Information Sciences 318 (2015) 14–27

Contents lists available at ScienceDirect

Information Sciences

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

Feature selection for clustering using instance-based learning by exploring the nearest and farthest neighbors

Chien-Hsing Chen

Department of Information Management, Ling Tung University, Taiwan

ARTICLE INFO

Article history:
Received 21 September 2011
Received in revised form 23 March 2015
Accepted 8 May 2015
Available online 14 May 2015

Keywords:
Feature selection
Instance-based learning
Neighbor
Mutual information
Clustering

ABSTRACT

Feature selection for clustering is an active research topic and is used to identify salient features that are helpful for data clustering. While partitioning a dataset into clusters, a data instance and its nearest neighbors will belong to the same cluster, and this instance and its farthest neighbors will belong to different clusters. We propose a new Feature Selection method to identify salient features that are useful for maintaining the instance's Nearest neighbors and Farthest neighbors (referred to here as FSNF). In particular, FSNF uses the mutual information criterion to estimate feature salience by considering maintainability. Experiments on benchmark datasets demonstrate the effectiveness of FSNF within the context of cluster analysis.

1. Introduction

The selection of salient features is an important issue in cluster analysis and is relevant for topics such as image representation [24], time-series prediction [45], machine learning [41] and natural language processing [5]. Typically, the use of a large number of features to represent a pattern is highly informative for a learning algorithm. However, in a high-dimensional dataset, some features are noisy; thus, learning algorithms are often biased by noisy features that affect the learning process. The goal of feature selection for clustering is usually to identify a subset of salient features from the original representation space; the identified salient features are helpful for data clustering that aims to maximize the between-cluster scatter and minimize the within-cluster scatter. A previous semi-supervised learning method [4] utilized the pairwise constraints between instances with class labels to identify salient features. Because the class labels would usually be unavailable in a real-world dataset, we consider an unsupervised learning method for the selection of salient features applied for data clustering.

Feature selection algorithms are classified into two primary categories: those based on the filter model and those based on the wrapper model [1,27]. The filter model requires an evaluator to measure the intrinsic characteristics of each feature. One of the most well-known criterion functions for evaluating features uses a relevance metric [37,41]; a feature that is either dependent (relevant, consistent, reliable, important or informative) on the target class label or conditionally independent of the other features is usually defined as a relevant (salient) feature. However, a feature subset selected by the filter model is problematic with respect to achieving the goal of feature selection for clustering because the number of clusters or the clustered structure cannot be effectively predicted in advance. This problem can be easily overcome when we have prior information implying that instances having the same class labels will belong to the same clusters; however, such information is unlikely to be available for a real-world dataset.

E-mail address: ktfive@gmail.com

http://dx.doi.org/10.1016/j.ins.2015.05.019
0020-0255/

The wrapper model requires a pre-specified classification (or clustering) learning algorithm trained on data instances and an evaluator to quantify the performance in terms of a generalization criterion (e.g., accuracy or between-cluster scatter). Using heuristic algorithms (e.g., forward or backward selection), the goal is to effectively visit the space of all possible feature subsets to identify the best feature subset, which is usually a local-optimal solution. However, the selected feature subset suffers from two biases. First, the feature subset depends strongly on the selection of the learning algorithms because different clustering algorithms produce differently shaped clusters [19] and favor different prior preferences (e.g., the specification of the number of clusters in *K*-means, visual surface inspection in a self-organizing map (SOM) and a hierarchy of clusters represented as a dendrogram in hierarchical clustering). We would be less likely to trust a feature subset selected by biased clusters, as calculated by an arbitrary clustering algorithm, when we lack prior information (such as natural shape, number of clusters and clustered structure), as is the case with a real dataset. Second, many studies based on the wrapper model for unsupervised learning partition a dataset into clusters in advance, using all of the features. Consequently, the selected feature subset identifies biased clusters due to their noisy features. Such shortcomings directly affect the flexibility of wrapper-based feature selection methods because the salient feature subset that results from a particular clustering algorithm may not be appropriate for use with another clustering algorithm.

In this paper, we present a new method to achieve the goal of feature selection for clustering without the need to explore the exact clustered information. Specifically, FSNF uses the mutual information criterion to identify salient features due to its robustness and popularity. The nearest and farthest neighbors help select the salient features; FSNF uses the mutual information criterion to assess features while considering these neighbors based on distinguishability and redundancy toward a robust statistical evaluator. FSNF then identifies a real-valued salience vector instead of heuristically visiting the space of all possible feature subsets. Once the salience vector, which is usually a local-optimal solution, is obtained, the salient features are used to perform clustering using learning algorithms.

The rest of this paper is organized as follows. Section 2 reviews work related to feature selection. Section 3 describes FSNF, which attempts to identify the best feature salience vector. Experimental results are presented in Section 4. Section 5 presents a discussion of the results and concluding remarks.

2. Related work

Feature selection for clustering is an active topic in the data-mining field. In general, feature selection algorithms are classified into two categories: those based on the filter model [8] and those based on the wrapper model [27]. A number of feature selection methods based on the wrapper model for classification [30,38] and clustering [52] have been proposed in the literature. In supervised learning, one can generally use the wrapper model to construct a classifier and use the criterion (e.g., accuracy) to observe how the features can positively predict class labels via this classifier [10,35,36,40]. Maldonado and Weber [31] introduced a new wrapper-based method in which the classifier was built using support vector machines with kernel functions. Su and Hsiao [42] developed a system for simultaneous multiclass classification and feature selection. Wang et al. [49] found the smallest set of genes that can ensure the highly accurate classification of cancers from microarray data using supervised machine learning algorithms. Additionally, various studies have identified salient features that are class-dependent [29,33,37] or class-separable [48,54] features. Nicolas et al. [11] proposed a new method for instance and feature selection based on supervised learning.

Class labels may be absent in a real-world dataset. Consequently, the wrapper model is integral to the field of unsupervised learning, and nonparametric techniques can be applied to many interesting problems [44]. Yang et al. [52] presented a feature selection method for selecting features by minimizing the Bayes error rate estimated with a nonparametric estimator. Lin et al. [25] also introduced a nonparametric technique for feature screening. Chen [4] developed a nonparametric technique using pairwise instance constraints to supervise the feature selection process.

Another option is to develop a clustering algorithm to partition a dataset into clusters and use internal or relative cluster validity as the criterion to observe whether a feature is informative with respect to the clustered information [23]. Chow et al. [7] proposed a selection method that considered the compactness and separation of clusters. Huang et al. [17] proposed a feature co-selection method for Web document clustering in which the clustering results in one type of feature space helped identify salient features in other types of feature spaces. Li et al. [22] proposed a new text-clustering method with feature selection and extended the chi-square term-category independence test to measure whether the dependency between a term and a cluster was positive or negative. Sanguinetti [39] presented a latent variable model to perform dimensionality reduction on a dataset that contained clusters; specifically, a variable was considered salient when it preserved clustered information by mapping an original representation space to a latent space. Mitra [32] proposed using structural similarity between clusters for feature selection, where the topological neighborhood information about pairs of instances was considered to assess the similarity. Furthermore, a feature selection method based on the wrapper model in semi-supervised feature selection, considering labeled and unlabeled examples, has also been described [50,53].

The filter model does not require a clustering (or classification) learning algorithm to provide cluster information in terms of cluster shape and number of clusters. It does require an evaluator to measure the intrinsic characteristics of each feature. Han et al. [14] presented a new criterion function that identifies features pertinent to the classification task at a very low computational cost. Chen [3] selected salient features using compactness and separability. Peng et al. [37] developed criteria based on mutual information for feature selection. Recently, a hybrid model that captures the advantages of the filter and wrapper models was developed to address the above-mentioned computational issue [27,55].