Engineering Applications of Artificial Intelligence 55 (2016) 93–102

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Relevant based structure learning for feature selection

Hadi Zare *, Mojtaba Niazi

Faculty of New Sciences and Technologies, University of Tehran, Iran

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form
21 March 2016
Accepted 6 June 2016
Available online 2 July 2016

Keywords:
Feature selection
Supervised learning
Relevant features
Mutual information
Structure learning
Graphical models

ABSTRACT

Feature selection is an important task in many problems occurring in pattern recognition, bioinformatics, machine learning and data mining applications. The feature selection approach enables us to reduce the computation burden and the falling accuracy effect of dealing with huge number of features in typical learning problems. There is a variety of techniques for feature selection in supervised learning problems based on different selection metrics. In this paper, we propose a novel unified framework for feature selection built on the graphical models and information theoretic tools. The proposed approach exploits the structure learning among features to select more relevant and less redundant features to the predictive modeling problem according to a primary novel likelihood based criterion. In line with the selection of the optimal subset of features through the proposed method, it provides us the Bayesian network classifier without the additional cost of model training on the selected subset of features. The optimal properties of our method are established through empirical studies and computational complexity analysis. Furthermore the proposed approach is evaluated on a bunch of benchmark datasets based on the well-known classification algorithms. Extensive experiments confirm the significant improvement of the proposed approach compared to the earlier works.

1. Introduction

Feature selection (or variable selection) has been considered as a primary step in machine learning, pattern recognition and data mining fields. It is used in a variety of applied domains such as text classification, micro-array analysis and image processing. Nowadays with the explosion of massive online data, choosing an optimal subset of features is a very crucial step (Wang et al., 2014; Tan et al., 2014). While predictive modeling with huge feature sets are common in recent years, it would cause heavy computational burden, interpretation difficulty, and weak results based on curse of dimensionality (Bishop, 2007; Zare et al., 2013). Not only the suitable feature selection process can provide efficient tools to remove irrelevant, redundant and noisy features, but it would improve the speed of learning phase and performance measures of the predictive task too. Based on learning language, the feature selection could be classified into supervised and unsupervised methods. Supervised feature selection approaches are mainly based on the relation between the features and the label to find the optimal feature sets (Tabakhi and Moradi, 2015; Peng et al., 2005; Wu et al., 2013; Zokaei Ashtiani et al., 2014; Wang et al., 2015). On the other hand, finding the optimal feature selection techniques for unsupervised problems are much harder than the supervised one's due to the ambiguous definition of the unsupervised learning, "discovering the interesting patterns from the data" (Murphy, 2012; Feng et al., 2016; Moradi and Rostami, 2015; Tabakhi et al., 2014; Perkins and Theiler, 2003). Here we concentrate on the feature selection for the supervised learning problems.

If we let the original feature set $F = \{f_1, f_2, \ldots, f_p\}$ and the class variable as Y, the aim of feature selection process is to find the optimal subset $S \subset F$ such that it has the best predictive accuracy based on the validation performance criteria. Supervised feature selection process typically can be divided into four primary steps (Liu and Yu, 2005),

(1) Evaluation criteria
(2) Search approaches
(3) Stopping criterion
(4) Validation methods

In the evaluation step, a criterion should be designed carefully to test the relevancy between the selected subset of features and the class variable. Because of the exponential computational complexity of searching through the complete subsets of the original set of features, search procedure for generating candidate subsets of features to evaluate them are devised in search step. The search

* Corresponding author.
E-mail address: h.zare@ut.ac.ir (H. Zare).

http://dx.doi.org/10.1016/j.engappai.2016.06.001
0952-1976/

and evaluation on the candidate subset of features are continued until the stopping criterion holds. Finally the selected feature set usually requires to be validated based on the dataset or prior domain expert knowledge. The evaluation criteria and search strategy are more important than the other steps in a feature selection process.

Based on different evaluation criteria, the feature selection techniques can be generally classified into three main types, the filter, the wrapper and the hybrid methods, Guyon and Elisseeff (2003), Liu and Yu (2005), Dash and Liu (1997) and Liu et al. (2015). The straightforward approach for evaluation criteria is to measure directly the performance of a subset of features based on classification accuracy with the aid of a predictive classifier to select the best subset of features (Kohavi and John, 1997). Although the most effective and optimal approach could be offered in a wrapper model, these techniques suffer from heavy computational burden of training classifier algorithms. The main idea of filter methods is the selection of the optimal features based on statistical or information theoretical evaluation criteria applied on the certain characteristics of the data without requirement of any classification algorithms. The hybrid (embedded) techniques that are somewhat similar but less computationally expensive compared to wrapper methods which measure optimal subset of features through the learning phase. Because of the time consumption of the wrappers and hybrid techniques, the filter methods are highly recommended for dealing with real applications using a variety of evaluation tools such as the Markov blanket based for streaming dataset (Wu et al., 2013), fuzzy-rough sets for feature significance (Maji and Garai, 2013), heuristic relevance based approach (Moradi and Rostami, 2015), divergence criterion (Bressan and Vitrià, 2003) and centrality based influence measure (Moradi and Rostami, 2015).

A variety of techniques are proposed for search strategies, such as exhaustive search (Oliveira and Sangiovanni-Vincentelli, 1992), ranking based among the feature based on the relevancy to the class variable (Lewis, 1992; Geng et al., 2007; Pinheiro et al., 2012). Because of the exponential computation time of exhaustive search approach and ignoring the redundant features in relevant ranking based methods, sequential greedy approaches are proposed to maximize the evaluation criteria in an iterative and incremental development manner (Liu and Motoda, 1998; Guyon and Elisseeff, 2003). Although the traditional forward greedy approaches are commonly used for dealing with huge number of features because of low computational burden and more robustness to over-fitting, they suffer from neglecting the impact of redundancy among features. Some methods (Koller and Sahami, 1995; Peng et al., 2005) have proposed the innovative information theoretic evaluation criteria in a sequential search vein to remedy the aforementioned problems.

Recently some feature selection methods are proposed for massive online dataset, where the number of features are increased with fixed number of observations, streaming cases or incremental observations (Wu et al., 2013; Wang et al., 2014; Wu et al.,) and in these works the evaluation criteria is based on the priorly defined probabilistic and information theoretical concepts in Koller and Sahami (1995) and Peng et al. (2005). The main problems in these recent works could be categorized in threefold, computational burden, streaming setting and optimality criteria.

In this paper we propose a novel feature evaluation criteria in a filter approach based on structure learning and information theoretical tools that can be adopted for streaming dataset and non-streaming dataset. In line with the proposed approach, "*structure learning for feature selection*", hereafter called as *SLFS*, that allowed us to choose more relevant less redundant features carefully within a negligible loss of total feature information, the computation time is reduced compared to the earlier works such as Peng et al. (2005) and Wu et al. (2013).

The structure of the paper is organized as follows. In Section 2, the related works on feature selection are reviewed and a motivation of the basic idea to solve the problem is presented. Section 3 is devoted to the theoretical foundation of the feature selection based on the Markov blanket approach and an overall scheme of our method. The proposed feature selection algorithms and their advantages compared to the previous ones are illustrated in Section 4. We present and describe the experimental results based on the state-of-the art datasets through the SLFS algorithm compared to the earlier works in Section 5. Finally Section 6 discusses the results based on the proposed framework as well as conclusions and future works on the field.

2. Related works and basic idea

2.1. Related works

Because of the importance of the feature selection problem, many researches have been done on various aspects of this fundamental topic. By the availability of massive number of features, reasonable to assume a large subset of features are either irrelevant or redundant for predictive modeling and only a small portion of relevant features yield more effective learning aims (Dash and Liu, 1997; Liu and Yu, 2005). On the one hand, most of the earlier researches have been concentrated on finding relevant features based on the high dependency to the class labels (Peng et al., 2005; Bressan and Vitrià, 2003; Hu et al., 2010). On the other hand, for a wide variety of applications, such as genomic microarray analysis (Xing et al., 2001; Baur and Bozdag, 2016), image representation (Aharon and Elad, 2008) and text categorization (Forman, 2003; Hva et al., 2013), there exist high redundancy among the features. Hence the feature selection algorithm based only on the relevance criteria can be resulted in suboptimal set of features (Kohavi and John, 1997; Pinheiro et al., 2012). There are many research efforts to consider the feature selection criteria with the redundancy and relevancy simultaneously, Markov blanket based approach (Koller and Sahami, 1995; Liu et al., 2015), max-dependency and min redundancy based on mutual information (Peng et al., 2005), and meta-heuristic greedy search (Tabakhi and Moradi, 2015). From the theoretical point of view, Markov blanket framework for feature selection would be yielded to the optimum subset of features and the remaining ones could be considered as redundant features. Because of the exponential computational complexity of finding the Markov blanket subset among the features, there exist a variety of efforts to approximate it such as linear correlation approach (Yu and Liu, 2004) and statistical χ^2-square test (Wu et al., 2013). Those works consider pairwise feature dependency rather than the joint consideration to find the Markov blanket.

2.2. Overall scheme of the idea

First, we define the mutual information between two features x and y,

$$I(x;\, y) = \int_x \int_y P(x,\, y)\log\frac{P(x,\, y)}{P(x)P(y)}dxdy$$
$$= H(x) - H(x|y) \qquad (1)$$

where $H(x)$ is the entropy of the x and $H(x|y)$ is the conditional entropy of x given y (Vergara and Estvez, 2013).The definition (1) of mutual information can be generalized for a vector of features $E = (f_1, f_2, \ldots, f_m)$ and class variable Y as follows: