Neurocomputing 153 (2015) 41–53

Contents lists available at ScienceDirect

Neurocomputing

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

Label propagation based on local information with adaptive determination of number and degree of neighbor's similarity

Seyed Alireza Saffari *, Abbas Ebrahimi-Moghadam

Electrical Engineering Department, Ferdowsi University of Mashhad, Iran

ARTICLE INFO

Article history:
Received 14 April 2014
Received in revised form
21 November 2014
Accepted 22 November 2014
Available online 2 December 2014

Keywords:
Semi-supervised learning algorithms
Graph-based algorithms
Sparse representation
Local information

ABSTRACT

In many practical applications of machine vision, a small number of samples are labeled and therefore, classification accuracy is low. On the other hand, labeling by humans is a very time consuming process, which requires a degree of proficiency. Semi-supervised learning algorithms may be used as a proper solution in these situations, where ε-neighborhood or k nearest neighborhood graphs are employed to build a similarity graph. These graphs, on one hand, have a high degree of sensitivity to noise. On the other hand, optimal determination of ε and k parameters is a complex task. In some classification algorithms, sparse representation (SR) is employed in order to overcome these obstacles. Although SR has its own advantages, SR theory in its coding stage does not reflect local information and it requires a time consuming and heavy optimization process. Locality-constrained Linear Coding (LLC) addresses these problems and regards the local information in the coding process. In this paper we examine the effectiveness of using local information in form of label propagation algorithm and present three new label propagation modifications. Experimental results on three UCI datasets, two face databases and a biometric database show that our proposed algorithms have higher classification rates compared to other competitive algorithms.

1. Introduction

Data classification is of interest to machine learning researchers. Many classification algorithms are offered and developed by the researchers. Usually, in these classification algorithms for test or unlabeled samples, we use training or labeled samples. Unfortunately, when the number of training samples is far less than the test samples, these methods perform poorly. In many practical applications in machine learning, the number of labeled samples is quite low, while a large number of samples are unlabeled. Therefore, a large number of samples must be labeled and used as training samples. Labeling process by human is a time consuming task, which requires skilled hand work. In this condition an appropriate approach is to employ both labeled and unlabeled samples for data classification. In semi-supervised learning, which is an active topic in machine vision [1–10], labeled and unlabeled samples are both employed. Since many unlabeled samples can be gathered only by measuring them without interpretation, semi-supervised learning methods are very useful. These methods are divided into two main groups. In the first group, we only estimate labels of unlabeled samples [11,12]. These methods are known as "*transductive algorithms*". In the second group of methods, known as "*inductive algorithms*" [13], a decision function with very low error-rate for all samples (labeled and unlabeled) is sought. Another semi-supervised learning method, which has been studied widely, is graph-based semi-supervised learning. In these algorithms, the knowledge of the mutual data similarity is represented by graphs. In this regard, graph $G=(V,E)$ in which vertex set V includes all labeled and unlabeled samples and edge set E which contains similarity between data corresponding to vertex set of that edge set is considered. The graph is called similarity graph. Different types of graph-based methods by defining different similarity graphs can be introduced all of which have the same goal of modeling the relationship between sample point and its neighbors. Two conventional similarity graphs are ε-neighborhood graph and k nearest neighborhood graph. In ε-neighborhood graph, vertices of each pair of samples that have distance less than ε are connected to each other. In k nearest neighborhood graph, corresponding vertex of samples that belong to one of k nearest neighborhood are connected to each other. A semi-supervised learning method can be defined as a mincut problem [14].

Label propagation methods, which propagate labels of the training samples to test samples [15–20], are among semi- supervised learning methods. In consistency method [15] Gaussian kernel is employed to determine edge weights. In fact, in this algorithm, edge weights are determined using $e_{ij}=\exp(-||x_i-x_j||^2/2\sigma^2)$, $i\neq j$ and $e_{ii}=0$. In [16,17] k nearest neighborhood graph is used as similarity graph.

* Corresponding author.
E-mail address: arsaffari@yahoo.com (S.A. Saffari).

http://dx.doi.org/10.1016/j.neucom.2014.11.053
0925-2312/

After determining k nearest neighbors for each sample, that sample is expressed as a linear combination of its neighbors and the weight vector obtained by this method is considered as edge weights. Usually, in semi-supervised learning methods, center of attention is on graph structure and weights of edges are defined separately. ε-neighborhood and k nearest neighbors' graphs, which are usually used in these methods, have the following disadvantages: 1- These graphs are constructed using pair-wise Euclidean distance which is very sensitive to noise. 2- Considering different sample distributions, to determine proper neighbors for each sample, ε and k must be defined adaptively for each sample, but in ε-neighborhood and k nearest neighbors' graphs, a fixed parameter is considered for all samples so the accuracies of these graphs are very low.

Recently sparse representation (SR) has found various applications in machine vision and statistical pattern recognition [21–23]. In [24], l_1-graph, which is based on SR, is employed to produce a graph based algorithm. In l_1-graph, graph structure and weights of edges are found simultaneously using l_1-minimization. SR graph based algorithms have the following attributes compared to the other graph-based algorithms: First, in SR graph-based algorithms, graph structure and edge weights are found simultaneously by l_1-norm minimization. Second, since Euclidean distance is not employed, SR graph-based algorithms have lower degrees of sensitivity to noise. And third, in SR graph-based algorithms the number and degree of similarity of samples to each other are determined adaptively, hence there is no metric to determine the number and the degree of similarity between each sample and its neighbors. So efficiency of SR graph-based algorithms is much higher than other graph based methods which are explained so far in this paper.

We can group graph based semi-supervised learning algorithms which use SR theory in two categories. First category is the graph reduction semi-supervised learning methods [25,26]. Usually when the size of test samples grow, graph based semi-supervised learning methods have two major weaknesses: (1) Possible outliers and noisy samples have negative effect on the construction of the similarity graph. (2) The evaluation of predictors learned from the graph for new samples can be time-consuming if the predictors involve computations on all the samples in the original graph. To solve these problems graph reduction semi-supervised learning methods were introduced. In [25] a graph reduction method based on manifold-preserving sparse graphs has been proposed, where the number of vertices is reduced while the edge weights from the original graph are remained unchanged. In [26] a sparse semi-supervised learning framework using Fenchel-Legendre conjugates is proposed. The main focus of [26] is to reduce the number and to choose the appropriate unlabeled samples. The purpose of the second category of the graph based semi-supervised learning algorithms which use SR theory is graph construction [27,28]. Although there has been a numerous graph based semi-supervised learning methods, there are still much to do about neighbor selection and the degree of their similarities for each sample. In [3] a semi-supervised classification algorithm called the Sparse Regularized Least Square Classification (S-RLSC) algorithm. In [4] a semi-supervised classification approach through kernel-based sparse representation is proposed. This method computes the sparse representation of data in the feature space, and then the learner is subject to a cost function which aims to preserve the sparse representing coefficients. Our proposed method in this paper can also be categorized into graph construction methods but based on LLC and not the SR method.

Although numerous works have been done in machine vision based on the SR theory [29–32], little is said about the shortcomings of SR based graphs. Recently, a type of signal representation namely Locality-constrained Linear Coding (LLC) [33] is introduced in which local information is utilized, instead of sparsity constraint. Various studies make use of local information in order to enhance the learning efficiency, like feature reduction [34,35], density estimation [36], anomaly detection [37] and data classification [38,39]. k nearest neighbor classifier is the most familiar instance for local information usage. In SR-based algorithm, due to over completeness of the dictionary matrix, sometimes the samples selected as neighbors are not actually close to the related sample; the sparsity constraint in SR is what that has forced choosing those samples as neighboring samples. In fact, SR does not preserve the samples' local information during the coding process. The second problem of SR-based algorithm is the absence of an analytical trouble-free solution. Solving SR requires a time consuming optimization process. In LLC, local information constraint is employed instead of sparsity constraint and simple analytical solution exists. In this paper, we study the efficiency of using local information, like in LLC, in a form of label propagation algorithm. Experimental results show that our proposed algorithms have better classification rate compared to the other label propagation methods.

The rest of this paper is organized as follows: in Section 2, after a brief review on SR theory and LLC coding, some label propagation algorithms are investigated. Our proposed algorithms are introduced in the third section of this paper. The fourth section contains results of our experiments. Finally in Section 5 conclusion and future works are discussed.

2. Review on related works

In this section first a brief review on SR theory is presented. Then LLC, in which the local information is used instead of sparsity constraint, is briefly reviewed and finally a number of label propagation algorithms are introduced.

2.1. SR theory in machine vision

In recent years, sparse representation has caught researcher's attention in different fields. Based on this representation, several algorithms have been introduced. One of the first algorithms, introduced in statistical pattern recognition based on SR theory, is Sparse Representation based Classification (SRC) [29], which provides excellent accuracy in face data classification. SRC is based on a simple assumption that samples of a specific class are in the same sub-space. As a result, a test sample can be well represented by the training samples of its own class. In SRC, each test sample is expressed by a sparse linear combination of all training samples, and the non-zero elements of a coefficient vector are expected to point to a specific class.

Assume that dictionary matrix

$$\mathbf{D} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \mathbf{x}_{\ell+1}, \ldots, \mathbf{x}_n\right] \in \mathscr{R}^{d\times n}$$

contains all labeled and unlabeled samples; first ℓ samples contain labeled and the rest are unlabeled samples. Each sample can be expressed as linear combination of other samples. When the samples are numerous, the resulting weight vectors are sparse, i.e., many of their elements are zero. In this condition, optimal weight vector can be found with the aid of SR theory. SR theory can be defined as the following optimization problem:

$$\underset{\mathbf{w}_i}{\operatorname{argmin}} \|\mathbf{w}_i\|_0 \ \ s.t. \ \ \|\tilde{\mathbf{D}}\mathbf{w}_i - \mathbf{x}_i\|_2^2 \le \varepsilon \tag{1}$$

where $\mathbf{w}_i$ is weight vector for i^{th} sample and $\tilde{\mathbf{D}} \in \mathscr{R}^{d\times(n-1)}$ is a dictionary matrix in which i^{th} sample is omitted. Moreover $\|\cdot\|_0$ represents ℓ_0- norm of $\mathbf{w}_i$ that gives the number of non-zero elements of $\mathbf{w}_i$. In fact, Eq. (1) finds $\mathbf{w}_i$ vector such that in addition to satisfying $\|\tilde{\mathbf{D}}\mathbf{w}_i - \mathbf{x}_i\|_2^2 \le \varepsilon$, number of its non-zero elements are minimum and $\mathbf{w}_i$ vector is sparse. But since $\|\cdot\|_p$