Information Sciences 269 (2014) 176–187

Contents lists available at ScienceDirect

Information Sciences

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

Embedded local feature selection within mixture of experts

Billy Peralta *, Alvaro Soto

Department of Computer Science, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 6904411 Santiago, Chile

ARTICLE INFO

Article history:
Received 17 March 2012
Received in revised form 18 April 2013
Accepted 1 January 2014
Available online 11 January 2014

Keywords:
Mixture of experts
Local feature selection
Embedded feature selection
Regularization

ABSTRACT

A useful strategy to deal with complex classification scenarios is the "divide and conquer" approach. The mixture of experts (MoE) technique makes use of this strategy by jointly training a set of classifiers, or experts, that are specialized in different regions of the input space. A global model, or gate function, complements the experts by learning a function that weighs their relevance in different parts of the input space. Local feature selection appears as an attractive alternative to improve the specialization of experts and gate function, particularly, in the case of high dimensional data. In general, subsets of dimensions, or subspaces, are usually more appropriate to classify instances located in different regions of the input space. Accordingly, this work contributes with a regularized variant of MoE that incorporates an embedded process for local feature selection using L_1 regularization. Experiments using artificial and real-world datasets provide evidence that the proposed method improves the classical MoE technique, in terms of accuracy and sparseness of the solution. Furthermore, our results indicate that the advantages of the proposed technique increase with the dimensionality of the data.

1. Introduction

Performing classification in scenarios with large and complex intra and inter-class variation is a challenging task for most classification methods. In these cases, different subsets of instances might respond to different patterns and, even more, these patterns might arise in different subsets of dimensions. As an example, in visual recognition, changes in illumination or pose conditions usually produce drastic variations in the visual appearance of relevant objects, affecting the discriminative properties of different visual features [35]. As a further example, in gene function prediction, the expression level of particular genes can change substantially under different experimental conditions, affecting the discriminative properties of different co-expression patterns that usually arise on subsets of experiments [15].

A useful strategy to deal with complex classification scenarios is the "divide and conquer" approach. Under this strategy, a complex problem is divided into multiple simpler problems. Decision trees (DTs) are one of the oldest and most widely used classification techniques based on this strategy [36]. This technique consists of building a tree using a partitioning scheme that recursively divides the input space and adjusts local classifiers within each partition. Interestingly, each branch of the resulting tree is in charge of classifying a different subset of instances. Furthermore, classification in each branch is performed using a particular subset of dimensions. We believe that this double "divide and conquer" strategy, that adaptively adjusts each branch of the tree to deal with a selected subsets of instances and dimensions, is one of the main reasons to explain the good performance shown by DTs and their later extensions based on ensemble strategies [8]. Unfortunately, the representational space and usual learning strategies used by DTs impose relevant limitations that affect their abilities to deal with complex classification scenarios. In particular, a DT embeds a hypothesis space given by a disjunction of conjunctions of constraints. These

* Corresponding author. Tel.: +56 2 354 4712; fax: +56 2 354 4444.
E-mail addresses: bmperalt@uc.cl (B. Peralta), asoto@ing.puc.cl (A. Soto).

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2014.01.008

constraints are usually based on single [36] or low dimensional [31] partitions of the input space. Furthermore, common training strategies are based on greedy schemes that can lead to suboptimal solutions. As an example, the greedy decision at the root node of the tree constrains the conjunctions embedded by all the branches of the tree.

A probabilistic approach to the "divide and conquer" strategy is the mixture of experts (MoE) technique [19]. In contrast to DTs, this technique uses a probabilistic framework that is advantageous in managing the intrinsic uncertainty in the data. MoE divides the data into multiple regions where each region has its own classifier or expert [19]. Each expert is specified by a probability distribution that is conditioned on class values. In the mixture, predictions of experts are weighed using a global model known as gate function. This function adaptively estimates the relevance or weight assigned to each expert for the classification of each input instance.

Both, DTs and MoE, use a "divide and conquer" strategy that divides the input space to perform classification, using a hard partitioning in the case of a DT and a probabilistic, or soft, partitioning in the case of MoE. A relevant difference arises in terms of how each technique handle the dimensionality of each instance: while a DT incorporates an embedded feature selection scheme, MoE does not. We believe that a suitable embedded feature selection scheme can be a useful tool to boost the performance of the MoE technique. In particular, in our experiments for the case of high dimensional datasets we notice that the traditional MoE technique has serious difficulties to learn adequate models. Also, as the number of parameters increases with the number of dimensions, the resulting MoE models become complex usually leading to overfitting problems.

This work contributes with a MoE model that incorporates embedded local feature selection using L_1 regularization. Our main intuition is that particular subsets of dimensions, or subspaces, are usually more appropriate to classify certain input instances. Consequently, we expect to improve the accuracy of traditional MoE models by introducing a technique that adaptively selects subsets of dimensions to train each expert in the mixture.

This paper is organized as follows. Section 2 presents background information about feature selection methods, in particular L_1 regularization. Section 3 describes relevant previous works. Section 4 presents the proposed approach. Section 5 presents and discusses the results of our experiments. Finally, Section 6 presents our main conclusions and future avenues of research.

2. Background

2.1. Feature selection

In classification problems, the goal corresponds to learn a mapping from an input vector x to an output value y, where x is a vector with D dimensions and y takes categorical values. If vector x is high-dimensional, one can usually improve classification accuracy by discarding irrelevant and redundant features [14,22]. This process is known as feature or variable selection. In general, there are three main methods for feature selection:

- *Filter methods* rank each input feature x_j in relation to predicting y using a metric of goodness, such as mutual information [3], Pearson correlation [16], Fisher score [10], and chi-square statistic [26], among others [14]. Next, the features are selected according to ranking results. These methods can be incorporated in a sequential forward selection in order to find a subset of discriminant dimensions [16]. Generally, the chosen metric is independent of the final classification model [14]. Filter methods are usually fast and simple, in comparison to alternative techniques.
- *Wrapper methods* search the feature space looking for possible subsets that improve performance. For each subset, these methods execute the classification model and evaluate its resulting predictive power [22], usually using accuracy or F-measure [46]. Then, the subset of features with greatest predictive power is chosen. Main issues associated with these methods are difficulties defining the best metric to measure predictive power, as well as the computational complexity associated to the evaluation of a large number of subsets of features, $2^D - 1$ subsets in the worst case (exhaustive search).
- *Embedded methods* combine feature selection and model fitting into a single optimization problem. DT [36] and Adaboost [12] can be considered embedded techniques, although an explicit criterium for feature minimization is not included during the training process. Two popular techniques to embed feature selection inside a classification algorithm are L_1-regularization [42] and automatic relevancy determination [27].

This paper concentrates on embedded models, specifically L_1-regularization, due to their computational tractability and formal soundness. Although filter methods are faster than alternative techniques, they are usually less effective, as they use a metric that is independent of the final classification scheme. On the other hand, wrapper methods are generally more reliable, as they can take advantage of robust classification algorithms. Nevertheless, these methods are slow due to the usually large number of subsets to explore, and the complexity associated to repeatedly training a robust classifier [22]. Embedded models are attractive because they use a reliable measure of goodness, similar to wrapper methods, but they avoid retraining a predictor for each feature subset explored.

2.2. L_1 regularization

Consider the context of linear models given by the expression $y = w^T x + b$, where $x \in \Re^D$ is the input vector, $y \in \Re$ is the output value, $w \in \Re^D$ is the vector of coefficients, and $b \in \Re$ is the bias [4]. Selecting features by means of regularization fits a