Information Sciences 370–371 (2016) 270–287

Contents lists available at ScienceDirect

Information Sciences

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

Combining neighborhood separable subspaces for classification via sparsity regularized optimization

Pengfei Zhu, Qinghua Hu*, Yahong Han, Changqing Zhang, Yong Du

School of Computer Science and Technology, Tianjin Univerity, Tianjin 150001, China

ARTICLE INFO

Article history:
Received 1 December 2015
Revised 27 July 2016
Accepted 2 August 2016
Available online 2 August 2016

Keywords:
Ensemble learning
Attribute reduction
Neighborhood rough sets
Joint representation
Group sparsity

ABSTRACT

The neighborhood rough set theory has been successfully applied to various classification tasks. The key concept of this theory is to find a sufficient and necessary neighborhood separable subspace for building a compact model. Given a classification learning task, there usually exist numerous neighborhood separable subspaces that maintain the discriminative ability of the original space with respect to a given granularity. These subspaces contain complementary information for classification. However, it is a challenging task to compute these subspaces efficiently. In this paper, we develop a fast neighborhood attribute reduction algorithm based on sample pair selection to find all reducts. Nevertheless, it cannot deal with large-scale data. Then we propose a randomized attribute reduction algorithm based on neighborhood dependency. The randomized algorithm can find a part of all reducts and is very efficient. A classification framework of joint subspace representation is proposed to fully exploit the complementary information in different subspaces. In addition, a weight matrix is learned to combine the representation residuals in the different subspaces via group sparsity regularization. The performances of the proposed attribute reduction algorithms are compared, and the influence of granularity on attribute reduction is discussed. Finally, the proposed technique is compared with other ensemble learning algorithms. Experimental results show that the proposed framework is superior to state-of-the-art classifiers.

1. Introduction

The rapid development of sensors and digital devices has contributed to the emergence of a wide variety of high-dimensional data, e.g., gene microarray data [36] and fault signals [55], which in turn has led to the so-called curse of dimensionality, including extremely high time complexity and storage requirements, as well as the failure of classification models [23]. Thus, to eliminate irrelevancy and redundancy in the feature space, it is necessary to establish low-dimensional structures in high-dimensional data. For this purpose, it is important to determine the characteristics of the raw feature space that should be preserved as well as those that should be removed. Moreover, although the diversity of feature subspaces has been widely exploited, e.g., random subspace method [47] and compressive sensing [52], it has not been investigated extensively from the viewpoint of classification tasks.

Neighborhood rough sets have been employed in vibration diagnosis [55], cancer recognition [20], and tumor classification[44]. The neighborhood rough set theory involves granulation of the feature space into a family of neighborhood

* Corresponding author. Fax.:+862227401839.
E-mail address: huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.ins.2016.08.004
0020-0255/

information granules and approximation of classification with these granules. If objects with similar feature values are grouped into the same class, the classification is said to be consistent; otherwise, it is inconsistent [20]. Consistency of the neighborhood information granules reflects the separability of the classes. If all the neighborhood information granules are consistent, the task is separable. Neighborhood dependency is essentially the percentage of consistent neighborhood granules [21,22]. It reflects the separation level of classification. Consequently, the neighborhood structure of the feature space can be introduced to evaluate the classification separability [22].

Redundant and irrelevant attributes in the feature space increase the computational complexity, and degrade classification performance. Various attribute reduction methods, including those based on the discernibility matrix [15,35,42,49], [54] and the attribute significance index [22], have been proposed for eliminating such attributes without reducing the approximation ability of the original feature space. In the former category of methods, the discernibility matrix is employed to construct a Boolean discernibility function, and all reducts can be obtained through the reduced disjunction of the discernibility function [29]. The disadvantage of such methods is the computation complexity of the discernibility matrix. It has been shown [10] that only the minimal elements in the discernibility matrix are useful for attribute reduction. Chen [9,10] proposed a fast attribute reduction method using rough sets and fuzzy rough sets to find sample pairs corresponding to minimal elements in the matrix. In the latter category of methods, different attribute evaluation indices, e.g., dependency, Shannon entropy, and mutual information, are used to develop heuristic algorithms with different search strategies [45]. For attribute reduction based on neighborhood rough sets, Hu [22] proposed a greedy algorithm based on neighborhood attribute significance. However, this algorithm can only find one reduct, and usually, it cannot find the optimal solution. According to the rough set theory, given a classification dataset, there are multiple attribute reducts that maintain the approximation ability of the raw feature space.

Obviously, it is useful to find the optimal reduct (i.e., the minimal reduct or some other reduct) for classification. Rough sets emphasize the approximation ability rather than the generalization ability of classification. In fact, different reducts provide diverse information, which describe the original task from different perspectives. Actually, different reducts can be considered as multi-view observations from different sources or sensors. Researches on cognitive psychology and neuroscience show that human brains can automatically store and combine multi-modal information [43]. Different areas of human brain cortex correspond to different modalities, e.g., text, audio, image,etc [39]. Additionally, compared to uni-modal case, multi-modal information can lead to shorter reaction time and higher recognition accuracy [16]. Hence, motivated by the multi-modal organization of the human brains, we can combine the information in different reducts and develop a robust classification model with high accuracy and efficiency.

In fact, subspace ensemble learning has attracted considerable attention, and it has contributed to significant improvements in classification performance. In general, an ensemble is built in three steps, i.e., generating subspaces, learning multiple base learners and combining their predictions [56]. First, subspaces, including sample subspaces and feature subspaces, are generated [28]. Sample subspace techniques perturb the training data with resampling methods, such as bootstrap sampling used in bagging [8]. Feature subspace methods introduce randomness in the feature space, e.g., random forests [6], random subspaces [19], and rotation forests [38]. In general, sample subspaces and feature subspaces are used together to increase the diversity of base learners [26,46,50,56]. As compared to random subspaces, neighborhood separable subspaces can guarantee diversity while maintaining the discrimination ability of the original feature space. Thus, they facilitate the implementation of effective subspace ensembles. Second, base learners are trained in these feature subspaces or sample subspaces. In the third step, majority voting is used in most methods for decision combination. In addition, the following ensemble pruning methods based on ordered aggregation have been proposed for selecting a subensemble: reduce-error pruning [31], orientation ordering [32], margin distance minimization [33], and boosting-based ordering [34]. All these methods attempt to treat base classifiers individually and ignore the class weights of the base classifiers.

In classification tasks, a basic problem is the development of a statistical, generative, or discriminative model to use the training samples to correctly label a test sample. The nearest neighbor classifier (NNC) [11] searches for the nearest sample, and the nearest feature subspace (NFS) [18] searches for the nearest linear subspace in which a class lies. Recently, representation-based classifiers (SRC[47]/CRC[53]) have been developed as a generalization of NNC and NFS; they treat the query sample as a linear combination of all the training samples and use the reconstruction residual of each class for classification. In essence, SRC/CRC search for the most similar samples, and the representation coefficients can be considered as the sample's importance. Representation-based classifiers exhibit excellent performance [17,30,48,60,62], and they can be applied to multi-task learning as well as tasks with nonlinear data distributions. Furthermore, joint representation has been proposed [51] to extend a single modal representation to multi-task presentation. Face blocks and different features are used to construct different dictionaries for recognition. If each neighborhood separable subspace is used as a dictionary, then a joint subspace representation-based classifier can be constructed.

In the present study, as shown in Fig. 1, first, efficient neighborhood attribute reduction algorithms were developed on the basis of the discernibility matrix, neighborhood dependency, and sample pair selection. Then, by considering each neighborhood separable subspace as a dictionary, a joint neighborhood separable subspace representation-based classifier (JNSSRC) was proposed for using the information in different subspaces. To consider the distinctiveness of different subspaces, a weight matrix was learned for combining the representation residuals of the different subspaces via group sparsity imposed on the weights. Finally, experiments were conducted to analyze the performances of the proposed attribute reduction methods, and JNSSRC was compared with some state-of-the-art classifiers.