A Feature Selection Method Based on Information Gain and BP Neural Network

Xingyun Wang, Min Zuo and Lihui Song

Abstract Data mining and machine learning fields are facing with a great challenge of mass data with high dimensionality. Feature selection can contribute a lot to address this issue with the concept of reducing the number of features by eliminating the redundant and irrelevant ones while preserving the information of original features maximally. This paper analyzes and compares two common feature selection methods, then puts forward a novel method for feature selection based on information gain and BP neural network (IGBP). The experimental result shows that IGBP method can reduce the time cost and improve the accuracy of the model at the meantime. The scientificity and superiority of IGBP are demonstrated in this paper, making it an efficient approach to deal with high-dimensional data.

Keywords Feature selection · Data mining · BP neural network · Information gain · IGBP method

1 Introduction

In recent years, data mining theories and methods have been constantly applied to practical fields, revealing a great value and potential in all walks of life [1, 2]. However in internet era, data mining task becomes extremely difficult even impossible to accomplish due to the rapid increase in data size and dimensionality and the restriction of computing capability on the contrary. Under this circumstance, data mining technology confronts a test of both time cost and accuracy.

X. Wang · M. Zuo (✉)
School of Computer and Information Engineering, Beijing Technology
and Business University, Beijing 100048, China
e-mail: zuomin1234@163.com

L. Song
College of Mechanical and Electrical Engineering,
Yanching Institute of Technology, Sanhe 065201, Hebei, China
e-mail: 651265386@qq.com

Y. Jia et al. (eds.), *Proceedings of 2017 Chinese Intelligent Systems Conference*, Lecture Notes in Electrical Engineering 459,
https://doi.org/10.1007/978-981-10-6496-8_3

As a key step of the data mining technology, feature selection has crucial impacts on training time and training results [3]. In real-world applications, the dataset of high dimensionality contains numerous features which could be irrelevant to the output classification and could have correlative dependence on each other leading to a redundancy. If all these features are adopted to conduct the data mining process, the redundant part will weaken the model's interpreting ability while the irrelevant part will result in a high time consumption [4]. Therefore, the central research of this paper is to find the optimal subset of features by using a feature selection method which is effective and efficient. And once the subset is clarified, we can get the consequence as accurately as possible in limited time without reducing the size of dataset.

2 Related Work

Feature selection is used to eliminate some irrelevant and (or) redundant features from a large number of them, making the chosen subset reserves the characteristics of the original set as much as possible [5–7]. Feature selection is helpful to reduce the dimensionality of data and remain a high relevance between features and the output classification, shorten the training time and improve learning performance.

2.1 Feature Selection Method

Feature selection methods are used to assess the subset of features with some criterions, determining the optimal subset directly. And the feature selection method can be categorized into two types: filter and wrapper. The filter method is independent of the selected classifier as a pre-processing step which evaluates each feature based on its inner effect on the output classification. The wrapper method uses a single learner as a black box to evaluate the subset of features according to their predictive performance [8].

2.2 Information Gain

Entropy as a wildly used measure standard in information theory can be taken when apply the filter method to evaluate features [9]. A high level of entropy means the feature contains more unpredictability than a lower one. Whereas, a low level of entropy stands for more precision of information.

The information gain represents the decrease of expected entropy of a feature when partition the dataset with it. In practice, features are ranked according to the information gain. The higher the feature scores in the ranking list, the less information remains after partition, and the more suitable of the feature for classification.

2.3 BP Neural Network

The feature selection method based on BP neural network is a typical one among numerous wrapper methods. BP neural network is capable of mapping any n-dimension to m-dimension [10]. It can classify inputs that have different values into different categories accurately after being training well enough. Moreover, BP neural network is suitable for both continuous and discretized features and has strong robustness on null values and even errors [11]. In the application of feature selection, BP neural network trains and builds network model through using different feature sets as the input, measuring the feature sets by the evaluation indicator of the model after building.

3 IGBP Method

This paper analyzed and compared the methods of filter and wrapper adequately, we found these two methods do not exist oppositely. The filter method proceeds prior to the data mining procedure while the wrapper method is conducted in the middle of data mining procedure. Additional, the methods of filter and wrapper have some peculiarities respectively. Applying the filter method based on information gain is convenient and requires small computational complexities. But the chosen subset of features may not be the optimum for a certain classifier. On the other hand, the wrapper method based on BP neural network has the advantage of accuracy and a weakness as well [12]. Theoretically, to validate all combinations of the feature sets requires the method of exhaustion. And it is impossible to complete the task in limited time when the number of the features is too large considering that the training speed will decrease sharply in every single procedure.

Considering that the method of filter is simple while the wrapper method is accurate, we can keep the advantages of both methods and weaken their disadvantages by using the two in a combination way. Therefore this paper proposed a novel method which combines the filter and wrapper methods. The combination method is based on information gain and BP neural network (IGBP). In detail, IGBP applies the filter method based on information gain to evaluate each feature firstly. Then it choses different feature sets according to the result from the first step to build BP neural network models. At last, the final evaluation of a certain feature set is judged by the performance of the model. IGBP method makes up the lack of using a single wrapper method and gets rid of the inaccuracy problem by using a single filter method.

4 Experiment

In this section, we conducted a data mining experiment to demonstrate the IGBP method. The dataset for experiments is extracted from CFDA (China Food and Drug Administration) Sampling and Monitoring System. This dataset contains 14,517 instances, recording food features and detection conclusions. There are nearly 70 features in the dataset, however most of them are not suitable for the data mining task. 20 features are chosen manually for the follow-up experiment. The detection conclusions are ‘qualified’ and ‘unqualified’ which correspond to the positive and negative class as the output classification of the BP neural network. Within the dataset, there are 10,000 instances with positive values and 4517 instances with negative values.

4.1 Pre-evaluation

After a pre-processing work, we imported the well-organized dataset into WEKA (Waikato Environment for Knowledge Analysis) and apply the InfoGainAttributes Eval algorithm to evaluate these 20 features. Table 1, 2 and 3 shows the result.

As the result shows, the initial 20 features have different influence degrees on the output classification. And they can be divided into 3 groups. Group A contains 6 features of which the score is greater than 0.05, representing that these features are strongly indicative of the output classification. The features in group A are {Food Category, Sampling Date, Sampled Co. Province, Production Co. Province, Production Date, Shelf Life}. Group B are {Sampling Package, Sample Form, Quality Rank, Annual Sales, Unit Price, Sampling Site, Food Function}. These 7 features score greater than 0.01 and less than 0.05 representing a less importance to the output classification. The features in group C hardly have any influence on the output classification. They score less than 0.01, including {Storage State, Sample Origin, Sample Package, Use of Food, Export or Not, Sampling Approach, Area Type}. The follow-up experiments are performed on the basis of the score result.

Table 1 The score result of features in group A

No.	Feature	Average merit	Average rank
1	Food Category	0.383	1
2	Sampling Date	0.171	2
3	Sampled Co. Province	0.129	3
4	Production Co. Province	0.125	4
5	Production Date	0.069	5
6	Shelf Life	0.068	6

Table 2 The score result of features in group B

No.	Feature	Average merit	Average rank
7	Sampling Package	0.026	7
8	Sample Form	0.025	8
9	Quality Rank	0.023	9
10	Annual Sales	0.023	10
11	Unit Price	0.022	11
12	Sampling Site	0.02	12
13	Food Function	0.019	13

Table 3 The score result of features in group C

No.	Feature	Average merit	Average rank
14	Storage State	0.007	14
15	Sample Origin	0.006	15
16	Sample Package	0.004	16
17	Use of Food	0.003	17
18	Export or Not	0.001	18
19	Sampling Approach	0	19
20	Area Type	0	20

4.2 Evaluation Model 1

We adopted total 20 features (20-subset) as the input set to train the network model (model 1) with the Multilayer Perceptron algorithm. The training process lasts 38 min and the result shows that model 1 has an accuracy of 90.5352% to classify instances into correct classifications.

The TPR (True Positive Rate) of model 1 is 0.981 with a 0.262 FPR (False Positive Rate). And its ROC (receiver operating characteristic) area is 0.9172 which represents an average performance of the model. The ROC curve of model 1 is shown in Fig. 1.

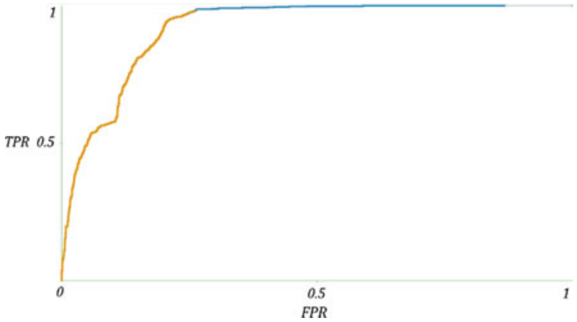

Fig. 1 The ROC curve of model 1

4.3 Evaluation Model 2

In the next stage, we removed the last 7 features in group C according to the score result and repeated the same training process above with the 13-subset of features to build the network model (model 2). The training process of model 2 lasts 30 min which is 21% time cost lower than model 1. Meanwhile, the accuracy of model 2 is 93.6006% which improves 3.1% compared with model 1.

Model 2 has a 0.964 TPR and a 0.126 FPR and its ROC area is 0.967 which is greater than model 1, namely the performance of model 2 improves significantly compared with model 1. Tues the chosen subset of features is more suitable for data mining tasks than the set contains the total 20 features. Figure 2 shows the ROC curve of model 2.

4.4 Evaluation Model 3

The last experiment chose 6 features (6-subset) in group A as the input to build the BP neural network (model 3). It takes 20 min and results in an accuracy of 89.2% which is the least in all these experiments.

Model 3 has the TPR of 0.921 and the FPR of 0.171 with the ROC area of 0.935. Figure 3 shows the ROC curve of model 3.

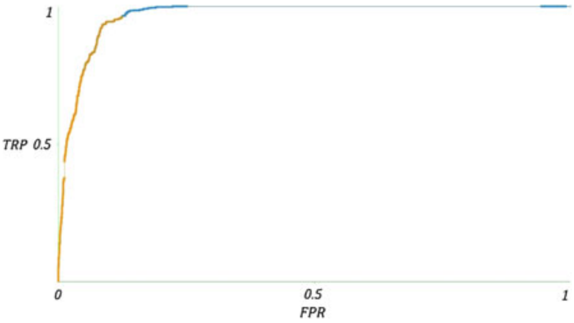

Fig. 2 The ROC curve of model 2

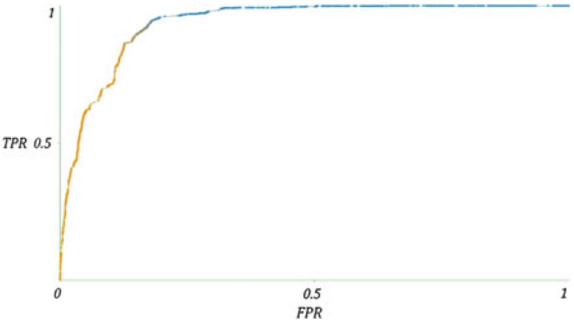

Fig. 3 The ROC curve of model 3

4.5 Discussion of Experimental Results

The summary of experimental results is given in Table 4.

According to the experimental results, the optimal subset of features is {Food Category, Sampling Date, Sampled Co. Province, Production Co. Province, Production Date, Shelf Life, Sampling Package, Sample Form, Quality Rank, Annual Sales, Unit Price, Sampling Site, Food Function}. When choosing the 13-subset of features as the input of BP neural network, it can result in a 93.6% accuracy and a satisfactory ROC area of 0.967. The model of 13-subset performs well enough to conduct subsequent data mining work.

As a typical case of the wrapper method, BP neural network requires a huge time cost in every training process due to its complicate structure. If we adopt BP neural network individually to verify every subset of features, the repeated training process will cost immeasurable time. Considering the 20-subset from the beginning, taking the method of exhaustion requires $2^{20} - 1$ times of training which is an impossible work in limited time. The IGBP method presented in this paper can find the optimum by eliminating features one after another according to the criterion of information gain. The worst case in the example above only takes 20 times of training, making the data mining task possible to accomplish.

On the other hand, IGBP method also avoids the issue of inaccuracy to adapt the filter method along. As the last experiment illustrates, if we chose the 6-subset in group A without any other evaluation to conduct the data mining process, the model will result in a low accuracy which means the 6-subset is not the optimum. Therefore IGBP method reserves the advantage of efficiency of the filter method and remedies its disadvantage of inaccuracy by combining the wrapper method.

Table 4 The summary of experimental results

Model	Indicator				
	Time cost (min)	Accuracy (%)	TPR	FPR	ROC
Model 1	38	90.5	0.981	0.262	0.917
Model 2	30	93.6	0.964	0.126	0.967
Model 3	20	89.2	9.921	0.171	0.935

5 Conclusion

Feature selection plays an important role in data mining and machine learning fields since it has great influences on both processes and results of the task. Choosing a proper subset of features to perform data mining could reduce the time cost and improve the performance of the model simultaneously. This paper proposed the IGBP method after making sufficient analyses on two types of feature selection methods. Through performing the evaluation processes repeatedly, the feasibility as well as the superiority of the IGBP method when facing the mass data have been proved, providing the subsequent data mining work with theoretical and practical basis.

Acknowledgements This work was supported by The National Key Technology R&D Program of China (2015BAK36B04).

References

1. Agarwal S. Data mining: data mining concepts and techniques. In: 2013 international conference on machine intelligence and research advancement (ICMIRA). IEEE; 2013. p. 203–7.
2. Almasoud AM, Al-Khalifa HS, Al-Salman A. Recent developments in data mining applications and techniques. In: 2015 tenth international conference on digital information management (ICDIM). IEEE; 2015. p. 36–42.
3. Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Springer Science & Business Media; 2012.
4. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(March):1157–82.
5. González A, Pérez R. Selection of relevant features in a fuzzy genetic learning algorithm. IEEE Trans Syst Man Cybern Part B (Cybern). 2001;31(3):417–25.
6. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res. 2004;5(October):1205–24.
7. Sheikhpour R, Sarram MA, Gharaghani S, et al. A survey on semi-supervised feature selection methods. Pattern Recogn. 2017;64:141–58.
8. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;97(1–2):273–324.
9. Wu G, Xu J. Optimized approach of feature selection based on information gain. In: 2015 international conference on computer science and mechanical automation (CSMA). IEEE; 2015. p. 157–61.
10. Najah A, El-Shafie A, Karim OA, et al. Application of artificial neural networks for water quality prediction. Neural Comput Appl. 2013;22(1):187–201.
11. Ennett CM, Frize M, Walker CR. Influence of missing values on artificial neural network performance. Stud Health Technol Inform. 2001;1:449–53.
12. Ding S, Li H, Su C, et al. Evolutionary artificial neural networks: a review. Artif Intell Rev. 2013;1–10.