Using Ensemble Feature Selection Approach in Selecting Subset with Relevant Features

Mohammed Attik

LORIA/INRIA-Lorraine, Campus Scientifique - BP 239 - 54506
Vandœuvre-lès-Nancy Cedex, France
Mohammed.Attik@loria.fr
http://www.loria.fr/~attik

Abstract. This study discusses the problem of feature selection as one of the most fundamental problems in the field of the machine learning. Two novel approaches for feature selection in order to select a subset with relevant features are proposed. These approaches can be considered as a direct extension of the ensemble feature selection approach. The first one deals with identifying relevant features by using a single feature selection method. While, the second one uses different feature selection methods in order to identify more correctly the relevant features. An illustration shows the effectiveness of the proposed methods on artificial databases where we have *a priori* the informations about the relevant features.

1 Introduction

Feature selection can be viewed as one of the most fundamental problems in the field of the machine learning. It is defined as a process of selecting relevant features out of the larger set of candidate features. The relevant features are defined as features that describe the target concept. Hence, two degrees of relevance are defined: weak and strong relevances. Strong relevance implies that the feature is indispensable in the sense that it cannot be removed without loss of classification accuracy. While, weak relevance implies that the feature can sometimes contribute to classification accuracy. For a discussion of relevance *vs.* usefulness and definitions of the various notions of relevance, see the review articles of [1] and [2]. Many potential benefits have been given by feature selection task which can be summarized as facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times and defying the curse of dimensionality to improve prediction performance. Feature selection methods can be divided into two main groups: filter and wrapper approaches. The main difference is that the wrapper method makes use of the classifier, while the filter method does not. The wrapper approach is clearly more accurate, but it tends to be more computationally expensive than the filter model. In this work, the accuracy and stability of the feature selection results are more important than the computation complexity where the wrapper methods are selected in order to understand data.

J. Wang et al. (Eds.): ISNN 2006, LNCS 3971, pp. 1359–1366, 2006.

Traditional feature selection methods are used to find the relevant subset from the original feature set which presents an influence on classification performance or which describes the target concept. Kohavi and John's [3] proved that the efficiency of a set of features to learning depends on the learning algorithm itself. The appropriate feature subset for one learning algorithm may not be the appropriate feature subset for another one. Hence, many works are proposed to combine feature selection techniques and classifier ensemble in order to improve the classification performance: random subspace methods, input decimation methods, feature subspace methods performed by partitioning the set of features and other methods for combining different feature sets using genetic algorithms [4]. As result of this combining approach, which is called *ensemble feature selection*, the selected subset of features is defined as a subset of features out of set of relevant candidate subsets, given by the ensemble feature selection approach, which maximize the classification performance. Furthermore, the optimal feature subset selection is defined as a small subset of features, out of set of subsets given by ensemble feature selection approach, that ideally is necessary and sufficient to describe the target concept [5].

In the following we present some important remarks which motivated our work: (1) All previous research works are interested only in selecting a relevant subset of features which describe the target concept, but not relevant subset of relevant features. Unfortunately, the process of selecting relevant features always remains a serious challenge. (2) Research works have shown that the best individual features do not necessarily constitute the best set of features, the relevance of a feature does not imply that it is in the optimal feature subset and, some what surprisingly, irrelevance does not imply that it should not be in the optimal feature subset [6]. (3) Generally, the measures used to evaluate the feature selection methods are classification performance and subset size. Unfortunately, no research work proposes the measures to evaluate the feature selection method result by comparing to correct solutions which contain only the relevant features.

According to above motivating remarks, we propose to: (1) determine relevant features. Roughly, we propose to determine all relevant features which describe the target concept. Ensemble feature selection approach is used to achieve these tasks, (2) select relevant subset with only relevant feature. Hence, this task uses the results obtained by the process of determining relevant features, (3) define other measures to evaluate methods of feature selection to determine the relevant features.

The next Section presents in detail our solutions for feature selection. In Section 3, an illustration of our solutions using Optimal Brain Surgeon (OBS) variants for feature selection will be given. These techniques are considered as wrapper methods for supervised neural networks. Finally, some conclusions are drawn in Section 4.

2 Our Propositions for Feature Selection

In this section we present our propositions to select relevant features.

2.1 Description of Solutions

We present in detail the following tasks: (1) building a set of relevant subset, (2) selecting relevant features, (3) estimating the number of the relevant features, (4) using multiple feature selection methods and (5) selecting subset with relevant subset.

Building a set of relevant subset task. We use a simple variant of ensemble feature selection approach (see Figure (1)) which is the base of all our solutions. It is defined by: (1) Training separately multiple component classifiers. These classifiers are defined by the homogeneous main properties: error function, error minimization algorithm, etc. The diversities are given by varying: training/validating/testing data subset, data training presentation order, etc. In this step we propose to fix a minimum classification performance threshold for all classifiers. (2) Applying an adapted feature selection method to these classifiers to obtain the relevant Subsets of features.

Selecting relevant features task. We propose a method to select relevant features (see Figure (2)). This method exploits the relevant subset set given by the previous task and require a priori information about the number of the relevant features. It is defined by: (1) Building the distribution of pruned/preserved features by exploiting the relevant subset set. This distribution informs us the importance of a feature compared to another, in this case we can make a ranking list of preserved features. (2) Determining the relevant features by using threshold which presents the number of relevant features. If we know a priori the number of relevant features (Nbr) it can be easily to select all relevant features.

The information about the distribution characteristics of the used feature selection method can represent the new evaluation measures i.e. the best feature selection method is the method which can inform us sufficiently about the relevant features.

Estimating of the relevant features number task. We propose a method (see Figure (2)) to estimate the number of the relevant features (Nbr). Nbr is useful for our proposed method of relevant feature selection described above. This estimating method exploits the relevant subset set given by ensemble feature selection. It is defined by: (1) Building the distribution of the number of pruned/preserved features. This distribution informs us about the different number of features which give the same classification performance. (2) Determining the number of relevant features. This step exploits the information about the characteristics of the used feature selection method. The simple proposed solutions to estimate the number of relevant feature, Nbr, are given by:

$$\begin{cases} Nbr \in [Min\ ,\ Max] \\ \text{or } Nbr \in [Min\ ,\ Peak] \\ \text{or } Nbr \in [Peak\ ,\ Max] \\ \text{or } Nbr \approx \frac{1}{2}(Min\ +\ Max) \end{cases}$$

where $Peak$ represent the height frequencies of preserved features, Min is the minimum number of preserved features and Max is the maximum number of preserved features.

The information about the distribution characteristics of the used feature selection method can represent the new evaluation measures i.e. the best feature selection method is the method which can inform us sufficiently about the number of relevant features.

Using multiple feature selection methods task. We propose to use multiple methods of feature selection in order to improve estimating the number of relevant features (see Figure (3)). Each feature selection method presents some characteristics which can be used to determine the number of relevant features. So, the different solutions can ensemble contribute to estimate more sufficiently Nbr. The simple solution of the number of relevant feature, Nbr, can be given by:

$$\begin{cases} Nbr \in [min(Peak_i)\ ,\ max(Peak_i)] \\ \text{or } Nbr \approx \frac{1}{N_m} \sum_i^{N_m} Peaks_i \end{cases}$$

where $Peak_i$ represent the height frequencies of preserved features, Min_i is the minimum number of preserved features, Max_i is the maximum number of preserved features for a method i and N_m is the number of the feature selection methods used.

Selecting subset with relevant features. After determining the set of relevant features, we can select subset of relevant features (see Figure (4)). This selection depends on the evaluation criteria, the decision maker preferences, the application's needs. We can apply: (1) the feature selection method in order to select a relevant subset with relevant features. (2) the ensemble feature selection in order to select the best or the optimal subset with relevant features .

2.2 Our Algorithms

According to different task given previously, we propose two algorithms in order to select relevant features. The first algorithm uses One Feature Selection Method to Select Relevant Features (OFSM-SRF), for more details see Algorithm (1). The second algorithm uses Multiple Feature Selection Methods to Select Relevant Features (MFSM-SRF), for more details see Algorithm (2).

3 Illustration

Feature selection methods for artificial neural networks have been first designed for topology optimization dealing with the over-training problem. Several heuristic methods based on computing the saliency (also termed sensitivity) have been proposed: Optimal Brain Damage (OBD) [7] and Optimal Brain Surgeon (OBS) [8]. These methods are known as pruning methods. The principle of these techniques is the weight with the smallest saliency will generate the smallest error

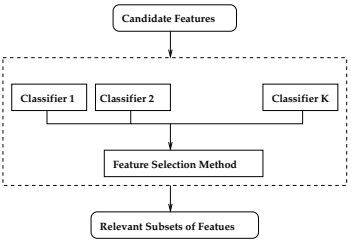

Fig. 1. The used ensemble feature selection scheme which is based on a single feature selection method

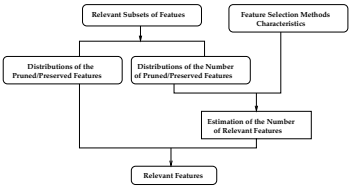

Fig. 2. The used scheme in selecting the relevant features and estimating the number of relevant features

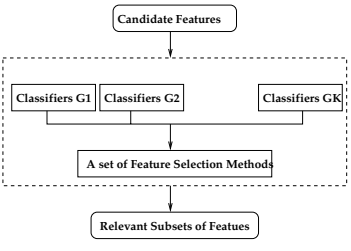

Fig. 3. The used ensemble feature selection scheme which is based on multiple feature selection methods

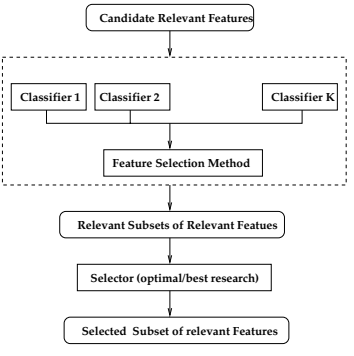

Fig. 4. The used scheme in selecting subset with relevant features

variation if it is removed. OBS has inspired some methods for feature selection such as Generalized Optimal Brain Surgeon (G-OBS), Unit-Optimal Brain Surgeon (Unit-OBS), Flexible-Optimal Brain Surgeon (F-OBS) and Generalized Flexible-Optimal Brain Surgeon (GF-OBS). G-OBS [9, 10] proposes to delete a subset of m weights in a single step. Unit-OBS [9] is based on G-OBS calculation to select input unit will generate the smallest increase of error if it is removed. The F-OBS [10], its particularity is to remove connections only between the input layer and the hidden layer. The GF-OBS is a combination of F-OBS and G-OBS [10], its removes in one stage a subset of connections only between the input layer and the hidden layer.

In this work we use the first Monk's problem. This well-known problem requires the learning agent to identify (true or false) friendly robots based on six nominal attributes. The attributes are head_shape (round, square, octagon), body_shape (round, square, octagon), is_smiling (yes, no), holding (sword, balloon, flag), jacket_color (red, yellow, green, blue) and has_tie (yes, no). The "true" concept for this problem is (head_shape = body_shape) or (jacket_color = red).

Algorithm 1. OFSM-SRF

1 - design a set of supervised classifiers according to the selected model.
2 - give a threshold which presents a minimum classification performance for all classifiers
3 - train the supervised algorithms with different initial conditions
4- apply feature selection method fs adapted to the classification algorithm
5 - build the distribution of the number of pruned/preserved features and the distribution of pruned/preserved features
6 - study these distributions and exploit the information regarding the used feature selection method needed to determine the relevant features.

Algorithm 2. MFSM-SRF

1 - select a set of different classification model M_i (not necessary homogeneous).
2 - associate for each M_i a set of adapted feature selection method fs_{ij}.
3 - apply OFSM-SRF for each M_i and fs_{ij}.
4 - combine the obtained results to determine relevant features

The training dataset contains 124 examples and the validation and test datasets contains 432 examples.

To forecast the class according to the 17 input values (one per nominal value coded as 1 or -1 if the characteristic is true or false), the Multilayer perceptron (MLP) starts with 3 hidden neurons containing a hyperbolic tangent activation function. This number of hidden neurons allows a satisfactory representation which able to solve this discrimination problem. The performance in classification are equal to 100% according to the confusion matrix. In this study, GF-OBS and G-OBS remove three weights at the same time. For each method we build 300 MLPs by varying: the training/validating subset, the initialization of the weight and the order of the data presentation.

Figures 5, 7, 6 and 8 give the results by applying Unit-OBS, F-OBS, GF-OBS, OBS and G-OBS methods, which it represent the distribution of pruned features and distribution of pruned feature number. For every algorithm, the number of pruned features differs from 2 to 12 according to each initialization. But the frequency depends on the algorithm. Unit-OBS presents the high frequencies for

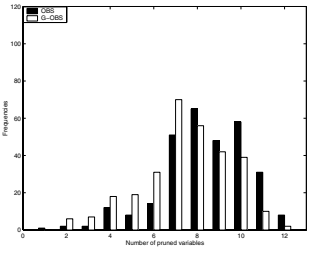

Fig. 5. OBS, G-OBS: number of pruned variables distribution

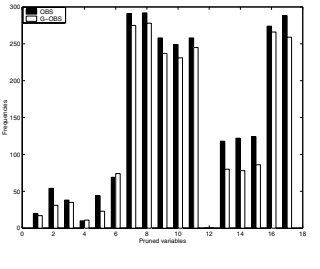

Fig. 6. OBS, G-OBS: pruned variables distribution

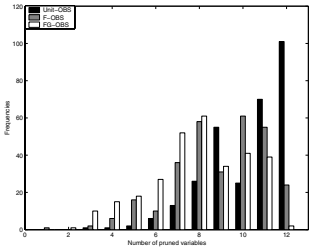

Fig. 7. Unit-OBS, F-OBS, FG-OBS: number of pruned variables distribution

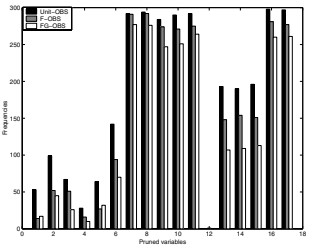

Fig. 8. Unit-OBS, F-OBS, FG-OBS: pruned variables distribution

a large number of pruned features compared to the other methods. According to the first Monk rules, the concept is true if the head shape is equal to the body shape i.e. if feature 1 = feature 4 and feature 2 = feature 5 and feature 3 = feature 6. Thus, a good method should preserve these features and feature 12 (indeed, the concept is also true if feature 12 is true). Nevertheless, it is possible to obtain the desired performance by eliminating some of the first six features. In this case, a cleaver algorithm will prune couples of features. For example, to prune feature 1 (head_shape=round) allows to prune feature 4 (body_shape=round). In other case the ideal method is to eliminate 10 features and to keep only 7 features.

If we analyze the distributions of the number of pruned features. We notice that for the first property, it is not really preserved by these techniques. For the second property, each method presents a $peak_i$ which informs us about the number of features to keep. According to the obtained results we have 5, 7, 9 variables which correspond to Unit-OBS, F-OBS, and OBS methods respectively. The peaks obtained by various methods can be presented an advantage to estimate the number of pertinent features Nbr: Unit-OBS gives the minimum of relevant features, F-OBS almost it estimates the relevant feature number and the OBS gives more than the relevant feature number. In this case we can determine the confidence interval of the relevant features number $Nbr \in [min(peak_i)\ ,\ max(peak_i)] = [5, 9]$ and the number of relevant feature is given by: $Nbr \approx peak_{F-OBS} = 7$ or $Nbr \approx \frac{1}{N_m} \sum_i^{N_m} peaks_i = \frac{1}{3}(5 + 7 + 9) = 7$ where N_m is the number of the feature selection methods used.

4 Conclusion

In this paper, we have focused on selecting the subset of relevant features in order to understand data. Understand data is helpful or necessary to make decision in different domains (e.g. marketing, biology, medicine, ...). We have presented novel solutions for achieve this task. Our solutions are a direct extension of ensemble feature selection approach. The first solution is based on using a single feature selection method. Conversely, the second approach is based on using multiple feature selection methods in order to determining more correctly all

relevant features. These two solutions require a priori informations about the number of relevant features, an estimation of this number is given. We have showed that the best feature selection method is the method which can inform us sufficiently about the relevant features or the number of relevant features. An overall experimentation of our solutions have been achieved on artificial data where the information about the relevant features and rules between these features are available and the data dimensionality is supported by the used feature selection methods of optimal brain surgeon family. It has clearly highlighted that these solutions represent an important step for data comprehension.

References

1. Kohavi, R., John, G.: Wrappers for feature selection. Artificial Intelligence **97** (1997) 273–324
2. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence **97** (1997) 245–271
3. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97** (1997) 273–324
4. Valentini, G., Masulli, F.: Ensembles of learning machines (2002)
5. Kira, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: Proc. of AAAI-92, San Jose, CA (1992) 129–134
6. Piramuthu, S.: Evaluating feature selection methods for learning in data mining applications. In: HICSS (5). (1998) 294–301
7. Le Cun, Y., Denker, J.S., Solla, S.A.: Optimal brain damage. In Touretzky, D.S., ed.: Advances in Neural Information Processing Systems: Proceedings of the 1989 Conference, San Mateo, CA, Morgan-Kaufmann (1990) 598–605
8. Hassibi, B., Stork, D.G.: Second order derivatives for network pruning: Optimal brain surgeon. In Hanson, S.J., Cowan, J.D., Giles, C.L., eds.: Advances in Neural Information Processing Systems. Volume 5., Morgan Kaufmann, San Mateo, CA (1993) 164–171
9. Stahlberger, A., Riedmiller, M.: Fast network pruning and feature extraction by using the unit-OBS algorithm. In Mozer, M.C., Jordan, M.I., Petsche, T., eds.: Advances in Neural Information Processing Systems. Volume 9., The MIT Press (1997) 655
10. Attik, M., Bougrain, L., Alexandre, F.: Optimal brain surgeon variants for feature selection. In: International Joint Conference on Neural Networks - IJCNN'04, Budapest, Hungary. (2004)