Feature Selection Algorithm Based on Association Rules

Yi Qu¹, Yu Fang¹a and Fengqi Yan²

¹Department of Computer Science and Technology, Tongji University, 201804 Shanghai, China
²Heyeah Information Technology Co., Ltd., 201804 Shanghai, China

a Corresponding author: fangyu@tongji.edu.cn

Abstract. Feature selection is one of the key issues in pattern recognition. The quality of the feature selection has a direct impact on the classification accuracy and generalization performance of the classifier. In order to reduce the size of the feature subset and improve the efficiency of the algorithm without reducing the accuracy, this paper proposes a feature selection algorithm based on association rules, ARFS. The algorithm uses association rules to mine the frequent 2-items set of the feature attributes and category in the dataset. Then the algorithm sorts the features according to the confidence of the frequent 2-items set, and then combines the sequential forward selection method, and uses the classification performance of the decision tree classifier as the evaluation criteria of the feature subsets. In terms of experiments, this paper selects five public datasets in the UCI Machine Learning Repository to conducted three comparison experiments of feature subset selection, learning algorithm accuracy and runtime for four feature selection algorithms. Experimental results show that the ARFS is superior to the contrast method in the feature subset size and the accuracy. However, the ARFS is slightly inferior to the ReliefF algorithm in the runtime.

1. Introduction

Feature selection is a method of selecting a minimum feature subset that does not lose the original value of the data in a specific application domain. John[1] thought that feature selection is a process of reducing feature dimensions without reducing the accuracy. Dash[2] made a comprehensive review of feature selection and presented the basic framework of feature selection, as shown in Figure 1. The basic framework has three steps: generation of candidate feature subsets (search strategy), evaluation criteria, and stop criteria[3-4]. According to the formation process of feature subsets, the search strategy can be divided into the following three kinds: global search, random search, and heuristic search[5]. The advantage of global search is that the optimal feature subset can be obtained, but the time complexity is too large. The random search strategy is a combinatorial optimization problem, which combines feature selection with Simulated annealing algorithm, genetic algorithm[6], or random sampling process[7-8], and takes the probabilistic reasoning and sampling process as the basis of the algorithm. The heuristic search strategy is a practical method to solve the approximate optimal feature subsets. This method includes: single optimal feature combination, sequential forward selection (SFS), sequence backward selection (SBS) and so on. The evaluation criteria determine the direction of further search. Common evaluation methods include distance metrics, information metrics, dependency metrics and consistence metrics[2]. The stop criteria are related to search strategies, evaluation criteria, and specific applications. Common stop criteria include cycle times or execution
times, feature thresholds and evaluation function thresholds. According to the evaluation criteria, the feature selection methods can be divided into the following: embedded[9], filter[10], and wrapper[11].

Fig. 1. Basic framework for feature selection.

In the Embedded, the feature selection algorithm is embedded as part of the classification algorithm. The most typical algorithm is the decision tree, and the process of generating the decision tree is also the feature selection process. Classical decision tree algorithms include Quinlan's ID3, C4.5[12], Breiman's Classification and Regression Trees (CART)[13]. ID3 uses information gain as a measure of selection of attributes. Its disadvantage is the tendency to choose attributes with a large number of values. C4.5 uses gain rate to expand, but it tends to split unbalanced. CART uses the Gini indicator as an attribute selection metric, which is biased toward multi-valued attributes. It has difficulty when the number of classes is large. Although the decision tree method is flawed, it is often quite effective in practice.

The evaluation criteria of filter are obtained directly from the dataset and are independent of the classification algorithm. They have the advantages of strong versatility, low algorithm complexity, high operating efficiency, and suitability for large-scale datasets. However, it ignores the performance of the selected feature subset on the classification algorithm. The selected feature subset may not be optimal. In 1992, Kira and Rendell proposed an effective feature selection algorithm, the Relief[14]. The feature selection criterion of the algorithm is relevance corresponding to a feature. The biggest limitation of the Relief is that the redundant features cannot be identified in the relevant feature sets, and generally can only be applied to the binary class data. Kononenko[15] extended the original Relief in 1994 and obtained the ReliefF algorithm, which enables Relief to handle multi-category, incomplete, and noisy data.

The wrapper was first proposed by John in 1994[1], which used the performance of the classifier as an evaluation criterion for feature selection. The wrapper is more accurate than the filter, but less efficient. Hsu[16] used a genetic algorithm to find a set of feature subsets with the lowest classification error rate for decision trees. Chiang[17] combined Fisher discriminant analysis with genetic algorithm to identify key variables during chemical failures, and achieved good results. Guyon[18] proposed the SVM-RFE (recursive feature selection based on support vector machine) method that uses the SVM discriminant function to sort features, and then uses RFE to select the least feature subset that is retained in the feature ranking process. Finally, a classifier with higher accuracy is constructed.

In order to reduce the size of the feature subset and improve the efficiency of the feature selection algorithm without reducing the accuracy, this paper proposes the ARFS. The algorithm uses association rules to mine the frequent 2-items set of the feature attributes and category in the dataset. Then the algorithm sorts the features according to the confidence of the frequent 2-items set. In this step, a feature sequence with the best ability to distinguish between categories is obtained. And then the feature sequence of frequent 2-items set combines the sequential forward selection method, and uses the classification performance of the decision tree classifier as the evaluation criteria of the feature subsets to select the least feature subset without reducing the accuracy as possible. The experimental results of the UCI Machine Learning Database[19] dataset show that the ARFS in this paper is superior to the contrast feature selection method in the feature subset size and the accuracy, and ARFS also has good generalization performance. However, the ARFS is slightly inferior to the ReliefF in the efficiency.
2. Association Rules

Association rule is an important research direction of data mining. Association rule mining algorithm was proposed by Agrawal et al. [20] in 1993. In this paper, association rules specifically refer to the existence of implicit relations between mining feature attribute items and category items. Association rules are often applied in the areas of market analysis, economic forecasting and decision support. The main task is to discover concurrent data items and events. The ARFS uses association rules to find frequent 2-items set of feature attributes and categories, and then calculates their confidence for feature selection. The associated definitions of association rules and frequent 2-items set mining algorithms used in this paper are given below.

2.1 Item set

Association rule is a method for obtaining association information between data items. In the ARFS, \( I = \{i_1, i_2, \ldots, i_m\} \) is a collection of items. In \( I \), each \( i_k \) is called a project, the number of items \( k \) is the length of item set \( I \). Each subset in item set \( I \) is marked as \( T \), and all correspond to a unique identifier, denoted as \( TID \), the whole of \( T \) constitutes the dataset \( D \) of item set \( I \), and the value of \(|D|\) is equal to the number of \( T \).

2.2 Association rules

The association rule \( R \) is an implication:

\[
R : X \Rightarrow Y
\]  

\( X \subset I, Y \subset I \) and \( X \cap Y = \emptyset \), it means that item set \( X \) appears in a certain \( T \), and causes \( Y \) to appear in \( T \) with a certain probability. Among them, \( X \) and \( Y \) are called antecedent and consequent of association rules respectively.

2.3 Support

For item sets \( X, Y \), where \( X \subset I \) and \( Y \subset I \), Let count \( (X \cap Y) \) denote the ratio of the intersection of the feature subsets of \( X \) and \( Y \) to \(|D|\), then the support of association rule \( R \) is:

\[
\text{Support} (X \Rightarrow Y) = \frac{\text{count} (X \cap Y)}{|D|}
\]  

The support of association rule \( R \) reflects the probability of simultaneous occurrence of \( X \) and \( Y \), the minimum support in \( R \) is the minimum support threshold of the item set, denoted as \( \text{sup}_\text{min} \), it represents the minimum standard for filtering association rules. The item set whose support is greater than or equal to \( \text{sup}_\text{min} \) is called frequent item set \( L \), and the frequent set whose length is \( k \) denotes as \( L_k \). The support of \( L_k \) is equal to the support of its association rule.

2.4 Confidence

For the association rule \( R \), the confidence refers to the ratio of the number of subsets containing \( X \) and \( Y \) to subsets of \( X \), which is:

\[
\text{Confidence} (X \Rightarrow Y) = \frac{\text{support} (X \Rightarrow Y)}{\text{support} (X)}
\]  

Confidence reflects the probability that \( T \) contains \( Y \) if it contains \( X \). The minimum confidence is also the minimum confidence threshold of the item set, denoted as \( \text{conf}_\text{min} \). In general, only association rules with high support and confidence are the rules that are eventually discovered.

2.5 \( L_2 \) Mining algorithm

In the ARFS proposed in this paper, first we need to find \( L_2 \) that is greater than or equal to \( \text{sup}_\text{min} \) between feature attributes and categories in \( D \), then calculate the confidence of these frequent 2-items set. In order to mine frequent 2-items set \( L_2 \), this paper applies the principle of Apriori algorithm and proposes an \( L_2 \) mining algorithm based on Apriori. Its pseudo code is shown in Figure 2.
In the above algorithm, $T_2$ is a 2-items subset of feature attributes and categories, many redundant feature attributes in it. In order to improve the efficiency of the algorithm and reduce the time complexity of the algorithm, in this paper, the parameter $sup_{\text{min}}$ is set to filter out a part of unrelated redundant features in advance, and then complete the subsequent feature selection work.

3. Feature Selection Algorithm Based on Association Rules

In order to reduce the size of the feature subset and improve the efficiency of the feature selection algorithm without reducing the accuracy, this paper proposes a feature selection algorithm based on association rules, ARFS. Although feature selection has been fully studied in the field of pattern recognition and data mining, the study of feature selection methods based on association rules is rare. The ARFS proposed in this paper firstly needs to calculate the confidence of the frequent 2-items set $L_2$ between the feature attribute and the category; then it obtains the maximum value of the confidence of all feature attributes in a certain feature as the confidence of the feature, which is denoted as $\text{Max.} (L_2, \text{Conf})$. The confidence of this feature is used to measure the correlation between features and categories; then we can reorder these features by assigning different weights to the features based on the relevance of each feature and category. Because in advance we reorder the features according to the correlation metrics, a heuristic search strategy is used to search for feature subsets, that is the sequential forward selection method SFS, also called the set increase method. Finally, the accuracy of the decision tree classifier is used as the final basis to select the corresponding feature subset. The feature selection starts with an empty set, and then a feature subset is formed by combining the features with a certain step size and the selected features to form a feature subset. This process is iterated until the feature subset with the highest accuracy is selected.

In the ARFS proposed in this paper, because the association rule can find the association between the feature attributes and the categories in the dataset, it uses the maximum strategy to calculate the confidence between feature attributes and categories, and uses this confidence to evaluate the correlation between features and categories. Then sort the features by the size of their relevance weights, and finally get an ordered sequence of features whose relevance is from largest to smallest. The feature sequence is sorted by the relevance weights, so selecting the SFS method on this basis can select the feature subset with the least scale as much as possible. Because the ARFS adopts the SFS method to match the ordered feature sequence, it can also reduce the time complexity of the search strategy. The pseudo code of the algorithm is shown in Figure 3.
In the above algorithm, when sorting according to the confidence, \( L_2 \) adopts the Max strategy that is to sort the features by the maximum confidence of feature attributes. In calculating the maximum accuracy \( \text{accuracy}_\text{max} \) of the feature subset, the Max strategy is also adopted. The \( \text{accuracy}_\text{max} \) is updated with the maximum accuracy calculated during each iteration. In the SFS, according to the ordered feature sequence, the feature is gradually added to the feature subset from front to back with a certain step length \( \text{divide}_\text{length} \). The default value of the \( \text{divide}_\text{length} \) is 1. In the process of iteratively selecting the optimal feature subset, the stopping criteria is that the accuracy difference between the current feature subset and the current optimal feature subset is less than 0, \( \Delta \text{acc} < 0 \); It must also meet the condition that the frequency of \( \Delta \text{acc} < 0 \) is less than the value of parameter \( \beta \), and the default value of \( \beta \) is the length of the original feature vector. When the above two conditions are satisfied, the iteration can be stopped, and the feature subset solved by the ARFS is finally output.

### 4. Experimental results and analysis

In order to evaluate the performance of the ARFS proposed in this paper, we conducted three comparison experiments of Evaluation feature selection, accuracy and Runtime on five public datasets of UCI machine learning database. Each experiment runs the following four algorithms: ARFS, RFE-SVM, ReliefF, and CART, which CART classifier is a benchmark comparison algorithm. These algorithms in this paper are implemented in the experimental environment of Intel Core i5-2450 2.50GHz CPU, 4GB memory, win10 64-bit operating system, PyCharm application software using Python programming language and scikit-learn[21] machine learning module.

The data types of the five public datasets of the UCI machine learning database used in the experiment are discrete type, where the cmc dataset is the abbreviation of the contraceptive method choice dataset; the monks-1 and the monks-2 datasets only use the test set; The spect-heart and spectf-heart datasets combine the test set and the training set together for training. The dataset details are shown in Table 1.

| DataSet   | Instances | Features | Classes |
|-----------|-----------|----------|---------|
| monks-1   | 432       | 7        | 2       |
| monks-2   | 432       | 7        | 2       |
| cmc       | 1473      | 9        | 3       |
| spect-heart| 267      | 22       | 2       |
| spectf-heart| 267     | 44       | 2       |

In order to obtain statistically significant experimental results, a 10-fold cross validation experiment was used in this paper. That is, the dataset is divided into 10 \( S_1, S_2, \ldots, S_{10} \). In the \( i \)th iteration, \( S_i \) is used as a test set, and the rest subsets are used to train the classifier. The average of 10
results is used as an estimate of the accuracy. The classifier uses the CART decision tree. In order to evaluate the pros the experimental results, this paper compares the CART, ReliefF, RFE-SVM and ARFS for the three evaluation indexes of feature attribute, accuracy and runtime, the attribute is the size of the optimal feature subset selected by the feature selection algorithm; the accuracy is the average value of the 10-fold cross validation experiment corresponding to the experimental result; and the runtime is the efficiency of the algorithm execution. Table 2-4 shows the results of the three groups of comparative experiments on the five datasets for the above four algorithms.

Table 2. Experimental results on number of attributes.

| DataSet    | attributes |
|------------|------------|
|            | CART | ReliefF | RFE-SVM | ARFS |
| monks-1    | 7    | 7      | 2       | 2    |
| monks-2    | 7    | 1      | 1       | 1    |
| cmc        | 9    | 8      | 6       | 6    |
| spect-heart| 22   | 14     | 4       | 1    |
| spectf-heart| 44   | 19     | 14      | 4    |
| Average    | 17.8 | 9.8    | 5.4     | 2.8  |

Table 3. Experimental results on classification accuracy.

| DataSet    | Accuracy/% |
|------------|------------|
|            | CART | ReliefF | RFE-SVM | ARFS |
| monks-1    | 85.6501| 85.6501| 93.0920 | 93.0920|
| monks-2    | 62.8647| 67.0772| 67.0772| 67.0772|
| cmc        | 42.2017| 42.3993| 42.5455| 43.7594|
| spect-heart| 72.1937| 76.6809| 78.3333| 79.4160|
| spectf-heart| 69.3162| 75.9687| 77.1368| 76.1254|
| Average    | 66.4453| 69.5552| 71.6370| 71.8940|

Table 4. Experimental results on runtime.

| DataSet    | Runtime/s |
|------------|-----------|
|            | CART | ReliefF | RFE-SVM | ARFS |
| monks-1    | 0.1155| 3.5018  | 8.4740  | 1.8397|
| monks-2    | 0.0257| 2.0547  | 0.3995  | 1.3306|
| cmc        | 0.2136| 33.3399 | 3.3094  | 32.1259|
| spect-heart| 0.2932| 0.8689  | 1.2267  | 6.3397|
| spectf-heart| 0.3528| 1.6437  | 234.9267| 26.4117|
| Average    | 0.2002| 8.2818  | 49.6673 | 13.6095|

The experimental results shown in Table 2, the ARFS is superior to the three feature selection algorithms in the selection of attributes. The average numbers of the features of the ARFS on the 5 groups of public datasets is 2.8, the result is the best, which indicates that the ARFS can select the least feature subset. Table 3 is the experimental results of the average accuracy after 10-cross validation, in addition to the spectf-heart dataset, the accuracy of the ARFS is slightly worse than the RFE-SVM. In the other datasets, the ARFS is superior to the other three contrast algorithms in the accuracy, the average accuracy of the ARFS on the 5 groups of public datasets is 71.8940%, the result is the best, which shows that the ARFS can improve the accuracy. The time efficiency of the CART is the highest, because the algorithm is the benchmark of our experiment. In the time efficiency of the algorithm, we only use the results of CART's runtime experiment as our optimization goal, but not compared with the results of the experiment. This paper compares and analyzes the runtime efficiency of the other three algorithms. From the experimental results of Table 4, it is known that the ARFS is superior to the RFE-SVM in the average runtime, and is slightly inferior to the ReliefF. ReliefF belongs to filter feature selection algorithm, which is not combined with classifier in the process of
feature selection, so the efficiency of runtime will be much higher. The ARFS and the RFE-SVM all belong to the wrapper algorithm, and the average runtime of the ARFS on the 5 sets of public datasets is 13.6095s, experimental results show that the average efficiency of ARFS is higher than RFE-SVM. Therefore, based on the above experimental results, the following conclusions are drawn: The ARFS can greatly reduce the number of attributes under the premise of guaranteeing the accuracy of the classifier, and improve the efficiency of the algorithm, provide an efficient pre-processing method for subsequent data mining, pattern recognition, or machine learning on the dataset.

5.Conclusion
The feature selection algorithm that this paper proposed maintains the accuracy of the classifier, reduces the number of features of the dataset greatly, improves the efficiency, and has a good Generalization performance. The algorithm first mines association rules of frequent 2-items set of feature attributes and categories in the dataset, then ranks feature based on the confidence of the frequent 2-item set, and then divides feature subset with the sequential forward selection method, finally, selects the features by the evaluation standard based on decision tree classifier classification performance. This paper selects five public datasets in the UCI Machine Learning Repository to perform three comparative experiments with CART, ReliefF, RFE-SVM, and the proposed ARFS. The experimental results show that the average of feature subset size is 2.8 features; the average accuracy is 71.8940%; the average value on the runtime is 13.6095s in ARFS. The ARFS proposed in this paper is superior to the contrast feature selection method in the feature subset size and the accuracy. However, ARFS is slightly inferior to the ReliefF in the efficiency of the feature selection algorithm.

The feature selection algorithm based on association rules in the scope of application also has some limitations that it only can be applied when the feature attribute value is a discrete dataset. In further work, feature selection will be implemented for continuous feature attributes.

Acknowledgments
We would like to thank reviewers for their comment. This work is supported by the Science and Technology Commission of Shanghai Municipality (Nos. 16511102800)

References
[1] G.H. John, R. Kohavi, K. Pfleger. Irrelevant Features and the Subset Selection Problem. Proc of the 11th Int Conf on Machine Learning, 121-129 (1994)
[2] D. Manoranjan, L. Huan. Feature Selection for Classification. Intelligent Data Analysis, 1(3): 131-156 (1997)
[3] H. Liu, H. Motoda. Feature selection for knowledge discovery and data mining. Boston: Kluwer Academic Publishers, 1998.
[4] L.C. Molina, B. Lluis, N. Àngela. Feature selection algorithms: A survey and experimental evaluation. Barcelona: Universitat Politecnica Catalunya, 2002.
[5] Z.H. Sun, B. George, M. Ronald. Object detection using feature subset selection. Pattern Recognition, 37(11): 2165-2176 (2004)
[6] L. Wang. Intelligent Optimization Algorithms with Applications. Beijing: Tsinghua University Publishers, 2004.
[7] A. Tsymbal, P. Seppo, W.P. David. Ensemble feature selection with the simple Bayesian classification. Information Fusion, 4(2): 87-100 (2003)
[8] B.L. Wu, A. Tom, F. David, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics, 19(13): 1636-1643 (2003)
[9] Z.H. Zhou. Machine learning. Beijing: Tsinghua University Publishers, 2016.
[10] H. Liu, R. Setiono. A Probabilistic Approach to Feature Selection: A filter Solution. Proc of Int Conf on Machine Learning, 319-327(1996)
[11] P. Langley. Selection of relevant features in machine learning. Proc of the AAAI Fall Symposium on Relevance, 1-5(1994)
[12] J.R. Quinlan. Induction of decision trees. Machine Learning, 1(1): 81-106 (1986)
[13] L. Breiman. Classification and Regression Trees. New York: Chapman and Hall/CRC, 2011.
[14] K. Kira, L.A. Rendell. The Feature Selection Problem: Traditional Methods and a New Algorithm. Proc of the 9th National Conf on AI, 129-134 (1991)
[15] I. Kononenko. Estimating Attributes: Analysis and Extension of Relief. Proc of the European Conf on Machine Learning, 171-182 (1994)
[16] W.H. Hsu Genetic wrappers for feature selection indecision tree induction and variable ordering in Bayesian network structure learning. Information Sciences, 163(17): 103-122 (2004)
[17] L.H. Chiang, R.J. Pell. Genetic algorithms combined with discriminant analysis for key variable identification. J of Process Control, 14(2): 143-155 (2004)
[18] I. Guyon, W. Jason, B. Stephen, et al. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1/2/3): 389-422 (2002)
[19] UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/MLRepository.html
[20] R. Agrawal, T. Imilinski, A. Swami. Mining association rules between sets of items in large database. Proc of the ACM SIGMOD Conf on MD, 207-216(1993)
[21] Scikit-learn. http://sklearn.apachecn.org/cn/0.19.0/