Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Feature subset selection from positive and unlabelled examples

Borja Calvo [a,*], Pedro Larrañaga [b], Jose A. Lozano [a]

[a] Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country, Paseo Manuel de Lardizabal, 1, E-20018 Donostia-San Sebastián, Spain
[b] Departamento de Ingeligencia Artificial, Universidad Politécnica de Madrid, E-28660 Boadilla del Monte, Spain

ARTICLE INFO

Article history:
Received 25 January 2008
Received in revised form 23 January 2009
Available online 12 May 2009

Communicated by L. Heutte

Keywords:
Positive unlabelled learning
Partially supervised classification
Feature subset selection
Filter methods

ABSTRACT

The feature subset selection problem has a growing importance in many machine learning applications where the amount of variables is very high. There is a great number of algorithms that can approach this problem in supervised databases but, when examples from one or more classes are not available, supervised feature subset selection algorithms cannot be directly applied. One of these algorithms is the correlation based filter selection (CFS). In this work we propose an adaptation of this algorithm that can be applied when only positive and unlabelled examples are available. As far as we know, this is the first time the feature subset selection problem is studied in the positive unlabelled learning context. We have tested this adaptation on synthetic datasets obtained by sampling Bayesian network models where we know which variables are (in)dependent of the class. We have also tested our adaptations on real-life databases where the absence of negative examples has been simulated. The results show that, having enough positive examples, it is possible to obtain good solutions to the feature subset selection problem when only positive and unlabelled instances are available.

1. Introduction

Synthesising the knowledge contained in databases into classification models is a very powerful tool that can be used in a wide range of applications, from genome analysis to spam filtering. In principle, one can be tempted to think that the more information we have, the better the model we can induce, but this is only partially true.

In supervised databases we have instances characterised by some features or predicting variables and a category or class variable (Bishop, 2006, Duda et al., 2001). Given a new instance, a classification model tries to predict its class based on the value of the features, but not all the features are equally useful for the classification purpose. Non informative (poorly correlated with the class) and redundant variables (highly correlated with other features) can be harmful for some model induction algorithms. Irrelevant and redundant features are not only harmful, but they also lead to models that are too complex and increase the computational time required to obtain the classifier. Therefore, producing a small set of predictive and non-redundant features is becoming a very important step in many machine learning applications.

Two main ways to reduce of the dimensionality of classification problems have been proposed in the literature: feature extraction (Liu and Motoda, 1998) and feature subset selection (FSS) (Liu and Motoda, 2008, Guyon and Elisseeff, 2003). The former consists of combining the features in the database to obtain new, better features. The main problem with this approach is that the meaning of the original variables is lost in the newly constructed features. The latter consists of selecting the best subset of features for the classification purpose. In this work we will focus on the FSS approach.

There are three main approaches to the FSS problem (Guyon and Elisseeff, 2003, Saeys et al., 2007), namely embedded, wrapper and filter methods. Some classifier induction algorithms, such as the C4.5 algorithm (Quinlan, 1993), do not use all the available variables. This sort of FSS is known in the literature as *embedded* FSS. The *wrapper* approaches (Kohavi and John, 1997) try to identify the subset of variables that, given a classification paradigm and a dataset, provides with the best classification function. The process consists of a search in the feature subset space guided by a performance measure (typically the accuracy, though other measures can be used). Each subset is evaluated by testing the performance of the chosen paradigm in the dataset, using only the variables in the subset at evaluation. The estimation of the performance of the classifiers requires a validation scheme, such as cross validation (Efron, 1983) or bootstrap estimation (Efron and Tibshirani, 1993). As a result, the evaluation of each subset involves the training and testing of several classification functions, increasing the computational time required for the FSS process. Besides, the search for the best subset is an NP-hard problem (Amaldi and Kann, 1998) and, thus, an exhaustive search quickly becomes computationally unfeasible and search heuristics have to be used to obtain a good feature subset in a reasonable time (Inza et al., 2000). This is the main drawback of these methods. Another characteristic

[*] Corresponding author. Fax: +34 943015590.
E-mail addresses: borja.calvo@ehu.es (B. Calvo), pedro.larranaga@fi.upm.es (P. Larrañaga), ja.lozano@ehu.es (J.A. Lozano).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.04.015

of the wrapper methods (that can be good or bad, depending on the point of view) is that the subset produced by the algorithm depends on the classification paradigm considered in the search. This means that the selection obtained with a classification paradigm cannot be applied to other classification paradigms, as the solution is tuned up for that particular paradigm.

The *filter* approaches search for the best subset of variables, independently of the classification paradigm, considering the relationship between the predicting variables and the class and (sometimes) the relationship among the predicting variables. One of the most simple approaches consists of ranking the variables according to their usefulness and then selecting only those variables on the top of the ranking. The usefulness of a feature is measured univariately by means of different metrics. For instance, information theory related metrics (Cover and Thomas, 2006) evaluate the usefulness of the feature by measuring the reduction on the uncertainty of the class variable when the value of the feature at evaluation is known (Ben-Bassat, 1982). Once the features are ranked, a threshold must be set to obtain the final subset. The ranking methods are only concerned with the relevancy of the features considered and, thus, they do not filter out redundant variables.

The problem of searching for relevant and non-redundant features can be solved by a multivariate filter method known as correlation based filter selection (CFS) (Hall and Smith, 1997). This method searches for the best feature subset guided by a metric that measures both the correlation between each variable and the class and the correlation among the selected variables. The aim is to obtain a subset of relevant variables (i.e., features strongly correlated with the class) without redundancies (i.e., with a small correlation between them).

Filter methods are much faster than wrapper approaches and they are independent of the classification paradigm. Therefore, once we have a subset of features, this subset can be used in the training of any sort of model.

All the filter FSS methods mentioned above require examples from all the classes in order to measure the correlation between each feature and the class, but in some real situations, getting examples from one or more classes can be difficult or even impossible. For instance, suppose that we have a set of papers about a particular topic and we want to retrieve, from a database of (unlabelled) papers, those related to the ones in our set. We could try to obtain a set of uninteresting documents by hand labelling some papers, but this can be a very tedious task. In addition, the hand-labelled negative examples have to be representative of all the possible negative instances and this is an even harder task. Another example where getting negative instances is impossible is the identification of cancer genes (Furney et al., 2008). If we want to identify which genes are related with cancer, we have a list of positive examples (genes that have already been identified as cancer-related), but for the rest of the genes we have no information about their label (it is not possible to ensure that a given gene is not related to cancer in any possible way) and, thus, we have no negative instances. Therefore, it would be interesting to be able to build a classifier only with positive and unlabelled examples.

We can overcome the lack of negative instances by training a classifier using only positive and unlabelled examples. The problem of learning binary classifiers from only positive and unlabelled examples, known in the literature as partially supervised classification (Liu et al., 2002) or positive unlabelled learning (Denis et al., 2002), deals with this sort of situation. Many new methodologies have been developed to solve this problem (Calvo et al., 2007, Denis et al., 2003, Liu et al., 2003). In this paper we tackle the FSS problem when only positive and unlabelled examples are available. To the best of our knowledge, this is the first time the FSS problem is explicitely addressed in the positive unlabelled learning framework.

In this work we present an adaptation of the CFS algorithm that can be used without negative examples. The results obtained with this new algorithm have been compared with the ones obtained with the original CFS. For this comparison, we have used synthetic datasets obtained sampling Bayesian network models and real-life data based datasets where the absence of negative examples has been simulated.

The rest of the paper is organised as follows. In Section 2 the CFS algorithm is described and our adaptation to the positive unlabelled learning context is presented. In Section 3 our proposal is compared with the original CFS on synthetic and real-life data based problems. Finally, in Section 4 some conclusions and ideas about the future work are provided.

2. CFS based feature subset selection

Before presenting the CFS algorithm and its adaptation to the positive unlabelled learning context, some basic notation has to be introduced. Instances are characterised by a feature vector $\boldsymbol{X}$ of n components $(X_1, \dots, X_n)$ and a class variable C that can take only two values, 0 and 1 (also referred to as negative and positive); each feature X_i can take t_i values. For the sake of simplicity, the probabilities $P(X_i = j), P(X_u = v|X_i = j)$ and $P(X_i = j|C = c)$ will be abbreviated as $P(x_{ij}), P(x_{uv}|x_{ij})$ and $P(x_{ij}|c)$. $P(C = 1)$ will be denoted as p.

2.1. The CFS metric

The CFS algorithm (Hall and Smith, 1997) is based on a metric that evaluates the merit of a given set of features. This metric is then used to guide a search for the best possible subset of variables. The merit function is based on the correlation between each feature and the class (relevancy) and on the correlation among the features in the subset (redundancy). This function can be expressed as:

$$G_S = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii'}}}}$$

where k is the number of variables in the subset S, $\overline{r_{ci}}$ is the average correlation between the features in S and the class, and $\overline{r_{ii'}}$ is the average correlation among the features in S.

In (Hall and Smith, 1997) the authors measure the correlation between two variables X_i and X_u by means of the uncertainty coefficient $U(X_u|X_i)$, which is based on the mutual information $I(X_u; X_i)$ and the entropy $H(X_u)$ (Cover and Thomas, 2006). When the features are represented as random discrete variables (either because they are discrete or because they have been discretised) $U(X_u|X_i)$ is defined as:

$$U(X_u|X_i) = \frac{I(X_u; X_i)}{H(X_u)} = \frac{H(X_u) - H(X_u|X_i)}{H(X_u)}$$

$$H(X_u) = -\sum_{v=1}^{t_u} P(x_{uv}) \log P(x_{uv})$$

$$H(X_u|X_i) = -\sum_{j=1}^{t_i} P(x_{ij}) \sum_{v=1}^{t_u} P(x_{uv}|x_{ij}) \log P(x_{uv}|x_{ij})$$

Sets of irrelevant (poorly correlated with the class) and/or redundant variables (with a high correlation among them) will have a small G_S value associated. Therefore, this metric can be used to guide the search for sets of relevant and non-redundant variables.

The CFS approach consists of a search in the feature subset space for a feature subset that maximises the G_S score. As this search is an NP-hard problem, search heuristics are required to