Knowledge-Based Systems 89 (2015) 641–653

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A manifold learning framework for both clustering and classification

Weiling Cai *

Department of Computer Science & Technology, Nanjing Normal University, Nanjing 210097, PR China

ARTICLE INFO

Article history:
Received 12 February 2015
Received in revised form 4 September 2015
Accepted 8 September 2015
Available online 14 September 2015

Keywords:
Pattern recognition
Clustering learning
Classification learning
Bayesian theory
Manifold Learning

ABSTRACT

In recent years, a great deal of manifold clustering algorithms was presented to identify the subsets of the manifolds data. Meanwhile, numerous classification algorithms were also developed to classified data shaped in the form of manifold. However, nearly none of them pay attention to the statistical relationship between the manifold structures and class labels, thus failing to discover the knowledge concealed in data. In this paper, a manifold learning framework for both clustering and classification is presented, which involves two steps. In the first step, the clustering through ranking on manifolds is executed to explore structures in data; in the second step, the class posterior probability is calculated by using the Bayesian rule. The core of this framework lies in employing the Bayesian theory to establish the relationship between manifolds and classes thus creates a bridge between clustering learning and classification learning. Our new manifold learning framework is interesting from a number of perspectives: (1) our algorithm can perform manifold clustering learning which can auto-determine the clustering parameters without manual determining; (2) our algorithm can perform manifold classification learning which models the posterior probabilities $p(\omega_l|x_i)$ by using the Bayesian rule; (3) our algorithm can provide the statistical relationship between the manifold structure and the given classes. Encouraging experimental results are obtained on 2 artificial and 16 real-life benchmark datasets.

1. Introduction

Data mining [1–3] is the discovery of interesting relationships and characteristics that may exist implicitly in data. Clustering and classification are two primary data-mining techniques [4,5]. The clustering approaches such as K-Means [6], Fuzzy C-Means (FCM) [7] and Gaussian Mixture Model [8] are widely utilized to discover the hidden structure in data. Whereas the classification approaches such as Multi-layer Perceptron (MLP) [9] and Support Vector Machines (SVM) [10,11] are successfully applied to determine the class labels of unseen samples. To fuse the advantages of clustering and classification together, numerous researchers studied on how to design a single approach for both clustering and classification. To bridge clustering and classification, Setnes and Babuŝka [12] proposed Fuzzy Relational Classifier (FRC) which attempted to utilize the fuzzy composite operators to construct the relationship between the cluster structures and classes. To enhance the robustness of FRC, in one of our previous works, we developed Robust Fuzzy Relational Classifier (RFRC) [13] by replacing FCM and hard class labels with Kernelized FCM (KFCM) [14,15] and soft labels, respectively. Another famous classifier is Radial Basis Function neural networks (RBFNN) [16,17] which extracts significant information from the observed data to construct its hidden layer.

However, all above algorithms are relatively suitable for the data shaped in the form of point clouds (group), but unsuitable for those data in the form of manifold structure. In real-life world, there are quite a number of data that form paths through a high-dimensional and expose manifold structure. For instance, motion segmentation problem in computer vision, the point correspondences in a dynamic scene can generally be represented as manifold; in classification of face images, the faces of person lie on the manifold. For these data exhibiting manifold structure rather than compact shape, a considerable number of clustering algorithms such as Spectral Clustering [18,19] have been presented to identify the subsets of the manifolds data. Numerous research studies proved that incorporating the structure information into a classifier can enhance its generalization ability, and this research finding is consistent with the famous No Free Lunch (NFL) theorem [20]. In the last decade, a number of manifold or subspace classification algorithms such as Plane-Gaussian Function Networks (PGFN) [21], Laplacian Regularized Least Square Classification (LapRLSC) [22,23] and Laplacian SVM (LapSVM) [24] were presented. These algorithms only attempt to integrate the manifold or subspace distribution information into the classification model.

* Tel./fax: +86 25 84441669.
E-mail address: caiwl@nuaa.edu.cn

http://dx.doi.org/10.1016/j.knosys.2015.09.010
0950-7051/

However, nearly none of them pay attention to the underlying relationship between the manifold distribution and given classes, thus unable to discover the knowledge concealed in data. As a result, an open and challenging problem is to design a framework for manifold data with the goal of combining the advantages of clustering and classification and meanwhile revealing the statistical relationship between manifolds and classes.

In this paper, we propose a manifold learning framework for both clustering and classification (MCC). MCC aims to discover the manifold structure hidden in data, design an effective and transparent classification mechanism and meanwhile exploit the relationship between manifolds and classes. To achieve these goals, our framework treats the manifold clustering learning and classification learning in a two-step *sequential* manner. In the first step, the clustering through ranking on manifolds is performed to explore structures in data; in the second step, by using the Bayesian rule, the class posterior probability is calculated to give class labels for unseen samples. It is worth mentioning that the number of manifolds (i.e. clusters) has a significant influence on the result of manifold clustering [25–27]. To auto-determine this parameter in our algorithm, the inter-cluster mean distance by ranking on manifolds is maximized and while the intra-cluster mean distance is minimized. As a result, our algorithm can auto-determine the clustering parameters without manual determining. Another key of this framework is to connect the multi-manifold with the given classes employ, and then establish a relationship between them. This relationship creates a bridge between clustering learning and classification learning. Based on such relationship, our framework cannot only group multi-manifold into different clusters, but also make classification decisions for unseen samples. More importantly, this relationship can successfully reflect the probability and statistics meaning between manifold structures and given classes, so that we gain some meaningful insights to make MCC prone to be transparent.

The new manifold learning framework for both clustering and classification is interesting from a number of perspectives:

(1) Our algorithm can perform manifold clustering learning which can auto-determine the clustering parameters without manual determining.
(2) Our algorithm can perform manifold classification learning which models the posterior probabilities $p(\omega_l|x_i)$ by using the Bayesian rule.
(3) Our algorithm can provide the statistical relationship between the manifold structure and the given classes.

The experimental results on both synthetic and real-life datasets all demonstrate the effectiveness and potential of MCC.

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3 describes the proposed manifold learning framework for both clustering and classification. Preliminary experimental results are shown in Section 4. Finally, we give concluding remarks and future work in Section 5.

2. Related works

There have been several recent related works to inherit the merits of both clustering and classification learning. We review the main works as follows.

2.1. Fuzzy relational classifier

Fuzzy Relational Classifier (FRC) [12] was proposed to provide a transparent alternative to the black-box techniques such as neural networks. As show in Fig. 1, in FRC, FCM is firstly adopted as the clustering criterion to discover the natural structure in data, and its objective function is as follows:

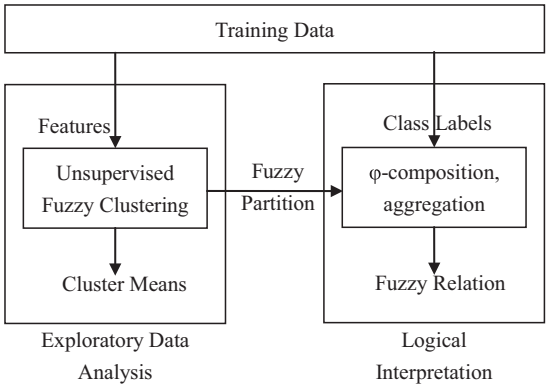

Fig. 1. Training process of FRC and RFRC.

$$J_{FCM}(U,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ji}^{2}\|\mathbf{x}_i-\mathbf{v}_j\|^2, \tag{1}$$

where $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ and $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_c\}$ are the training samples and cluster centers, respectively; and u_{ji} is the fuzzy memberships of $\mathbf{x}_i$ to $\mathbf{v}_j$. By definition, each sample $\mathbf{x}_i$ satisfies the constraint $\sum_{j=1}^{c}u_{ji}=1$. And then, a relation matrix $\mathbf{R}$ is computed for the obtained fuzzy partition and the given hard class labels. In FRC, FCM is unable to group the datasets consisting of the non-spherical clusters, so that the interpretation of the clustering or classification results may be biased.

Afterwards, we have presented Robust FRC (RFRC) [13] to improve both clustering and classification performance of FRC in our previous work. Specifically, in the clustering phase, the robust Kernelized FCM (KFCM) [14,15] is adopted to replace FCM which can be described as below:

$$J_{KFCM}(U,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ji}^{m}\|\phi(\mathbf{x}_i)-\phi(\mathbf{v}_j)\|^2, \tag{2}$$

where ϕ is an implicit nonlinear map from the input space to a rather high dimensional feature space. Compared to FCM, KFCM based on RBF kernel is a robust estimator according to M-estimator and is more flexible for clustering non-spherical data. Next, in the classification phase, the soft class label motivated by the fuzzy k-nearest-neighbor [28] is employed to replace the hard class label. With the incorporation of both KFCM and the soft class labels, RFRC makes the constructed relation matrix $\mathbf{R}$ more really reflect the relationship between the classes and clusters, and thus significantly boosts the performance of FRC.

It is worth to point out that in FRC and RFRC, the entries in the relation matrix $\mathbf{R}$ lack the statistical meaning, thus it is difficult to judge whether the obtained relationship is really reliable.

2.2. Radial basis function neural networks

Radial Basis Function neural networks (RBFNN) [16,17], as shown in Fig. 2, is a feed-forward multi-layer network. It usually consists of three layers: input layer, hidden layer and output layer. Each basis function Φ_k corresponds to a hidden unit and w_{kl} represents the weight from the kth basis function or hidden unit to the lth output units.

In the training phase of RBFNN, the basis function Φ_k for each hidden node can be determined by

$$\Phi_k^{RBF}(\mathbf{x},\mathbf{v}_k)=\exp\left(\frac{-\|\mathbf{x}-\mathbf{v}_k\|^2}{2\sigma^2}\right), \tag{3}$$