Neurocomputing 139 (2014) 423–431

Contents lists available at ScienceDirect

Neurocomputing

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

Representative selection based on sparse modeling

Yu Wang [a,b], Sheng Tang [a], Yong-Dong Zhang [a,*], Jin-Tao Li [a], Dong Wang [c]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] Graduate University of Chinese Academy of Sciences, Beijing 100190, China
[c] Huawei Technologies Co., Ltd, Beijing, China

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form
22 January 2014
Accepted 5 February 2014
Communicated by M. Wang
Available online 13 April 2014

Keywords:
Representative selection
Sparse modeling
Dictionary learning

ABSTRACT

Selecting representatives for multimedia analysis applications could greatly reduce the time and memory consumption. Many representative selection methods have been proposed to select a subset from the database as representatives. However, current methods cannot guarantee that the selected subset could represent the global distribution of the entire dataset. In order to evaluate how well the subset represents the global distribution of the whole dataset, we use the distance metric: Kullback–Leibler (KL) divergence between the distribution of the fake dataset reconstructed from the subset and the distribution of the true dataset. In this work, we propose a sparse modeling based method to select representatives. The proposed method formulates the representative selection problem as a discrete dictionary learning problem. Based on the assumption that the dataset can be approximately reconstructed by linear combinations of dictionary items, we design a two-step iterative representative selection algorithm, which can minimize this KL divergence. Experiments evaluate the proposed algorithm in several multimedia analysis applications, including image and video summarization, classification using representative images and classification using representative features, and our method is shown to outperform state-of-the-art methods.

1. Introduction

Multimedia content analysis and understanding is a fundamental research problem. Recent studies have shown that large-scale training dataset could greatly benefit the performance and beat the clever learning algorithm [1,2]. However, dealing with massive dataset is time and memory consumption. In fact multimedia data are very redundant, and could be represented well by a relative small subset [3–6], e.g. key frames in videos and prototypes in image set. Using a relatively small representative subset for multimedia analysis applications can greatly reduce the memory requirement and computational time, and how to select representative subset is the issue of representative selection problem.

Many algorithms have been proposed for representative selection problem. According to whether class labels of training data are available, these algorithms can be roughly grouped into two families, i.e. supervised and unsupervised representative selections. Generally speaking, supervised representative selection usually yields more reliable performance [3,7,8]. Given sufficient labels, it is possible for supervised representative selection to find a discriminative subset, and thus better performance can be obtained with this subset than with the whole dataset [8]. However, for real world applications most of the training data have not been labeled. Therefore, unsupervised representative selection turns out to be a more general way. Although unsupervised representative selection does not always improve the performance, it can still improve the efficiency like the supervised one.

Most prior unsupervised representative selection methods are based on following assumptions [4]: representatives are either in a low-dimensional space or distributed in high density region. When data meet the low-rankness assumption, Rank Revealing QR algorithm [9] can select a few data points by finding a permutation of data matrix that gives the best conditioned submatrix. Randomized algorithm [10] and Greedy algorithm [11] have also been proposed for the column selection problem from a low-rank matrix. Methods following the second assumption select data centers that are mainly located in high density region as representatives. Kmedoids [12] can be considered as a variant of Kmeans. Similar to Kmeans it is an iterative algorithm to find cluster centers, but those centers of Kmedoids are selected from data points. When similarities between pairs of samples are given, Affinity Propagation [13] uses a message passing algorithm to find data centers. However, if the selection method only focuses on high density region, it would under-represent discriminant medium density and low density ones. As pointed in [14] medium density region is highly effective for image

[*] Corresponding author. Tel.: +86 10 62600666; fax: +86 10 62601356.
E-mail addresses: wangyu@ict.ac.cn (Y. Wang), ts@ict.ac.cn (S. Tang), zhyd@ict.ac.cn (Y.-D. Zhang), jtli@ict.ac.cn (J.-T. Li), dave.wangdong@huawei.com (D. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.02.013
0925-2312/© 2014 Elsevier B.V. All rights reserved.

classification task. Data points in low density region have been proven to be informative [15,16]. In addition, those low density points on class boundaries are the answers which SVM (Support Vector Machine) is looking for [17]. Thus, We claim that a good subset should represent all the regions of the dataset distribution, not only the high density region of the dataset distribution.

Random representative selection does not rely on prior assumptions, and has shown superior performance in image classification tasks [18]. Recently the proposed method [4], which formulates the problem of representative selection as a sparse multiple measurement vector problem, is also assumption-free. However, those methods cannot guarantee to represent the global distribution of the whole dataset.

In this paper, we propose a sparse modeling based method to find a compact dictionary, whose items are representatives. The whole dataset could be approximately reconstructed by linear combinations of items of dictionary. In order to evaluate how well the subset represents the global dataset distribution, we use the distance metric: Kullback–Leibler (KL) divergence which is a measure of the difference between two probability distributions [19]. By minimizing the KL divergence between the distribution of the reconstructed dataset and the distribution of the true dataset, the global distribution of the whole dataset could be represented well by the subset. Thus, representative points selected by the proposed algorithm are not necessarily data centers in high density region, and many representatives would be selected to represent medium density and low density regions. Our work has following contributions with respect to the state of the art:

- We do not assume representatives are either in a low-dimensional space or distributed in high density region. We design a representative selection algorithm to represent the global distribution of the dataset, including medium density and low density regions.
- We formulate the representative selection problem as a discrete dictionary learning problem, and propose a discrete dictionary learning algorithm.
- We demonstrate the proposed algorithm in several multimedia analysis applications, including image and video summarization, classification using representative images and classification using representative features.

2. Related works

This work is closely related to active learning, sparse modeling and unsupervised representative selection.

2.1. Active learning

Active learning is a machine learning technique that selects the most informative samples for labeling and uses them as training samples [20]. A typical active learning system contains two parts, that is, a learning engine and a sample selection engine. The learning engine can adopt any existing classification algorithm, while the sample selection engine should be designed according to certain strategy. Generally used sample selection strategies in active learning are as follows: risk reduction, uncertainty, diversity, density and relevance. The first two strategies are learner related. Risk reduction is consistent with the learner to reduce the expected risk of labeling unlabeled samples [21]. Uncertainty criterion means to select those samples whose predicted labels are most uncertain using the learner [22]. The third and fourth strategies are data related. Diversity criterion requires that the selected samples should be diverse [23]. Density criterion indicates that the samples within high density region should be selected [24,25]. The last strategy, relevance criterion, is usually applied in multi-label tasks, and in these tasks samples that are relevant to the given concept or query should be chosen [26]. The proposed method is different from current active learning since it does not need the learning engine and aims to select representatives that can represent the global distribution of the dataset.

2.2. Feature selection

Not all features are important, i.e. some features may be redundant, irrelevant and noisy. Thus, feature selection aims to determine a minimal feature subset while retaining a suitably high accuracy in representing the original feature. There are two different types of feature selection approaches [27]: those which maximize clustering performance using an index function, and those which consider features to preserve the geometrical structure of the data space. The first category includes sequential unsupervised feature selection algorithm [28], maximum entropy based method [29] and the recently proposed rough set based unsupervised feature selection algorithm [30]. The second category of approaches selects the most representative features which can best preserve the geometrical structures of the data space. Laplacian score algorithm [31] and its extensions [32] have been proposed to select those features which can reflect the manifold structure of the data space. Recently the proposed algorithm [27], motivated from experimental design, selects a feature subset by minimizing the size of the parameter covariance matrix of the regularized regression model, to reflect the underlying manifold structure and at the meantime improve the learning performance. The feature selection aims at reducing the high dimensionality of the dataset in the feature-space. Different from it, the proposed method aims to solve another important problem related to large datasets, which is to find a subset of the data that appropriately represents the whole dataset in the object-space.

2.3. Sparse modeling methods

Recent researches have shown that sparsity can help to improve the performances of various machine learning problems [33,34]. Considering a set of data points in R^m arranged as columns in data matrix $X = \{x_1, \ldots, x_n\}$, the formulation of sparse modeling is as follows:

$$\min_{D,a} \sum_{i=1}^{n} \|x_i - Da_i\|_2^2, \quad \text{s.t. } \|a_i\|_0 \le s, \tag{1}$$

where in sparse modeling D is named dictionary, a_i is the named sparse representation and s is the maximum number of nonzero items in a_i. According to how dictionary is formed, existing sparse modeling methods can be generally classified into following categories:

(1) Entire dataset is formed as dictionary, and then each data point is sparsely represented by a linear combination of the rest [35,36].

(2) Items of dictionary is learnt from data. Dictionaries, learned from different class data, are used for clustering problem [37]. Transfer learning task builds a common dictionary from a few datasets to find new features [38]. Using online optimization based on stochastic approximation to learn dictionary from large-scale dataset is suitable for large-scale task [39]. K-SVD algorithm [40] uses SVD decomposition of the error matrix to learn dictionary from redundant signals.

Our method is different from current methods, since our dictionary is selected from data. In implementation, our algorithm can be considered as a variant of K-SVD, with an additional