Neurocomputing 142 (2014) 499–507

Contents lists available at ScienceDirect

Neurocomputing

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

Multiple perceptual neighborhoods-based feature construction for pattern classification

Guihua Wen [a,*], Lijun Jiang [a], Jun Wen [b]

[a] South China University of Technology, 510641 Guangzhou, China
[b] Hubei Institute for Nationalities, 445000 Hubei, Ensi, China

ARTICLE INFO

Article history:
Received 20 April 2012
Received in revised form
26 March 2014
Accepted 5 April 2014
Communicated by J. Zhang
Available online 2 June 2014

Keywords:
Classification
Feature construction
Feature selection
Neighborhoods

ABSTRACT

Feature construction is much critical to support classification tasks when a combination of the original features carries more discriminative information. However, the construction of features usually implies searching a very large space of possibilities and is often computationally demanding. Besides, some approaches require domain knowledge and the underlying principles of some approaches are hard to interpret. This paper presents a simple and efficient feature construction approach, which is independent of concrete classifiers and data domains. It begins with generating the features by calculating the distances between the sample and its neighborhoods in each class as features, as they have abilities to distinguish the sample. These features are then applied to combine with the original features of the sample to form the new feature vector for the sample. The novel work of this method lies in that a new general framework to create features for the given sample and a simple rule to combine the generated features with the original features are presented. This approach has been validated by applying it to a local classifier in experiments. The results suggest that the proposed method can be applied to nicely deal with the sparse and the noisy data.

1. Introduction

Many excellent classifiers cannot be comparable with human being when performing classification on the sparse and the noisy data [13]. When only few training samples compared to the intrinsic dimensionality of the feature space are available, many classifiers are not guaranteed to obtain the optimal results [10]. Many classifiers are also sensitive to outliers or noises as they equally treat the noisy samples and normal samples. The main reason is that it is much difficult to distinguish the noisy samples from the normal ones [45]. One way to solve these problems is to design the better classifiers while the data itself are not changed. The second method is to change the data usually through the feature selection that tries to make the data more separable. Feature selection aims to select a subset of relevant features for building robust learning models [31,20,5,12]. For instance, of all the millions of tests and observations that could be made on a human body, only a few need to determine whether a patient has a cold: checking for sore throat, fever, and runny nose. Thus, the various methods have been used to obtain high quality features. They can be categorized into three groups: feature selection, feature extraction, and feature construction [17,23]. Feature selection aims to reduce dimensionality by selecting a subset of original variables [19]. Feature extraction aims to reduce dimensionality by (linear or non-linear) projection of high dimensional vector onto lower dimensional vector. Linear discriminate analysis and principle component analysis are very common methods for feature extraction [24,48]. Feature construction is a process that discovers missing information about the relationships between features and augments the space of features by inferring or creating additional features [23].

Despite feature selection and feature extraction are useful for classification, they are not sufficient for learning concepts with primitive data representation when the underlying relations among features and the target concept are opaque and difficult to learn [37]. In such case, the feature construction is required, which has received increased attention in recent years [34,23]. The power of constructed features can be well illustrated by a simple example, the exclusive-or (XOR) problem, which has two features: x_1 and x_2. As posed, the two classes are not linearly separable and cannot be segregated in a two-layer neural network. However, a new non-linear feature, created by multiplying the values of x_1 and x_2 for each data point, transforms the classification problem to a space in which it is trivial [34]. Currently there many approaches can be applied to create new features, including principal components analysis, information theoretic tools [11], manifold learning methods [41,35,45,14], linear independent component analysis

* Corresponding author.
E-mail address: crghwen@scut.edu.cn (G. Wen).

http://dx.doi.org/10.1016/j.neucom.2014.04.007
0925-2312/

model [21], explanation method [25], frequent itemsets method [9], case-based reasoning [28], genetic algorithm [37,38], model-based approach [15], hierarchical methods [18,34], Bayesian method [26], graph mining approach [36], and genetic algorithm based approach [23]. Unfortunately, the construction of features usually implies searching a very large space of possibilities and is often computationally demanding [28]. For example, searching even a modestly expressive space of nonlinear functions for high information ones can be intractable [25]. Besides, some approaches require domain knowledge [28,25] and the underlying principles of some approaches are hard to interpret [15].

Different from the existing approaches, this paper proposes a new simple and efficient approach which applies the multiple neighborhoods to perform feature construction. This approach is then applied to improve a local classifier as validation.

2. Related work

Some feature construction approaches are based on the available domain knowledge. For example, the task-relevant discriminative features can be automatically constructed, guided by an explanation-based interaction of training examples and prior domain knowledge [25]. In the data mining, the frequent itemsets are also applied to be taken as new features [9]. However, they cannot be easily applied to the other domains without appropriate modification. Some other methods create features for dimensionality reduction, such as principal components analysis, information theoretic tools [11], and manifold learning methods [41,35,45,14]. For example, a linear independent component analysis model has been applied to interactive feature construction of binary data [21], employing a multivariate Bernoulli likelihood and independent Beta source densities. A dynamic causal model of neurophysiological data is also applied to construct features through trial-wise estimates of the model parameters [15]. Now there are some hierarchical feature construction methods, which repeatedly create higher-level features from lower-level features [34]. They are somewhat similar to deep learning approaches [18], based on the network structure and the recursive computing. One of the examples is Bayesian method that applies Bayesian networks to represent the conditional independence assumptions of a set of the original features, thus generating new ones [26]. In graph mining systems, a simple and efficient approach is to compute maximum common subgraphs from randomly selected pairs of data samples and to directly use them as features [36]. This kind of approach generally requires to construct a network from the given data set, whose computation time is much expensive. To decrease the computational demanding, a case-based reasoning is utilized to perform the feature construction. Learning tasks are stored together with a corresponding set of constructed features in a case base and can be retrieved to speed up feature construction for new tasks [28]. Another way is based on evolution computation [37,38]. For example, genetic programming is used to construct new features from those available ones in the data and a genetic algorithm is used to determine which such features are the most predictive [40].

Different from these existing works, our approach does not require domain knowledge and is computationally cheap, while it can significantly improve the predictive accuracy. Although the kernel technique is an effective way to solve insufficient feature problem by implicitly mapping the samples from the original space to a higher dimensional space, the optimal kernel function and the optimal parameters cannot be determined easily for the given data. Our approach not only obviously modifies the original space, but also can be further enhanced by kernel techniques.

The proposed approach in this paper has a very close connection with stacking. Stacked generalization or stacking is a type of ensemble method that uses the outputs of one classifier as inputs to another classifier [47]. There are two crucial issues in stacking: the type of generalizer that is suitable to derive the higher-level model, and the kind of attributes that should be used as its input. Typical attributes that are used at the meta-level are the class predictions of the base-level classifiers. It is also proved that the use of class probabilities is more crucial for the successful application of stacked generalization in classification tasks [46]. The other techniques such as fuzzy technique have been recently applied to stacking generalization. The method ensembles a set of fuzzy classifier whose fuzzy membership values are concatenated to form the feature vectors for a meta-layer classifier [27].

In this paper, the idea of stacking is used to generate new distinguished features, but with the different approaches. In this way the proposed approach can be extended with stacking framework such as taking the SVM as the meta-layer classifier.

3. New feature construction method

To nicely perform classification on the data with insufficient features, it is natural to learn the abilities from human being. Humans routinely classify objects according to both their individual attributes and membership in the higher order groups, where individual attributes may be influenced and regulated by their group [1]. This idea is illustrated in Fig. 1. When we observe the circle x, it looks bigger than its original size as it is surrounded by the smaller circles. In contrast, when the circle y is observed, it appears smaller than its original size as it is surrounded by the bigger circles. Consequently, when we observe x and y simultaneously, x is perceived to be bigger than y, although they are of the same size. This characteristic from the higher order group is very important for us to distinguish an object from its surrounding objects. We can formalize this characteristic to process the data more efficiently. One method is to define a transformation on the original space to build a new space whose dimensions are composed of distances between all sample points in the original space [8]. The newly created space is called the relative space that can be generated through relative transformation:

$$f^r : X \to Y \subset R^n$$

$y_i = f^r(x_i) = (d_{i1}, \ldots, d_{ij}, \ldots, d_{in}) \in Y, d_{ij} = \| x_i - x_j \|$, where $n = |X|$ is the number of elements in data set X, the sample point x_i in the original space is mapped to the sample point $y_i \in R^n$ in the relative space, and $\| \cdot \|$ is the distance norm.

Relative transformation can make the data distinguishable that cannot be identified in the original space. The relative transformation is also simple and efficient in dealing with noisy data or outliers, as shown in Fig. 2. In the original space the sample point x_4 can be regarded as a noisy point or an outlier since it is far away from the other three sample points. However, we have $d(x_3, x_1) = d(x_3, x_4)$ in the original space. This means that the sample point x_4 has the same possibility as the point x_1 to be taken as a nearest neighbor of the sample point x_3. This is not consistent with our intuition. In the relative space, $d(y_3, y_1) < d(y_3, y_4)$, the outlier or noisy sample point

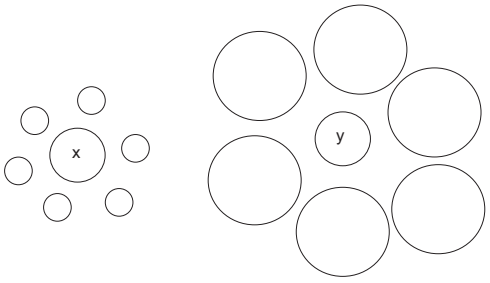

Fig. 1. Human visual perception is relative.