Knowledge-Based Systems 67 (2014) 328–341

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A non-parametric feature assessment mechanism by identifying representative neighbors for image clustering

Chien-Hsing Chen *

Department of Information Management, Ling Tung University, Taiwan

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form 24 January 2014
Accepted 17 April 2014
Available online 5 May 2014

Keywords:
Feature assessment
Representative neighbors
Image clustering
Sensitivity restriction
Scalability restriction

ABSTRACT

Unsupervised feature selection methods based on a non-parametric model usually focus on using the neighbors of each point for identifying salient features, which are helpful for performing clustering to satisfy both within-cluster and between-cluster scatter criteria. However, these methods usually suffer from two restrictions, i.e., sensitivity restriction and scalability restriction, in choosing the neighbors and determining the number of neighbors. In this paper, we propose a new non-parametric mechanism based on unsupervised learning for feature assessment and selection in image clustering. Our mechanism potentially overcomes the existing restrictions by identifying representative neighbors that primarily divide a dataset into subsets. We subsequently present a new solution used to minimize the number of representative neighbors by searching for a hyperplane that separates two linearly separable and neighboring clusters for which the distances of the representative neighbors from the hyperplane are minimal. We finally present a wrapper-based method that uses a backward strategy from our feature assessment and selection process to consider these representative neighbors. In tests on benchmark image datasets, the experimental results indicate that our method performs better in terms of relative cluster validations and statistical hypothesis testing than mutual information statistics for both discovery of interesting patterns and selection of features for cluster analysis.

1. Introduction

Feature selection based on unsupervised learning is an active research topic in the data-mining field and is often aimed at selection of salient features (or attributes) that are effective for discovering natural clusters while satisfying the within-cluster scatter (e.g., the points within a cluster are similar) and between-cluster scatter criteria (e.g., the points between clusters are different). The literature describes many feature selection applications for pattern recognition [1], gene expression [2], image annotation [3] and text clustering [4]. We assume that there exists a gray-scale image dataset that must be divided into clusters and that the size of each image is 50×100 pixels. The size of the vector used to represent this image is $50 \times 100 \times 256 = 1{,}280{,}000$. The clusters are built using all features, which include noise and thus usually suffer from weak interpretability of the cluster results because the instances in this image dataset have high data dimensionality. Most existing technologies for identification of salient features use a variance metric to evaluate an intrinsic variable. For example, a random variable with a large variance will increase only the between-cluster scatter, whereas another variable that follows a uniform distribution will increase only the within-cluster scatter [5]. Recently, non-parametric methods that consider the neighbors of each point have offered new options for feature selection [6–9].

Using the non-parametric model to consider the neighbors of each point for identification of salient features, our previous method [10] based on a filter model studied the basic characteristic of clustering by assuming that an instance usually belongs to a cluster that includes its nearest neighbors and belongs to different clusters than its farthest neighbors. Without the aid of any clustering learning algorithm to provide cluster information (e.g., cluster shape and number of clusters) for training, the salient features are thus selected from those that maximally satisfy this characteristic. We later proposed a new feature evaluator [5] that considered compactness (i.e., the average similarity from an instance to its nearest neighbors) and separability (i.e., the average distance from an instance to its farthest neighbors) to produce a feature salience vector. Furthermore, side information [6,7] (e.g., pairwise constraints between instances) was used to purify the nearest and the farthest neighbors to achieve feature assessment [11]. The nearest and the farthest neighbors appear to be helpful in

* Tel.: +886 912311895.
E-mail address: ktfive@gmail.com

http://dx.doi.org/10.1016/j.knosys.2014.04.026
0950-7051/

identification of salient features, which further aids in satisfying the within-cluster scatter and between-cluster scatter criteria in the clustering process.

Although many non-parametric methods that consider neighbors for feature assessments have been proposed, most suffer from two restrictions, i.e., sensitivity restriction and scalability restriction. First, because neighbors are usually chosen to observe how the features describe the data instances in a cluster or in different clusters, the optimal set of neighbors can be generalized such that each instance and its nearest neighbors will be included in a cluster, and each instance and its farthest neighbors will be located in different clusters. These chosen neighbors must thus be quite sensitive to feature assessment because different neighbors must result in different salient features. Second, because the number of these neighbors must be an empirical parametric value, the above-mentioned non-parametric methods would thus be non-scalable when the number of neighbors should be large. We assume a set of n data instances, and for each, we consider K neighbors. If K is sufficiently large (e.g., approximately n), the computational time involved in choosing these neighbors for feature assessment will be approximately $O(n^2)$, making data storage notably expensive.

In this paper, we propose a new non-parametric mechanism for feature assessment and selection based on unsupervised learning. This mechanism overcomes the existing restrictions with respect to two major components. First, we obtain clusters for a dataset and search for the representative neighbors that divide this dataset by clustering into subsets (i.e., clusters). In this specific use, we build a graph composed of a set of vertices and a set of weighted edges, which are connected to estimate the distances between the vertices. In this way, the estimated distances involve summing up the weights on the edges by traversing the vertices in the graph. We therefore obtain a distance matrix whose size is $n \times n$. This matrix is subsequently input to the visual assessment of cluster tendency (VAT) method [12,13], which is effective for estimating the number of clusters. The clusters are then obtained using a clustering learning algorithm with respect to the estimated number of clusters. The representative neighbors for each instance are thus chosen and are guaranteed to be positioned on the border of those clusters.

Second, we present a new solution for minimizing the number of representative neighbors. We assume the existence of a hyperplane that has a maximal margin between two linearly neighboring clusters. The support vectors on the two sides of the separating hyperplane must exhibit the shortest distance among all of the distances between the instances on the different sides. We thus choose two representative neighbors from the two opposing sides such that the distance between these two neighbors is minimal. Our solution considers only two representative neighbors and thus should be more scalable because the computational time required to choose representative neighbors for our feature assessment becomes $O(2n)$. Finally, we present a wrapper-based method that uses a backward strategy with consideration of representative neighbors for feature assessment and selection.

This paper is organized as follows. Section 2 introduces the background studies related to neighbor-based learning for feature selection. Section 3 explains our feature assessment and selection methods, and Section 4 presents the experimental results. A brief discussion is given in Section 5, and finally, Section 6 details the conclusions.

2. Background study on neighbor-based learning for feature assessment

We study selected state-of-the-art methods with respect to three major components that are closely related to the development of our method. First, we study a neighbor-based learning method that chooses neighbors for feature assessment. To improve the effectiveness of choosing those neighbors, we use graph theory to build a graph for estimating the distances between instances. Finally, we apply the neighbor-based learning method with consideration of clusters for choosing neighbors and determine the number of neighbors, a method that suffers from sensitivity and scalability restrictions. These three components stimulate us to develop extensive solutions that overcome these restrictions in feature assessment by identifying representative neighbors.

2.1. Feature selection based on neighbor-based learning

In this section, we study the background of neighbor-based learning for feature assessment. Assume that we have a dataset X, including n data instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = [x_{1,i}, \ldots, x_{j,i}, \ldots, x_{d,i}]^T$ denotes the ith instance with d dimensions. We follow our previous method [11] with revisions to briefly consider the nearest and the farthest neighbors and to find a new feature space that results from a feature vector $\mathbf{w}$ such that the loss between two sets of neighbors in this new space is minimized. Let us assume that a hyperplane exists with a maximal margin between two sets of neighbors. The support vectors on the two sides of the separating hyperplane have the shortest distance between the instances on the different sides among all of the distances. We thus define a margin-based loss function for the neighbors to obtain the support vectors of $\mathbf{x}_i$, and the function is written as follows:

$$\rho(\mathbf{x}_i) = \underset{\substack{l\in\{1,\ldots,L\}\\ k\in\{1,\ldots,K\}}}{\operatorname{argmin}}((\mathbf{w}^T \cdot \varepsilon(\mathbf{x}_i, \mathbf{x}_{i\to l}^{\phi})) - (\mathbf{w}^T \cdot \varepsilon(\mathbf{x}_i, \mathbf{x}_{i\to k}^{\Theta}))) \quad (1)$$

where $\rho(\mathbf{x_i})$ measures the margin and has the shortest distance between two sets of neighbors for $\mathbf{x}_i$ in the new feature space. The function $\varepsilon(.,.)$ is an element-wise absolute operator (e.g., $\varepsilon([0.2, 0.5]^T, [0.3, 0.8]^T) = [0.1, 0.3]^T$) for which the Manhattan metric is used to compute the distance, ϕ represents the farthest operator, and Θ represents the nearest operator. In this work, $\mathbf{x}_{i\to l}^{\phi}$ is the lth farthest neighbor, and $\mathbf{x}_{i\to k}^{\Theta}$ is the kth nearest neighbor of $\mathbf{x}_i$. Both L and K represent the number of neighbors. The number of farthest neighbors or nearest neighbors is defined according to $\pi(l) = \sum_{r=1,r\neq i}^{n} I((dist(\mathbf{x}_i, \mathbf{x}_r) \geqslant dist(\mathbf{x}_i, \mathbf{x}_{i\to l}^{\phi}))$ or $\psi(k) = \sum_{r=1,r\neq i}^{n} I((dist(\mathbf{x}_i, \mathbf{x}_r) \leqslant dist(\mathbf{x}_i, \mathbf{x}_{i\to k}^{\Theta}))$, where $\pi()$ and $\psi()$ are used to choose the number of neighbors. The function $dist(.,.)$ measures the distance between instances, and $I[]$ outputs 1 when the condition is satisfied and assigns a value of 0 otherwise.

We briefly illustrate an example to represent feature salience. For example, we obtain the first item $\varepsilon(.,.) = [0.6, 0.1]^T$ and the second item $\varepsilon(.,.) = [0.1, 0.7]^T$ under $\mathbf{w} = [0.5, 0.5]^T$. To observe $[0.6, 0.1]^T$, we note that the first dimension primarily drives the data into different clusters (i.e., the between-cluster distance is large). To observe $[0.1, 0.7]^T$, we note that the first dimension leads to the data located in a single cluster (i.e., the within-cluster distance is small). The first dimension is thus more salient than the other because a large $\varepsilon(\mathbf{x}_i, \mathbf{x}_{i\to l}^{\phi})$ would maximize the between-cluster distances and a small $\varepsilon(\mathbf{x}_i, \mathbf{x}_{i\to k}^{\Theta})$ would minimize the within-cluster distances.

2.2. Searching for neighbors while violating the triangle inequality

We implement graph theory to estimate the distances between instances by searching for neighbors $\mathbf{x}_{i\to l}^{\phi}$ and $\mathbf{x}_{i\to k}^{\Theta}$. The graph is composed of vertices and weighted edges. The vertices represent data instances, and the weighted edges represent the distances between any two vertices calculated by the distance metric. Use of the built graph should be effective in searching for those neighbors because it is commonly used to overcome the shortcomings of