapper, approach forward search and GA is applied. The procedure is experimented using phishing datasets which contains 177 initial variables. The variable obtained are further examined using NB, LR, and RF ML models [14]. The author addresses the problem with the intrusion in network security. With the large variable dataset in the IDS, the ML models applied exhibit more time-consuming in execution. A variable selection mechanism is applied to choose the best attribute set to improve the accuracy and performance of IDS. In variable selection, genetic wrapper approach is applied with the LR as the wrapper learner algorithm The approach is evaluated using UNSW-NB15 and KDD99 dataset and the attribute subset generated is further analyzed using C4.5, NBTree and RF. The output achieved from the suggested approach is compared using other variable selection to examine the efficiency of the proposed mechanism.

3 Naive Bayes

NB belongs to a supervised approach based on the theorem of Bayes. NB is a fast, reliable and accurate classifier applied in a wide range of real-world application datasets. The NB is a popular model in ML applications due to simplicity and computational efficiency. NB is widely applied due to 1. Small learning data, 2. Easy to build, 3. Simple computing, 4. Time efficiency, and 5. Handle large datasets [15, 16]. The problem of predicting the instance, $X = \{x_1, \ldots, x_n\}$, where n denotes the input predictors and the input predictors are independent of each other.

$$p(C_k|x_1, \ldots, x_n) \tag{1}$$

where C_k denotes the classes.

Bayes theorem is described as

$$p(C_k|X) = \frac{p(C_k)p(X|C_k)}{p(X)} \tag{2}$$

where

X denotes the instances, $p(X)$—probability X, $p(X|C_k)$—probability X$inthe$ hypothesis C_k,$p(C_k)$—prior probability, C_k—the hypothesis X.

Then the above Formula is described as

$$\text{posterior} = \frac{\text{prior} \times \text{ likehood}}{\text{evidence}} \tag{3}$$

The evidence for each class label is fixed for the one sample. The posterior value is compared with other posterior class value, to compute the class of the sample which is to be classified. Then Bayes formula is carried by describing $(C_k|x_1, \ldots, x_n)$ using the multiplication method,

$$p(C_k|x_1, \ldots, x_n) = \text{p(C}_\text{k}\text{)}\ \text{p}(x_1, \ldots, x_n|C_k) \tag{4}$$

$$= p(C_k)p(x_1|C_k)p(C_k|x_1 \ldots x_n|C_k, x_1) \tag{5}$$

$$= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)p(x_3.\ldots.x_n|C_k, x_1, x_2) \tag{6}$$

$$= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)p(x_3|C_k, x_1, x_2)p(x_4 \ldots x_n|C_k, x_1, x_2, x_3) \tag{7}$$

$$= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)p(x_3|C_k, x_1, x_2)...p(x_n|C_k, x_1, x_2, x_3, \ldots x_{3n-1}) \tag{8}$$

Equations 8 becomes more complex which will affect probability value and the computing becomes too complex to carry out. Then based on the NB independence assumption that each variable $(x_1, \ldots, x_n)$ is independent of each other. Using the NB assumptions [17, 18],

$$= \frac{p(x_i \cap x_j)}{p(x_j)} = \frac{p(x_i)p(x_j)}{p(x_j)} = p(x_i) \tag{9}$$

For $i \neq j$, so that

$$\text{Arg max} : p(x_i|C_k, x_j) = p(x_i|C_k) \tag{10}$$

From Eq. 10, it makes it easy to perform the calculation, by applying NB independence assumption

$$\text{Arg max} : p(C_k|x_1, \ldots x_n) = p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k) \tag{11}$$

$$= p(C_k)\prod_{i=1}^{n} p(x_n|C_k) \tag{12}$$

Then the Eq. 12 is used for the classification process. From the above process, the NB Eq. 12 is derived based upon independence assumption within the input predictors. Violation of NB assumption makes the classifier to witness poor perform

[4]. The datasets which hold large-dimensional variables will possibly hold uncertainties in datasets and these datasets must be preprocessed before modeling with the NB model. To remove the uncertainties in the dataset and to choose the best variable subset, the use of a variable selection mechanism is applied. By choosing the independent attributes using a variable selection, mechanism will help to satisfy the NB assumption and makes the NB perform efficiently in the customer analyzes datasets [10]. The researchers suggest a new hybrid variable selection mechanism which performs effectively to get the best variable subset to enhance the NB prediction and also the suggested methodology gets the best set of variables and works in better time complexity to compare to both filter and wrapper mechanism.

4 Feature Selection

In traditional days, the generation of customer interaction with the enterprises is less, and due to underdevelopment of the technology, the methodology to store and to access the customer data is less in volume. Due to back lack of technology, the access to the customer data is difficult one. But in the fast-evolving commercial business world, the data associated with the customers are generated and stored in large volumes. With these large-dimensional customer data, many uncertainties are associated. In the customer data, there is possible to retain correlated variables. The existence of uncertainties and correlated variables cannot be directly handled by the NB model. So to remove the problem associated with these customer data, a variable selection mechanism is applied. In ML, the variable selection also named as attribute or feature or subset selection. Variable selection approach is a process of choosing the independent variables from the dataset either manually or automatically to maximize the accuracy of the applied ML model [12, 19]. The objective of the variable technique is to decrease the feature space of the dataset by selecting the variable set that is highly associated with the output class and using the variable subset with the ML model to minimize the computational performance and to obtain efficient prediction results. Consider the customer dataset used in the research $D = \{y_1, \ldots y_n | C_k\}$ where $(y_1, \ldots y_n)$ denotes the input predictors and the C_k denoted the output class. The aim of feature selection to use some kind of evaluation or search approach to choosing the variable subset which makes to increase the prediction and also increase the model performance [13]. In the variable selection, there are four phrases involved namely 1. Generation of subset, 2. Subset evaluation, 3. Criteria for stopping, 4. Validation of results. The general approach of variable technique is described below.

4.1 Variable Selection Algorithm

Inputs: $Y = (y_1, \ldots, y_n)$ denotes input predictors, $SGO =$ Successor genarator operator, $EM =$ evaluation measure, $\emptyset =$ stopping criteria.

1. Initialize:
2. $y^{/} = \text{begin}(y)$
3. $y_{vs} = \{\text{best of } y^{/} \text{ using } EM\}$
4. Repeat
5. $y^{/} = \text{search strategy}\left(y^{/}, SGO(EM), y\right)$
6. $y_{vs} = \{\text{best of } y^{/} \text{using } EM\}$
7. if $EM\left(y^{/}\right) \geq EM(y_{vs}) or (EM\left(y^{/}\right) == EM(y_{vs}) \& \left|y^{/}\right| < |y_{vs}|$
8. $y_{vs} = y^{/}$
9. Until stop criteria is not found

Output: $y_{vs} =$ best variable subset is obtained

In the variable selection mechanism, the generation of a subset is a search approach which uses a search strategy. Then using the $EM =$ evaluation measure the generated variable subset is evaluated using the previously obtained variable subset. If the subset of the variables obtained is better than the previously obtained optimal variable subset, then the new subset variables are replaced with the subset of the existing variables. The approach continues until $\emptyset =$ stopping criteria is reached. Lastly, y_{vs} best attribute subset is obtained and the variables are validated. In general, variable mechanism is categorized into wrapper and filter based upon the practice [5]. The filter approach is completely based upon the characteristics of data and using some statistical measure to compute the worthiness of variables with an output label and the highly independent attributes are considered. In the filter approach, no interference of the ML model is involved to choose the variable subset. These characteristics make the filtering mechanism work fast in selecting the subset of the variables.

4.2 Filter Algorithm

Inputs: $D = \{y_1, \ldots y_n | C_k\}$ - Learning set with n variable subset and C_k denotes the output label, $y^{/} -$ inital variable set, $\emptyset =$ stopping criteria.

1. Begin
2. set $y_{vs} = y^{/}$
3. $\gamma_{vs} = \text{eval}\left(y^{/}, m\right)$ // $\left(\text{examine } y^{/} \text{using an indpendent measure}\right)$
4. do begin
5. $\delta =$ generate$(y_1, \ldots y_n)$
6. $\gamma = \text{eval}(\delta, m)$
7. If$(\gamma > \gamma_{vs})$

8. $\gamma_{vs} = \gamma$
9. $\gamma'_{vs} = \gamma_{vs}$
10. repeat($\emptyset$ = stopping criteria is not reached)
11. end
12. return γ'_{vs}; end

Output: γ'_{vs} optimal variable subset is generated using filter approach.

The general filter algorithm in described in Fig. 2. Consider the learning set $D = \{y_1, \ldots y_n | C_k\}$. Using independent measure examine each individual generated subset δ = generate($y_1, \ldots y_n$) and compared with the previously generated subset. Then the process continues until ($\emptyset$ = stopping criteria) is reached. Finally, γ'_{vs}optimal variable subset is obtained. In wrapper mechanism, the variable subset generated is based upon some search strategy and by using some induction algorithm [11–18, 20–22].

4.3 Wrapper Algorithm

Inputs: $D = \{y_1, \ldots y_n | C_k\}$ - Learning set with n variable subset and C_k denotes the output label, $y' -$ inital variable set, $\emptyset$ = stopping criteria.

1. Begin
2. set $y_{vs} = y'$
3. $\gamma_{vs} = \text{eval}(y', I)$ // (examine y' using any induction algorithm I)
4. do begin
5. $\delta = \text{generate}(y_1, \ldots y_n)$
6. $\gamma = \text{eval}(\delta, I)$
7. If($\gamma > \gamma_{vs}$)
8. $\gamma_{vs} = \gamma$
9. $\gamma'_{vs} = \gamma_{vs}$
10. repeat($\emptyset$ = stopping criteria is not reached)
11. end
12. return γ'_{vs}; end

.

Output: γ'_{vs}optimal variable subset is genearted using wrapper approach

The general wrapper algorithm is described in Fig. 3. The difference between the filter and wrapper is based upon the assessment criteria and use of the induction learner. Consider the learning set $D = \{y_1, \ldots y_n | C_k\}$. The wrapper mechanism applies some evaluation measure (induction algorithm I). to selects the optimal feature subset. Normally, best variable set is generated by the wrapper technique [14].

The variable set obtained using wrapper methods is optimal compare to the filter method. But in both filter and wrapper approaches, there are some drawbacks linked with it. In the filter approach, the methods work fast to select the variable subset, but the output obtained is not satisfactory, and also in the filter approach, they need a set of the cut-off value to select the optimal variable subset [9]. While considering the cut-off value, better care should be given. In the wrapper mechanism, there exhibits a high computational time. This happens due to the use of an induction algorithm and search strategy in the wrapper to select the best variable subset [13]. To overcome the above problem with the filter and wrapper, a hybrid variable selection mechanism is proposed. This method selects the best optimal variable subset with efficient time complexity to compare to filter and wrapper. The hybrid approach is proposed based upon the considering of both filter and wrapper mechanism. The hybrid approach consists of two stages. First, the customer dataset is examined using a filter approach and using the cut-off value optimal variable subset obtained. Further preselected variable subset obtained from the filter approach is processed using wrapper procedure to get optimal variable subset.

5 Hybrid Variable Selection Methodology

To generate the best variable subset with efficient time and to reduce the search strategy in the wrapper approach, the hybrid mechanism is proposed. The proposed structure hybrid approach is shown in Fig. 1. The proposed hybrid variable selection approach is based upon the filter and wrapper mechanism [23]. The proposed approach consists of two stages to select the optimal variable subset. In the first stage, the customer dataset is processed through a filter approach. In the filter approach ReliefF-based filter approach is applied to score the variables accordingly to relevance with the class label.

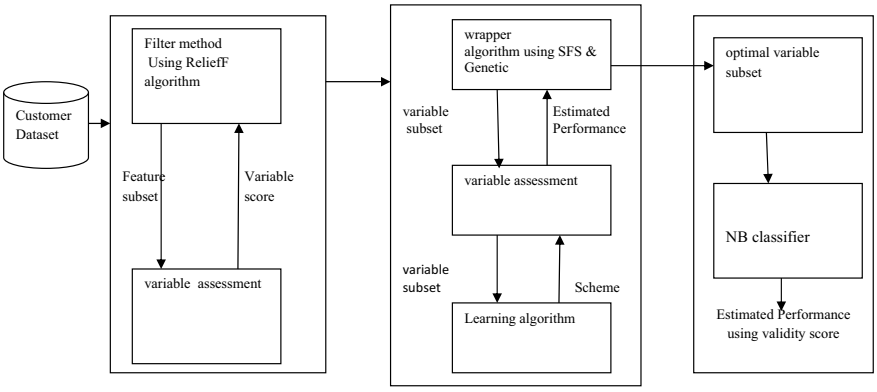

Fig. 1 Overall structure of the proposed hybrid approach

5.1 ReliefF Pseudocode

1. Assign an initial weight to all variables $(A) : W[A] = 0$
2. For i:=1 to m do begin
3. Randomly choose instance R_i
4. Find k nearest hits, H_j
5. For each class $C \neq \text{class}(R_i)$ do
6. Identify k nearest misses $M_j(C)$ using class C
7. For $X > 1$ to a do begin
8. $$W[A] := W[A] - \sum_{j=1}^{k} \frac{\text{diff}(A, R_i, H_j)}{m.k} + \sum_{c=\text{class}(R_i)} \frac{P(C)}{1 - P(\text{class}(R_i)} \sum_{j=1}^{k} \text{diff}\big(A, R_i, M_j(c)\big)/(m.k)$$
9. End

The ReliefF is proposed based upon the Relief algorithm and overcomes many disadvantages of the Relief approach [20]. The approach first initially assign zero scores to all the variables and randomly consider the instance R_i. Then the approach identifies the nearest H_j hits and misses $M_j(C)$ for the class C. Then using Eq. 8 as described, the weight of the variables is updated. Then the procedure is conducted for m times. Then from the ReliefF approach, scored variable set obtained. Next, to select the variable subset from the scored variable list, there needs a threshold value. Here, two threshold value is used to choose the variable subset. Then the variable subset generated is examined using the NB to check the performance prediction. Next, the preselected variable subset generated is processed through two different wrapper approaches. One is SFS and the second one is the genetic algorithm.

5.2 Sequential Forward Selection(SFS)

1. $V \leftarrow \{\emptyset\}$, repeat
2. for each $Y_j \notin V$;
3. $R_j \leftarrow R\big(V \cup \{Y_j\}\big)$;
4. Let $R^{/} \leftarrow \text{argmax}\{R_j\}$;
5. $V^{/} \leftarrow V \cup \{Y_j\}$
6. if $R\big(V^{/}\big) > R(V)$then
7. $V \leftarrow V^{/}$;
8. $R\big(V^{/}\big) \leftarrow R(V)$;
9. until $R\big(V^{/}\big) \leq R(V)\big|\big|V^{/}\big| == d)$ end

The SFS is based upon the wrapper approach and uses an induction algorithm to select the optimal variable subset [21]. The approach works by starting with the

empty set and uses a search strategy with a bottom-up procedure to add variables to set by using some evaluation function. The major disadvantage of SFS procedure does not eliminate the variables that are already included in the variable set.

5.3 Genetic Algorithm

1. Initialize population
2. While (Stopping criteria is not reached)
3. Selection
4. Reproduction
5. Replacement
6. Evaluate
7. End while

GA approach is one of the advanced heuristic algorithms in the variable selection to generate the best variable subset and rely upon the biological evolution and natural genetics. The algorithm works by starting with initialization, fitness assignment, selection, crossover, mutation and algorithm stops once the stopping criteria is reached [24]. Now, the variable subset generated from hybrid approach(ReliefF and SFS) and the hybrid approach(ReliefF and genetic) is examined individually with the NB classifier, to check how the performance is model is obtained through a proposed hybrid approach. Then performance prediction obtained using the filter-NB, wrapper-NB and proposed hybrid-NB is compared. The output obtained is compared using the validity scores.

6 Experimental Design

The procedure to improve the NB prediction in the customer analysis is performed in a detailed study and the procedure is presented in Chapter 5 (Hybrid variable selection methodology). The experiment is conducted using filter-NB, wrapper-NB, and lastly proposed hybrid-NB. Further, the experiment is performed using NB without using any variable selection.

6.1 Experimental Procedure

1. The customer dataset applied in this work is obtained from UCI and consists of 17 variable sets with 45,211 instances.
2. First, the NB model is constructed using a customer dataset without using any variable selection technique. Then next, using the ReliefF filter approach, the

variable subset is generated. By using ($\log_2$ n,& 65%), cut-off value variable subset is obtained from the scored variable list. Further, the NB is modeled variable subset obtained from the ReliefF approach.

3. Next, by using SFS and genetic wrapper approach, the variable subset is generated individually. Then NB is modeled using variable set generated from SFS with NB and variable set generated from genetic with NB
4. The variable subset generated from the proposed hybrid approach is modeled using NB. The hybrid approach involves two stages. In the first stage, ReliefF approach is used to score the variable accordingly to significance with the output class. The threshold value is applied to choose the best variable set(here 65% threshold value is considered). Then variable subset generated from the filter approach is processed through SFS and genetic wrapper procedure. Next, the variable subset obtained from SFS and genetic approach is individually examined using the NB classifier.
5. The experimental output is compared using NB without any variable selection approach, filter-NB, wrapper-NB, and lastly proposed hybrid-NB.
6. Using the validity scores, the experimental output is compared and the results are presented. The research applies different validity scores like accuracy, TPR, sensitivity, PPV, FNR, FPR[25].

6.2 Experimental Results

The results presented from Table 1 show the NB without using any variable selection approach obtains the accuracy of 88.0073%. But evaluation of NB using the variable subset obtained using ReliefF filter approach gets 88.94% accuracy. Here, ($\log_2 n$) threshold value is applied to get variable set.

The results presented from Table 2 show the NB without using any variable selection approach obtains the accuracy of 88.0073%. But evaluation of NB using

Table 1 Evaluation of NB using the variable subset obtained using ReliefF filter approach with ($\log_2 n$) threshold value

	Accuracy	Recall	Specificity	Precision	FNR	FPR
ReliefF and NB	88.9452	0.287	0.9693	0.553	0.713	0.030
NB without FS	88.0073	0.528	0.926	0.488	0.472	0.074

Table 2 Evaluation of NB using the variable subset obtained using ReliefF filter approach with (65%) threshold value

	Accuracy	Recall	Specificity	Precision	FNR	FPR
ReliefF and NB	89.5755	0.463	0.9530	0.567	0.537	0.0469
NB without FS	88.0073	0.528	0.926	0.488	0.472	0.074

Table 3 Evaluation of NB using the variable subset obtained using SFS and genetic wrapper approach

	Accuracy	Recall	Specificity	Precision	FNR	FPR
Genetic and NB	90.0135	0.414	0.964	0.607	0.586	0.035
SFS and NB	89.852	0.458	0.956	0.548	0.542	0.043
NB without FS	88.0073	0.528	0.926	0.488	0.472	0.074

Table 4 Evaluation of NB using the variable subset obtained using a hybrid wrapper approach

	Accuracy	Recall	Specificity	Precision	FNR	FPR
Genetic and ReliefF	90.0356	0.427	0.9630	0.605	0.573	0.036
SFS and ReliefF	89.916	0.412	0.9637	0.601	0.588	0.036
NB without FS	88.0073	0.528	0.926	0.488	0.472	0.074

the variable subset obtained using ReliefF filter approach gets 89.5755% accuracy. Here, (65%) threshold value is applied to get a variable set.

The results presented from Table 3 show the NB without using any variable selection approach obtains the accuracy of 88.0073%. But evaluation of NB using the variable subset obtained using SFS and genetic wrapper approach gets 89.85 and 90.01%, respectively. Compare to wrapper approaches, genetic and NB perform efficiently.

The results presented from Table 4 shows the NB without using any variable selection approach obtains the accuracy of 88.0073%. But evaluation of NB using the variable subset obtained using SFS and ReliefF hybrid approach gets 89.916% and genetic and ReliefF hybrid approach gets 89.916% respectively. Compare to hybrid approaches, genetic and NB performs efficiently.

The result presented in Table 5 show overall accuracy comparison obtained from the filter-NB, wrapper-NB and the hybrid-NB approaches. The results conclude the proposed hybrid approach genetic and ReliefF gets better results to compare to other approaches.

Table 5 Comparison of accuracy obtained from different approaches and proposed hybrid approach

	ReliefF and NB ($\log_2 \boldsymbol{n}$)	ReliefF and NB (65%)	Genetic and NB	SFS and NB	Genetic and ReliefF	SFS and ReliefF	NB without FS
Accuracy	88.9452	89.5755	90.0135	89.852	90.0356	89.916	88.0073

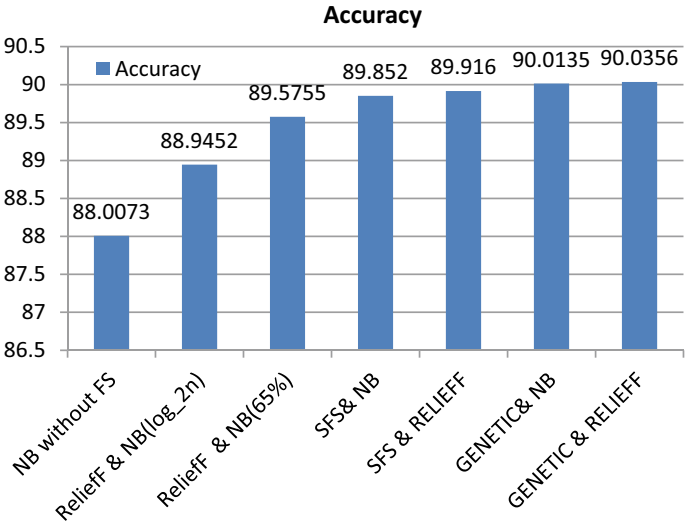

Fig. 2 Comparison of accuracy obtained from different approaches and proposed hybrid approach

6.3 Result Analysis and Discussion

The results obtained are presented in Tables 1, 2, 3, 4. Tables 1 and 2 represent the output obtained using ReliefF and NB with ($\log_2 n$, and 65%) cut-off value. The experiment shows the filter approach gets high accuracy of 89.5755% using 65% threshold value. Table 2 represents the output generated using the wrapper and NB procedure. In the wrapper, two different approaches are applied. One is SFS and the other is the genetic algorithm. The SFS with NB gets an accuracy of 89.852% and the genetic approach with NB gets an accuracy of 90.0135%. Compare to both approaches, genetic with NB performs better. Table 4 represents the hybrid approach with the NB. In hybrid, two approaches are proposed. one is genetic and ReliefF with NB and the other is SFS and ReliefF. The genetic and ReliefF with NB gets 90.0356% accuracy and the SFS and ReliefF with NB gets 89.916% accuracy. Compare to both hybrid approaches, genetic and ReliefF with NB performs better. Experiments using NB without FS obtain 88.0073% accuracy. Analyzes of customer datasets using (NB without FS, filter-NB, wrapper-NB,and lastly proposed Hybrid-NB) different approaches are performed and the results are presents. The empirical output represents that compare to all approaches, the proposed hybrid approaches get high accuracy and performs better. From the results, the research clearly makes to understand that:

1. The use of filter is simple and fast. But the results obtained are not satisfied. Then the use of threshold value must be given more importance. Since results are changed accordingly to a threshold value.

2. The wrapper approaches get optimal variable subset, but the approaches take more time to obtain the variable subset. This increases the complexity of the approaches. In the wrapper approach, NB is applied as an induction algorithm. However, different induction algorithm can be applied to check how the variable subset selected varies accordingly to the induction algorithm applied.
3. The proposed hybrid approach gets the best optimal variable subset with compare to filter and wrapper. The proposed approach is simple and fast as compare to filter and also gets better variable subset and time efficiency to compare to the wrapper approach. This research comes to a discussion that the proposed hybrid works well and improves the NB prediction in the customer dataset used in the experiment.

6.4 Conclusion

The research analyzes how customer prediction can be improved using the NB classifier. Due to the availability of redundant, missing, and insignificant variables in the dataset and the violation of NB assumption makes the NB witness poor prediction in performance. To overcome the problem, hybrid variable selection is proposed. The hybrid approach is based upon both filter and wrapper which consider the advantages of both approaches. The proposed technique comprises of two phases approach. In phase one, using ReliefF filter, the customer dataset is analyzed and ranked accordingly to relevance to the class label. Then using threshold value, optimal variable sets are obtained. The next phase uses the preselected variable set to process through SFS and genetic wrapper approach individually to obtain the best optimal variable subset. The experiment is conducted using filter-NB, wrapper-NB, and lastly proposed hybrid-NB. Further, the experiment is performed using NB without using any variable selection. The results reveal a hybrid approach selects the best attribute set and also improves the NB classifier better to other approaches. The hybrid approach is efficient in time and performs fast to select the variable subset. The work can be extended by examining using other different combinations of filter and wrapper approaches. In the wrapper approach, different induction learners can be applied to select the variable subset.

References

1. Chen IJ, Popovich K (2003) Understanding customer relationship management (CRM): people process and technology. Business Process Manage J 9(5):672–688. https://doi.org/10.1108/14637150310496758
2. Prabha D, Subramanian R (2017) A survey on customer relationship management. 1–5. https://doi.org/10.1109/ICACCS.2017.8014601
3. Pouria K, Sunita D (2017) Short survey on naive bayes algorithm. Int J Adv Res Comput Sci Manage 04

4. Sona T, Musa M, Bagirov AM (2011) Improving Naive Bayes classifier using conditional probabilities. In: 9th Australasian data mining conference. vol 121. pp 63–68
5. Aliezanejad M, Enayatifar R, Motameni H, Nematzadeh H (2019) Heuristic filter feature selection methods for medical datasets. Genomics. https://doi.org/10.1016/j.ygeno.2019.07.002
6. Rozlini M, Yusof MM, Noorhaniza W (2018) A comparative study of feature selection techniques for bat algorithm in various applications. In: MATEC Web of Conferences. 150:06006. https://doi.org/10.1051/matecconf/201815006006
7. Singh DAAG, Balamurugan SAA, Leavline EJ (2015) An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers. Int J Autom Comput 12(5):511–517
8. Dimitris F, Dimitris M, Spiridon L (2005) Best terms: an efficient feature-selection algorithm for text categorization. Knowl Inf Syst 8:16–33. https://doi.org/10.1007/s10115-004-0177-2
9. Dilwar M, Ramachandran V (2019) An enhanced feature selection filter for classification of microarray cancer data. ETRI J. 41. https://doi.org/10.4218/etrij.2018-0522
10. Siva Subramanian R, Prabha D (2020) Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-01961-9
11. Randall W, Taghi K, Napolitano A (2014) Optimizing wrapper-based feature selection for use on bioinformatics data. IN: Proceedings of the 27th ınternational florida artificial ıntelligence research society conference, FLAIRS 2014. pp 288–293
12. Yang Q, Salehi E, Gras R (2010) Using feature selection approaches to find the dependent features. In: Rutkowski L, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM (eds) Artificial ıntelligence and soft computing. ICAISC 2010. Lecture notes in computer science, vol 6113. Springer, Berlin, Heidelberg
13. Basnet RB, Sung AH, Liu Q (2012) Feature selection for ımproved phishing detection. In: Jiang H, Ding W, Ali M, Wu X (eds) Advanced research in applied artificial ıntelligence. IEA/AIE 2012. Lecture Notes in computer science, vol 7345, Springer, Berlin, Heidelberg
14. Chaouki K, Saoussen K (2017) A GA-LR wrapper approach for feature selection in network ıntrusion detection. Comput Secur 70. https://doi.org/10.1016/j.cose.2017.06.005
15. Aji W, Ahmad K, Della M, Risky A, Sandika P, Sulton K, Youngga N (2019) Naïve bayes classifier for journal quartile classification. Int J Recent Contrib Eng Sci IT (iJES). 7:91 https://doi.org/10.3991/ijes.v7i2.10659
16. Chen S, Webb GI, Liu L et al (2019) A novel selective naïve Bayes algorithm. Knowled-Based Syst 105361. https://doi.org/10.1016/j.knosys.2019.105361
17. Zhang H, Jiang L, Yu L (2020) Class-specific attribute value weighting for Naive Bayes. Inf Sci 508:260–274. https://doi.org/10.1016/j.ins.2019.08.071
18. Schneider KM (2005) In: Techniques for improving the performance of naive Bayes for text classification. pp 682–693
19. Xiaoping L, Yadi W, Rubén R (2020) A survey on sparse learning models for feature selection. IEEE Trans Cybernet. 1–19. https://doi.org/10.1109/TCYB.2020.2982445
20. Yang F, Cheng W, Dou R, Zhou N (2011) An improved feature selection approach based on ReliefF and Mutual Information. In: International conference on information science and technology
21. Marcano-Cedeno A, Quintanilla-Dominguez J, Cortina-Januchs MG, Andina D (2010) Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: IECON 2010-36th annual conference on ıeee ındustrial electronics society
22. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M (2019) Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 106839. https://doi.org/10.1016/j.csda.2019.106839
23. Wei G, Zhao J, Feng Y, He A, Yu J (2020) A novel hybrid feature selection method based on dynamic feature importance. Appl Soft Comput 106337. https://doi.org/10.1016/j.asoc.2020.106337

24. Khaled S, Mohamed A-N, Pierre T, Chelouah R (2013) Immune genetic algorithm for scheduling service workflows with QoS constraints in cloud computing. South African J Indus Eng 24:68–82
25. Prabha D, Ilango, K (2013) Customer behavior analysis using rough set approach. J Theo Appl Electron Commerce Res 8:21–33. https://doi.org/10.4067/S0718-18762013000200003