Neurocomputing 73 (2009) 295–303

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Boosting feature selection using information metric for classification

Huawen Liu [a], Lei Liu [a,*], Huijie Zhang [b]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, PR China
[b] Department of Computer Science, Northeast Normal University, Changchun 130021, PR China

ARTICLE INFO

*Article history:*
Received 10 September 2008
Received in revised form
15 July 2009
Accepted 25 August 2009
Communicated by G. Palm
Available online 25 September 2009

*Keywords:*
Classification
Boosting
Feature selection
Information metric
Filter model

ABSTRACT

Feature selection plays an important role in pattern classification. Its purpose is to remove redundant features from data set as many as possible. The presence of useless features may not only deteriorate the performance of learning algorithms, but also obscure important information (e.g., intrinsic structure) behind data. Along with new and emerging techniques, data sets in many domains are becoming larger and larger and many irrelevant features are often prevailing in these data sets. This, however, poses great challenges to traditional learning algorithms, such as low efficiency and over-fitting. Thus, it becomes apparent that an efficient technique is needed to eliminate redundant or irrelevant features from the data sets. Currently, many endeavors to cope with this problem have been attempted and various outstanding feature selection methods have been proposed. Unlike other selection methods, in this paper we propose a general scheme of boosting feature selection method using information metric. The primary characteristic of our method is that it exploits weight of data to select salient features. Furthermore, the weight of data will be dynamically changed after each candidate feature has been selected. Thus, the information criteria used in feature selector can exactly represent the relevant degree between features and the class labels. As a result, the selected feature subset has maximal relevance to the class labels. Simulation studies carried out on UCI data sets show that the classification performance achieved by our proposed method is better than those of other selection methods in most cases.



## 1. Introduction

As one of fundamental tasks in pattern recognition, classification mainly concerns the issue to make a prediction on the basis of currently available knowledge in classification model, which is induced from known data [1]. The induction procedure of model is termed as learning, that is, obtaining system behavior from the past actions. Given a sample (or data set) $\mathcal{D}$ with features $\mathcal{F}$ and the class labels $\mathcal{C}$, a deterministic inducer from $\mathcal{D}$ is a hypothesis from $\mathcal{F}$ to $\mathcal{C}$. During past decades, numerous techniques, such as decision tree and support vector machine (SVM), have been introduced to construct classifier [1], and most of them work well when the number of features is small. Unfortunately, in many domains, data sets are often filled with hundreds or even tens of thousands of features. Such massive features, on one hand, provide much more potential discrimination power for classification tasks. On the other hand, they may deteriorate classification performance due to the limited number of training data. As Duda et al. highlighted in [1] that a higher probability of error may arise when the number of features in a sample beyond a certain point. Indeed, many features are irrelevant or redundant and their presence may obscure important intrinsic structure behind data. Hence, it is necessary to remove redundant features from data sets as many as possible. One of such solutions is the concept of feature selection [2].

Feature selection refers to the process of obtaining an optimal subset from the original feature space, according to given criteria to determine which features are important and informative [3]. It cannot only effectively discard irrelevant features to lessen the problem of "the curse of dimensionality" [1], but also bring potential advantages to learning algorithms, such as lower computational cost and more efficient. In addition, the induced classifier has better classification performance, easier comprehensibility, higher generality, and more robust to noises [4,5]. Just owing to this, currently many fruits about feature selection have been reaped. Roughly speaking, they fall into four categories in terms of evaluation manner [5–7]: embedded, wrapper, filter and hybrid methods [8,9]. For embedded and wrapper models, they heavily rely on specific learning algorithm whose computational cost may be very expensive [2,4]. As a result, they are less general and unscalable well on data sets with high dimensionality.

Contrastively, filter method is independent of specific learning algorithms and it evaluates the interestingness of feature by the measurements of data content [3], such as Euclidean distance, Fisher score and correlation coefficient [10]. A typical illustration

* Corresponding author. Tel./fax: +86 431 85159373.
*E-mail addresses:* Huaw.Liu@gmail.com (H. Liu), Liulei@jlu.edu.cn (L. Liu), Zhanghj167@nenu.edu.cn (H. Zhang).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.08.012

is Relief [11], where Euclidean distance has been adopted to weight the significance of features. Comparing to other metrics, information ones seem much more popular. The underlying fact is that they can measure non-linear correlation between features [12]. Several extensive experiments (see, e.g., [13,14]) have demonstrated that information metrics work well in many cases. Loosely, filter model can obtain a feature subset with more robust and general property, nevertheless it does not consider the bias of learning algorithm in selecting features.

Recently, many researchers resort to sophisticated techniques to pick salient features. For example, Sebban and Nock [8] firstly estimated evaluation criterion with minimum spanning tree for each feature, and then chose interesting features by using statistical test in a forward selection way. Das [15] pointed out that boosting method is a good choice to obtain an optimal feature subset. While Li and Yang [16] employed bootstrapping sampling technique to obtain multiple feature subsets by virtue of mutual information criterion, and then optimally integrated them into one using SVM.

Unlike other methods, in this paper, we propose a new boosting scheme for feature selection based on information metric, which is dynamically estimated on the weight of data. One may observe that in conventional filter approaches, the weight of data is never altered. Consequently, the value of evaluation criterion, e.g., mutual information, estimated on the whole sampling space is determined once a training data has been given. However, it is unreasonable because the criterion cannot exactly measure the relevance between features as the selection procedure continues. To alleviate this troublesome issue, in our method, we continuously adjust the weight of data after each candidate feature has been chosen. The proposed architecture is similar to the boosting selection methods (e.g., BDSFS [15]) in some aspects, but shows substantial differences. The most primary distinctness is that the output generated by our method is only a feature subset, not one or several classifiers.

The structure of the rest is organized as follows. Section 2 gives some basic concepts about information metrics in feature selection. In Section 3, previous related work about ensemble feature selection are briefly reviewed. Section 4 provides a new feature selection scheme on information theory by using reweighed technique. Experimental results conducted to evaluate the effectiveness of our approach are presented in Section 5. Finally, conclusions are summarized in the end.

## 2. Background

For the sake of simplification, here we only deal with discrete random variables with finite values. Suppose $X$ and $p(x)$ are a discrete random variable and its marginal density, respectively. The information amount of $X$ can be measured by *information entropy* $H(X)$, where $H(X) = -\sum p(x)\log p(x)$. Further, *mutual information* mainly quantifies how much information is shared, i.e., the relevant degree, between different variables. Given $X$ and $Y$, their mutual information is $I(X;Y) = H(X)+H(Y)-H(X,Y)$, where $I(X;Y) = 0$ indicates that they are totally irrelevant with each other. Otherwise they share more common information and highly relevant [12,17].

As mentioned above, given a data set $\mathcal{D}$ with features $\mathcal{F}$, feature selection is to identify a feature subset $S \subseteq \mathcal{F}$ such that $J(S)$ is maximal while its cardinality is minimal, where $J(S)$ is the criterion function of $S$ and a higher value of $J(S)$ indicates a better feature space. For the purpose of classification, it is advisable if $J(S)$ involves information of both input features and the class labels. Information metric based on entropy is a such criterion. Generally speaking, most of information metric based feature selection methods (MIFS) take mutual information $I(S;C)$ and its variants as evaluation criterion, that is, $J(S)$ often takes the form of $I(S;C)$. However, identifying the best subset $S$ from $\mathcal{F}$ is usually intractable in an exhaustive way. Additionally, the estimated value of $I(S;C)$ on a limited training data is incredible. To cope with these problems, many heuristic subset search or selection strategies, such as branch and bound search, beam search, probabilistic search and random search, have been addressed [18]. Currently, the common assumption behind MIFS is to select individual feature at each time in a greedy manner [3]. In this case, only $I(f;C)$ is needed to be calculated for each feature $f \in \mathcal{F}$. More specifically, the selection procedure of MIFS is briefly described as follows [3,17]:

- Initialize relative parameters: $S = \emptyset$, $F = \mathcal{F}$.
- For each candidate $f \in F$, calculate its criterion $J(f)$.
- Select the feature $f$ with the largest $J(f)$, i.e., $S = S \cup \{f\}$ and $F = F - \{f\}$.
- If $|S| < \delta$, goto the second step to select the next feature; Otherwise, $S$ is the desired subset.

During past years, many outstanding MIFS methods have been witnessed. For example, BIF is the most naive MIFS method [19], whose evaluation criterion is mutual information, i.e., $J(f) = I(C;f)$. Since the metric $J(f)$ in BIF does not concern the redundancy among selected features, Peng et al. assigned $J(f)$ with $I(C;f) - 1/|S| \cdot \sum I(f;s)$ in mRMR [20] and then chose salient features by wrapping a learning algorithm. Furthermore, Novovičová et al. took the relevance between $s \in S$ and $C$ into account in their selector called mMIFS-U [21] and their evaluation criterion is $J(f) = I(C;f) - \max(I(f;s) \cdot I(C;s)/H(s))$. Recently, Cai et al. [22] employed conditional mutual information as their evaluation function, i.e., $J(f) = I(C;f,s) - I(C;s)$. Although these criteria take different forms, they primarily consist of two basic ingredients, i.e., the relevance of candidate feature $f$ with $C$ and its redundancy with the already selected subset $S$. Liu et al. in [17] summarized most information metrics in MIFS to a general one.

## 3. Related work

Since traditional filter selection methods provide limited contribution to the performance of classifiers, many endeavors have been attempted to improve classification performance further by sophisticated techniques. As an example, Tieu and Viola [23] obtained informative features by AdaBoost learning on the training data. In this method, the feature with the least error rate will be picked at each round, and then the weight of data is re-calculated by virtue of the error rate. Finally, a classifier is built on these selected features. This method, however, is similar to the one proposed by Das [15], where the base classifier is Decision-Stump. Yin et al. [24] made use of a variant boosting to combine features, where at each round, several base classifiers are built on different feature subsets and then synthesized by weighted voting. Redpath and Lebart [25] identified feature subset by the regularized version of Boosting, i.e., Adaboost$_{Reg}$. Additionally, their search strategy is the floating feature search, not the sequential one. Instead of depending on specific learning algorithm, our proposed method is an independent one, where the error rate is estimated by information metric, not the base learning algorithm. Thus, it can be taken as a pre-processing step for classification issues and integrated with any classifier.

Bootstrapping technique has also been used to identify salient features in literatures. For instance, Xu and Zhang [26] incorporated bootstrap with boosting selection procedure, where a bootstrap sample is generated randomly and then the feature