Constructive Meta-level Feature Selection
Method Based on Method Repositories

Hidenao Abe and Takahira Yamaguchi

1 Department of Medical Informatics, Shimane University,
89-1 Enya-cho Izumo Shimane, 693-8501, Japan
abe@med.shimane-u.ac.jp
2 Faculty of Science and Technology, Keio University,
3-14-1 Hiyoshi Kohoku Yokohama, 223-8522, Japan
yamaguti@ae.keio.ac.jp

Abstract. Feature selection is one of the key issues related with data pre-processing of classification task in a data mining process. Although many efforts have been done to improve typical feature selection algorithms (FSAs), such as filter methods and wrapper methods, it is hard for just one FSA to manage its performances to various datasets. To above problems, we propose another way to support feature selection procedure, constructing proper FSAs to each given dataset. Here is discussed constructive meta-level feature selection that re-constructs proper FSAs with a method repository every given datasets, de-composing representative FSAs into methods. After implementing the constructive meta-level feature selection system, we show how constructive meta-level feature selection goes well with 32 UCI common data sets, comparing with typical FSAs on their accuracies. As the result, our system shows the highest performance on accuracies and the availability to construct a proper FSA to each given data set automatically.

1 Introduction

Feature selection is one of the key procedures to get a better result from the data mining process. However, it is difficult to determine the relevant feature subset before the mining procedure. At practical data mining situations, data miners often face a problem to choose the best feature subset for a given data set. If it contains irrelevant or redundant features, a data miner can’t get any satisfactory results from mining/machine learning scheme. Irrelevant features not only lead to lower performance of the results, but also preclude finding potentially existing useful knowledge. Besides, redundant features not affect the performance of classification task, but influence the readability of the mining result. To choose a relevant feature subset, data miners have to take trial-and-error testing, expertise for the given feature set, or heavy domain knowledge for the given data set.

Feature selection algorithms (FSAs) have been developed to select a relevant feature subset automatically as a data pre-processing in a data mining process.
The performance of FSA is always affected by a given data set. To keep their performance higher, a user often tries to execute prepared FSAs to his/her dataset exhaustively. Thus a proper FSA selection is still costly work in a data mining process, and this is one of the bottle necks of data mining processes.

To solve the above problems, we have developed a novel feature selection scheme based on constructive meta-level processing. We have developed a system to construct proper FSAs to each given data set with this scheme, which consists of decomposition of FSAs and re-construction of them. To de-compose current FSAs into functional parts called ‘methods’, we have analyzed currently representative FSAs. Then we have constructed the feature selection method repository, to re-construct a proper FSA to a given data set.

After constructing the feature selection method repository, we have implemented a system to choose a proper FSA to each given data set, searching possible FSAs obtained by the method repository for the best one. Taking this system, we have done a case study to evaluate the performance of FSAs on 32 UCI common data sets. As the result, the performance of FSAs has achieved the best performance, comparing with representative higher performed FSAs.

2 Related Work

After constructing a feature set to describe each instance more correctly, we take a FSA to select an adequate feature subset for a prepared learning algorithm.

To improve classification tasks at data mining, many FSAs have been developed [2, 3, 4]. As shown in the survey done by Hall [5], wrapper methods [6] such as forward selection and backward elimination have high performance with high computational costs. Besides, filter methods such as Relief [7, 8]. Information Gain and FOCUS [9] can be executed more quickly with lower performance than that of wrapper methods. Some advanced wrapper methods such as CFS [10], which executes a substitute evaluator instead of a learned evaluator, have lower computational costs than wrapper methods. However, these performances are still non-practical, comparing with wrapper methods.

We also developed a novel FSA called ‘Seed Method’ [1]. Seed Method has achieved both of practical computational cost and practical performance, because it improves wrapper forward selection method, determining a proper starting feature subset for given feature set. With an adequate starting subset, this method can reduce the search space of $2^n$ feature subsets obtained by $n$ features.

To determine an adequate starting subset, the method extracts a feature subset with Relief.F and C4.5 decision tree [11] from given feature set.

Although studies done by [6, 12, 13] have shown each way to characterize FSAs, they have never discussed any way to construct a proper FSA to a given data set. So, a data miner still selects FSA with exhaustive executions of prepared FSAs, depending on his/her expertise. Weka [14] and Yale [15] provide many feature selection components and frameworks to users. We can construct several hundred FSAs with these materials. However, they never support to choose a proper one.