Expert Systems With Applications 108 (2018) 61–73

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Generation of compound features based on feature interaction for classification

Sreevani[a,*], C.A. Murthy[a], Bhabatosh Chanda[b]

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 108, India
[b] Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata 108, India

ARTICLE INFO

Article history:
Received 19 November 2017
Revised 26 April 2018
Accepted 26 April 2018
Available online 30 April 2018

Keywords:
Feature extraction
Feature selection
Compound features
Semi-features
Information theory
Feature interaction
Mutual information

ABSTRACT

Dimensionality Reduction (DR) is an important preprocessing step in supervised learning. It involves reducing/mapping of high-dimensional data into a low-dimensional space by preserving important information. The classical DR paradigm can be divided into Feature Selection (FS) and Feature Extraction (FE) approaches. So far, these approaches have been studied extensively but independently and the reduced set contains either original or transformed features. And, it is well known that FS and FE approaches based on information theoretic measures are considered to be the most effective approaches as these measures are able to compare the nonlinear relationships between random variables. Herein, we present a novel scheme to generate reduced compound (both original and transformed) feature set based on such measures in supervised learning. This method considers information theoretic measure Mutual Information (MI) and MI based interactions between features and it is able to produce maximum informative and less redundant compound features. The performance of the proposed algorithm is compared with state-of-the-art DR methods using multiple classifiers on UCI machine learning repository and face and object recognition and bio-microarray data sets.

1. Introduction

With the recent explosion of the size of the available data sets, high dimensional data is becoming more prominent in both supervised and unsupervised learning paradigms. The proliferation of high dimensional data causes serious problems to many Machine Learning algorithms with respect to scalability and learning performance (Van Der Maaten, Postma, & Van den Herik, 2009). It is very important to reduce the dimensionality of the data to decrease the training time and enhance the performance of the learning algorithms (Guyon & Elisseeff, 2003; Liu & Yu, 2005). Dimensionality Reduction (DR) methods are widely used in many application areas relevant to expert and intelligent systems, such as machine learning, image processing, anomaly detection, bioinformatics and text mining. These DR methods can be divided into two major categories: feature selection and feature extraction. Feature selection methods reduce the dimensionality by selecting a subset of important features. Feature extraction methods transform existing features into a new feature space of lower dimensionality. Unlike feature extraction, feature selection does not alter the original data and feature extraction may be preferred when only discrimination is needed (Jain, Duin, & Mao, 2000).

Feature selection has been reviewed in a number of recent review articles (Guyon & Elisseeff, 2003; Vergara & Estévez, 2014). Usually, feature selection methods are divided into two groups based on the evaluation process, classifier dependent ('wrapper' and 'embedded') or classifier independent ('filter') methods. Wrapper methods select a feature subset using the prediction accuracy of a classifier and perform well as the quality of the selected subset is optimized for the classification algorithm. But, these methods may suffer from over-fitting to the learning algorithm and very expensive in terms of computational complexity (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013). Embedded method is combined with the learning stage and less expensive in terms of computational complexity and less prone to over-fitting. Filter methods rank features according to their relevance to the class label. The relevance score is calculated using distance, information, correlation and consistency measures (Liu, Lin, Lin, Wu, & Zhang, 2017). The main advantages of the filter methods are their computational efficiency and independence from the classifier. The popular filter feature selection methods usually employ feature selection criteria based on the first and second-order statistics computed from the empirical distribution (Liu, Motoda, & Dash, 1998; Narendra & Fukunaga, 1977). The major drawback of these feature

* Corresponding author.
E-mail addresses: sreevani_r@isical.ac.in (Sreevani), murthy@isical.ac.in (C.A. Murthy), chanda@isical.ac.in (B. Chanda).

https://doi.org/10.1016/j.eswa.2018.04.033
0957-4174/

selection criteria are sensitive to data noise and data transformation (Dash & Liu, 1997).

Feature extraction creates new variables by combining the original features to reduce the dimensionality of the data. Feature extraction algorithms can be categorised in to linear and non-linear methods. Non-linear techniques are mostly kernel based, for example, kernel PCA (KPCA), kernel Discriminant Analysis (KDA), kernel MMC (KMMC) (Lee & Verleysen, 2007). Other non-linear techniques which preserve local properties of the data are, Locally Linear Embedding (LLE), Laplacian Eigenmaps (Tenenbaum, De Silva, & Langford, 2000), Hessian LLE (Donoho & Grimes, 2003), local tangent space analysis (Zhang & Zha, 2004). Linear feature extraction methods assume that the data lies on a lower-dimensional linear subspace. Popular linear dimensional reduction algorithm is Principal Component Analysis (PCA). However, PCA does not use the class label information which results in loss of some useful discriminating information for classification. A popular linear dimensionality reduction algorithm for supervised FE is Linear Dimensionality Analysis (LDA). LDA has been proven to be more effective than PCA as it uses discriminatory information. Some of the other linear FE methods are, Maximum Margin Criterion (MMC), Angle Linear Discriminant Embedding (ALDE), Marginal Fisher Analysis (MFA) (Fukunaga, 2013), Sparse Linear Discriminant Analysis (SLDA) Clemmensen, Hastie, Witten, and Ersbøll (2011). These well-known linear FE methods rely on first- and second-order statistics (remote means and low variances), which provide some clues about the linear separability between classes. But, FE methods, in spite of being linear, must retain the discriminative information even for the non-linear class boundaries.

From the above, one can conclude that a DR (FS or FE) method must consider high-order moments to model the discrimination with in the data. For this purpose, Shannon's information theory provides a powerful tool Mutual Information (MI), which can capture non-linear dependency between variables (Battiti, 1994). It can be considered as higher order statistics and is more robust to noise and data transformation (Principe, Xu, & Fisher, 2000). Recently, it has been widely used both in FS and FE (Battiti, 1994; Bennasar, Hicks, & Setchi, 2015; Torkkola & Campbell, 2000).

All the above mentioned methods provide reduced set either with original or transformed features. Also, the transformed features are the combinations of all of the original features. Recently, a new strategy for DR with the aim of providing reduced set with both original and combinations of features (compound features), without losing orthogonality among the selected combinations of features (semi-features) and original features, is proposed (Sreevani & Murthy, 2017). This method can keep balance between discrimination and interpretation of the resulting features as it provides both original and combinations of features. The algorithm, named as MPeMR, is based on recursively combining (PCA/ LDA) original/derived feature pairs and removing redundant features by selecting the representative one. This algorithm is an iterative method and in each iteration, semi-features are extracted considering only pairs of features i.e., each iteration generates semi-features which are only a 2-linear combination of features derived from the previous iteration. Moreover, MPeMR generates semi-features by considering only second-order statistics to model the discrimination present in the data. In this work, compound features are generated by considering information theoretic measure, higher order statistic MI and MI based feature interactions. Instead of considering pairs of features, here features are grouped so that the extraction of semi-features is performed on that group and semi-features are linear combinations of all the features present in the group. The proposed method extracts informative semi-features from the features which are 3-way interacting with the class label and eliminates both irrelevancy and redundancy among the generated compound features. Finally the algorithm provides maximum informative and less redundant compound features.

The rest of the paper is organized as follows. In Section 2, a summary of information theory concepts is provided and Section 3 explores the related work based on information theory. In Section 4, we describe the proposed method which uses MI concepts for compound feature generation in classification. Section 5 includes a set of experiments to show the feasibility of our method. Finally conclusion has been made in Section 6.

2. Background theory

As the interest of the article lies in providing compound feature subsets based on information theoretic measures, firstly we present the definitions of semi-feature and compound feature set (Sreevani & Murthy, 2017), then provide a brief introduction to information theoretic concepts.

Let $O = \{f_1, f_2, \ldots, f_D\}$ be the given set of features.

Semi-feature A feature s is called a *semi-feature* if it is a combination of only a *proper subset* of original features i.e., $s = a_{i1}f_{i1} + a_{i2}f_{i2} + \cdots + a_{ik}f_{ik},\ f_{i1}, \ldots, f_{ik} \in O,\ a_{ij} \in \mathbb{R}$ and $k < D$.

Compound Feature Set (CFS) A set R which contains both original (f) and semi-features (s) while maintaining orthogonality among them i.e.,

$$R = \Big\{ u \mid u \in \{f_{i1}, \ldots, f_{ip}, s_{j1}, \ldots, s_{jq}\}, \text{any two } u'\text{s} \\ \text{are orthogonal\& } p + q \leq D \Big\}.$$

Compound Feature Generation (CFG) Process of generating compound features without losing orthogonality among the selected original and combinations of features.

Assume a random variable X representing continuous-valued random feature, and a discrete-valued random variable $\mathcal{C}$ representing the class labels. The *entropy* of a random variable is a measure of its uncertainty (Cover & Thomas, 2012). The *entropy* of the class label $\mathcal{C}$ is denoted by $H(\mathcal{C})$ and is defined as:

$$H(\mathcal{C}) = -\sum_{c \in \mathcal{C}} p(c) log\ p(c),$$

where $p(c) =$ probability$\{\mathcal{C} = c\}$ represents the probability of the discrete random variable $\mathcal{C}$.

The uncertainty about $\mathcal{C}$ given X is measured by the *conditional entropy* as,

$$H(\mathcal{C}|X) = -\int_x p(x)\Big(\sum_{c \in \mathcal{C}} p(c|x) log\ p(c|x)\Big)dx,$$

where $p(c|x)$ is the conditional probability for the variable $\mathcal{C}$ given X. The *conditional entropy* is the amount of uncertainty left in $\mathcal{C}$ when a variable X is introduced, so it is less than or equal to the entropy of both variables. The *conditional entropy* is equal to the entropy if, and only if, the two variables are independent.

Mutual Information (MI) is the amount of information that both variables share, and is defined as:

$$I(X; \mathcal{C}) = H(\mathcal{C}) - H(\mathcal{C}|X),$$

and after applying the identities $p(c, x) = p(c|x)p(x)$ and $p(c) = \int p(c, x)dx$, *MI* can be expressed as,

$$I(X; \mathcal{C}) = \sum_{c \in \mathcal{C}} \int_x p(x, c)\ log\ \frac{p(x, c)}{p(c)p(x)}\ dx.$$

This is the difference of two entropies - the uncertainty before X is known ($H(\mathcal{C})$), and the uncertainty after X is known ($H(\mathcal{C}|X)$). This can also be interpreted as the amount of uncertainty in $\mathcal{C}$ which is removed by knowing X, which is the amount of information that