Information Sciences 198 (2012) 48–61

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Feature selection using structural similarity

Sushmita Mitra [a], Partha Pratim Kundu [a,*], Witold Pedrycz [b,c]

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
[b] Electrical & Computer Engg., University of Alberta, Edmonton, Canada T6G 2G7
[c] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Received 2 May 2011
Received in revised form 6 January 2012
Accepted 24 February 2012
Available online 3 March 2012

Keywords:
Structural similarity
Multi-objective optimization
Feature selection
Proximity
Membership

ABSTRACT

A new method of feature selection is developed, based on structural similarity. The topological neighborhood information about pairs of objects (or patterns), to partition(s), is taken into consideration while computing a measure of structural similarity. This is termed proximity, and is defined in terms of membership values. Multi-objective evolutionary optimization is employed to arrive at a consensus solution in terms of the contradictory criteria pair involving fuzzy proximity and feature set cardinality. Results for real and synthetic datasets, of low, medium and high dimensionality, show that the method led to a correct selection of the reduced feature subset. Comparative study is also provided, and quantified in terms of accuracy of classification and clustering validity indices.

1. Introduction

Feature selection is essential in analyzing large data, particularly being a preprocessing step for reducing dimensionality, removing irrelevant features, reducing storage requirements and enhancing output comprehensibility. It is a process that selects a minimum subset of n' features from an original set of n features ($n' < n$), such that the feature space is optimally reduced according to a certain predetermined evaluation criterion. This often involves selecting subsets of features useful to build good predictors [16].

Search is a key issue in feature selection, involving search starting point, search direction, and search strategy. One also needs to measure the goodness of the generated feature subset. Feature selection can be supervised as well as unsupervised, depending on class information availability in data. The algorithms are typically categorized under filter and wrapper models [23], based on whether or not the learning methodology is used to select the feature subset. The wrapper methods assess feature subsets according to their usefulness to a given predictor. However selecting a good set of features is usually suboptimal for building a predictor, particularly in the presence of redundant variables. Since finding the best feature subset is found to be intractable or NP-hard [1], therefore heuristic and non-deterministic strategies are deemed to be practical.

Feature selection can be supervised or unsupervised. Supervised feature selection mostly depends on the performance of a chosen classifier. In the absence of class information, the unsupervised techniques use some intrinsic property of the data [26]. Here, no external information like class label of an instance is needed. Related literature on feature subset evaluation include Category Utility score [9], Fisher's feature dependency measure [37,11], entropy based unsupervised feature ranking [6], and generally proceed by selecting the subset(s) of features while preserving the inherent characteristic of data. In Ref. [38], the authors use an unsupervised method that assumes a linear model to choose a subset of features which can

* Corresponding author. Tel.: +91 3325753109.
E-mail addresses: sushmita@isical.ac.in (S. Mitra), pkundu2003@gmail.com (P.P. Kundu), wpedrycz@ualberta.ca (W. Pedrycz).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.02.042

approximate the original data. Zhao et al. [39] have proposed an embedded model which evaluates a feature subset based on their capability of preserving sample similarity.

The use of soft computing is an interesting proposition along this direction [2], in order to arrive at an acceptable solution at a lower cost. This is of particular interest towards the efficient mining and analysis of large data. We can utilize the uncertainty handling capacity of fuzzy sets [25] and the search potential of genetic algorithms for efficiently traversing large search spaces. When there are two or more conflicting characteristics to be optimized, often the single objective optimization function requires an appropriate formulation in terms of an additive combination of the different criteria involved. In such cases a *multi-objective* optimization becomes more appropriate. Multi-objective GAs (MOGAs) [7] may be used as a tool, while efficiently searching for optimal solutions.

An interesting way of looking at feature selection is to aim at preserving the structural similarity of data clusters, while mapping a high-dimensional feature space to a lower-dimensional one. In other words, a pair of objects (or patterns) belonging to the same partition in the original high-dimensional space is expected to be retained in the same partition in the reduced domain as well. By considering such similarity or proximity between all object pairs as a guideline [29], one can hope to eliminate some of the less important features. The aim is to retain those features which allow the similarity between the partitioning, in the original and reduced spaces, to be high. This can also help in improving the computational efficiency in the lower dimensional space, given that the mapping is nearly lossless as measured in terms of the similarity measure used.

In this article, we propose such a method of feature selection, based on structural similarity. The topological neighborhood information about pairs of objects (or patterns), to partition(s), is taken into consideration while computing a measure of structural similarity. This is termed proximity, and is defined in terms of membership values of the corresponding patterns. For a dataset with N input patterns we can define an $N \times N$ symmetric matrix, referred as proximity matrix P, whose (i,j) th entry represents the similarity (or dissimilarity) measure for the ith and jth patterns for $i, j = 1, \ldots, N$. Typically distance functions are used for the purpose. The proximity matrix is a pertinent construct that allows us to deal with structural information inherent in the data. In the fuzzy perspective the concept of similarity boils down to the membership values.

We focus on the use of proximity relationship, as a similarity measure, from the viewpoint of fuzzy sets. This is used as one of the objective functions, during multi-objective optimization, for evaluating the fitness of the feature subsets of varying cardinality. The use of fuzziness allows us to efficiently model uncertainties and ambiguities inherent in real life overlapping data. The proximity of a pair of patterns in the original feature space is compared with that in the reduced subspace of selected features. If they are similar, as measured in terms of their belonging to the same cluster (both before and after feature selection), then this implies that the eliminated feature(s) are not so relevant to the decision making process. The second criterion is the cardinality of the selected feature subset. This is sought to be minimized, and serves as a penalty to the objective function. A close observation reveals that these two criteria are of a conflicting nature. A smaller subset of features is likely to result in a reduced proximity, and hence reduced classification accuracy (as compared to the original feature space).

Multi-objective optimization is employed to arrive at a consensus solution in terms of this contradictory criteria pair, involving fuzzy proximity and feature set cardinality. Here MOGA is used as a tool for the multi-objective optimization, and any other technique could also have sufficed. The user does not need to specify the desired number of features, as it is embedded in the optimization process. The algorithm terminates when an optimal subset of features is obtained, according to the fitness criteria of the multi-objective genetic optimization. Experimental results indicate correct selection of the reduced feature subset. Validation of the selected set of features is reported in terms of classification accuracy using WEKA [17] implementation of several well-known classifiers, as well as internal and external clustering validity indices.

The rest of the paper is organized as follows. In Section 2 we present the proximity-based methodology for feature selection and outline the background on multi-objective optimization. The experimental results and comparative study are described in Section 3, on various real and synthetic datasets. Finally, Section 4 concludes the article.

2. Proximity-based feature selection

Let us consider Fig. 1 to explain the concept of structural similarity between clusters in the context of feature selection. Using this crude example, we have discussed that the idea of preserving cluster structure of original feature space in a feature subset, would actually lead to feature selection. Removing irrelevant feature(s) does not affect much the internal characteristics of data. Three patterns $X1$, $X2$ and $X3$ are seen to be partitioned into the same cluster in the three-dimensional feature space of part (a). The three features are aligned with three reference axes *i.e.* x-axis, y-axis and z-axis of this dataset. If the least important feature *i.e.* the feature aligned with y-axis is eliminated, the cluster structure is expected to remain unaltered; implying that the single cluster would still contain the same distribution of pattern points as depicted in part (b) of the figure. Here the three- to two-dimensional mapping is said to be almost lossless, such that the clustering structures in the two subspaces are very similar. The clustering structure is said to be preserved in the transformation between the two subspaces. On the other hand, if an important feature *e.g.* the feature aligned with z-axis is eliminated then the mapping is bound to disrupt the cluster structure since important information gets lost in the process. From part (c) of the figure we observe that the similarity between the partitioning, in the two subspaces, is now no longer high. In other words, the distance between the partitioning is higher; with the pattern points getting redistributed into two different clusters *i.e.* cluster structure of original space is not preserved here.