Applied Soft Computing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Feature unionization: A novel approach for dimension reduction

Q1 Abbas Jalilvand[a,b,*], Naomie Salim[a]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Department of Computer Engineering, Hashtgerd Branch, Islamic Azad University, Alborz, Iran

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form 27 June 2016
Accepted 16 August 2016
Available online xxx

Keywords:
Dimension reduction
Feature selection
Feature unionization
Sentiment classification

ABSTRACT

Dimension reduction is an effective way to improve the classification performance in machine learning. Reducing the irrelevant features decreases the training time and may increase the classification accuracy. Although feature selection as a dimension reduction method can select a reduced feature subset, the size of the subset can be more reduced and its discriminative power can be more improved. In this paper, a novel approach, called feature unionization, is proposed for dimension reduction in classification. Using union operator, this approach combines several features to construct a more informative single feature. To verify the effectiveness of the feature unionization, several experiments were carried out on fourteen publicly available datasets in sentiment classification domain using three typical classifiers. The experimental results showed that the proposed approach worked efficiently and outperformed the feature selection approach.

1. Introduction

Q3

With rapid advance of internet technologies, the number of electronic documents has hugely increased worldwide. The fundamental goal of the text document classification is to assign appropriate classes to electronic text documents. Text classification has many applications in natural language processing tasks such as sentiment classification [1–3], spam e-mail filtering [4], topic detection [5], news filtering [6], web page classification [7], and document organization [8]. Since text data is unstructured, Bag of Words (BOW) model (i.e., term-based Vector Space Model (VSM)) has been most popular for document-feature representation [9]. In the BOW model, each word (term) or phrase is considered as a unique feature. Thus, text collection would result in tens or hundreds of thousands of features. In theory, having more features should improve the efficiency of classifier; however, it is not always true practically. More features may confuse the learning algorithm because most of the features are irrelevant or redundant, which may lead a classifier to over-fitting. Moreover, a large number of features impose a high computational cost on the learning step. Accordingly, dimension reduction is needed to remove unnecessary features to improve classifier's generalization ability and computational efficiency. Dimension reduction for high dimensional datasets is a significant challenge in statistical machine learning. This approach falls into feature selection and feature extraction. The feature extraction (FE) methods create a new reduced feature space from transformation of original high-dimensional feature space. Two commonly-used techniques of feature selection are principal component analysis (PCA) and linear discriminant analysis (LDA), which have been widely used by researchers. On the other hand, the feature selection (FS) methods aim to select a subset from the original set of features according to discrimination capability [10]. In these methods, a subset of features is constructed through identifying and removing as many irrelevant and redundant features. The relevant features are those that are highly correlated with the target class and distributed most differently among all classes, whereas redundant features are those that do not add anything new to the target class [11]. Accordingly, the FS methods attempt to increase relevancy and reduce redundancy as much as possible to construct an optimal feature subset. Based on feature selection and transformation, a new approach was proposed to transform the selected feature subset using combination of features in order to construct a more reduced feature subset. Taking into account the synonym words that can be considered as a feature, the basic idea of the proposed approach was to reduce dimensionality based on finding and combining features that can construct a more informative single feature based on a feature relevancy criterion. There are implicit and explicit relations between occurring words (features) in the same class. For example, synonyms, or the words of the same group, may tend to occur in the same class. In sentiment classification domain, two words, for example, 'good' and 'great'

Q2 [*] Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
E-mail address: abbas.jalilvand.ai@gmail.com (A. Jalilvand).

http://dx.doi.org/10.1016/j.asoc.2016.08.031
1568-4946/© 2016 Published by Elsevier B.V.

usually indicate positive class that can be unionized to make an individual feature. Practically, capturing these relations is not easy to do because most of them are latent. The proposed feature unionization approach can capture the relation of features according to their relevancy to the target class in a way to construct an informative feature. Since combination of features is carried out by union operator, the redundancy can also be removed due to inherent characteristic of unionization. In this paper, our focus is on the problem of supervised feature unionization and we propose a solution that is suitable for binary datasets.

1.1. Motivation

Reduction of feature dimensionality is very important for classification task to reduce the computational complexity and avoid overfitting problem, which improves the generalization ability of classifier. For having better generalization performance of the classification, the number of features should be reduced as much as it is required for the number of training samples. Although feature selection, as a dimension reduction method, can select a reduced feature subset, the size of this subset needs to be more reduced; meanwhile, the discriminative power of features should not be weakened. These reasons motivated us to propose an effective approach to construct a more compact and discriminative feature subset by features combination idea.

1.2. Contribution

A feature unionization algorithm was proposed in this research to combine features while keeping their discriminative information. The effectiveness of this algorithm, in terms of solution quality and computational efficiency, was experimentally demonstrated on a wide variety of datasets. Feature dimension space was significantly reduced because multiple features were unionized into a single feature. The performance of classification was improved due to transformation of feature space to a more discriminative subset. Considering the increased demand for analyzing data with large feature dimensionality in some emerging domains such as text classification, we expect widespread use of this algorithm in these applications. Moreover, this paper can provide a new direction of dimension reduction research in addition to providing some new algorithms.

The remainder of the paper is organized into four sections. Section 2 describes the theory related to feature selection and presents a literature survey of the existing methods. In Section 3, a new dimensionality reduction (DR) approach termed as feature unionization (FU) is proposed. Section 4 discusses experimental results obtained from fourteen sentiment datasets. The conclusions are presented in Section 5.

2. Related work

In classic supervised learning, a classifier is learnt based on training dataset to predict the labels of new instances (examples). Training dataset contains M instances described by N features (attributes) and L class labels. The goal of feature selection is to select relevant features and remove the redundant ones. Feature X_i is strongly relevant to sample S if there exist examples A and B in S, which differ only in their assignment to X_i and have different labels. Feature X_i is weakly relevant to sample S (or to target C and distribution D) if it is possible to remove a subset of the features so that X_i becomes strongly relevant. Features that are strongly relevant are generally important to keep no matter what, at least in the sense that removing a strongly relevant feature adds ambiguity to the sample. Depending on removed features, weakly relevant features may become important. Notions of feature redundancy are normally in terms of feature correlation. It is widely accepted that two features are redundant to each other if their values are completely correlated [12].

In the context of classification, feature selection techniques can be divided into three groups: filter, wrapper, and embedded techniques [13–15]. Filter techniques assess a feature (subset of features) based on various measures of the general characteristics of the training data (e.g., distance, information, dependency, and consistency) [16,17]. They work independently of any specific classification algorithm and can act as a preprocessing step to a learning algorithm. In contrast to filter techniques, wrapper and embedded methods are dependent on classification algorithms. Wrapper techniques search the space of feature subsets to find an optimal subset based on wrapping around a particular classifier using the training/validation accuracy measure [18]. The embedded techniques evaluate the subsets of feature in the training process of learning model to select the best one [19]. In practice, although the wrapper and embedded techniques are better than the filter techniques in terms of classification accuracy, they are much more time-consuming and even intractable in case of high dimensional dataset. Moreover, the filter techniques usually provide more generic knowledge of the data due to their independence from classifiers. Thus, the filter techniques are more popular, especially in high dimensional datasets. Based on considering feature dependencies, the filter techniques, in turn, can be divided into two groups: univariate and multivariate techniques [17]. Univariate techniques evaluate features separately to remove irrelevant ones using a certain criterion. Although they can determine effectively relevant features, they fail to handle redundancy due to ignoring feature dependencies. Information gain (IG) and Chi-square (CHI2) are the most popular univariate filter techniques that evolved from either the information theory or the linear algebra literature [12,20]. CHI2 and IG can be computed using Eqs. (1) and (2)

$$\mathrm{IG} = P(t_k, c_i)\log\frac{P(t_k, c_i)}{P(t_k).P(c_i)} + P(\bar{t}_k, c_i)\log\frac{P(\bar{t}_k, c_i)}{P(\bar{t}_k).P(c_i)} \quad (1)$$

$$\mathrm{CHI2} = \frac{N.[P(t_k, c_i).P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i).P(\bar{t}_k, c_i)]^2}{P(t_k).P(\bar{t}_k).P(c_i).P(\bar{c}_i)} \quad (2)$$

where t_k denotes a term; c_i stands for a category; $P(t_k, c_i)$ signifies the probability of documents from category c_i where term t_k occurs at least once; $P(\bar{t}_k, c_i)$ represents the probability of documents not from category c_i where term t_k occurs at least once; $P(t_k, \bar{c}_i)$ denotes the probability of documents from category c_i where term t_k does not occur; $P(\bar{t}_k, \bar{c}_i)$ represents the probability of documents not from category c_i where term t_k does not occur.

In contrast, multivariate filter techniques, with considering feature dependencies, can select a better feature subset by elimination of irrelevant and redundant features. CFS [21], FCBF [22], and CMIM [23] are examples that take into consideration the redundant features. The hypothesis in CFS [21] is that a good feature subset is one that contains features highly correlated with the target class, yet uncorrelated with each other. FCBF [22] is a fast filter technique that can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. CMIM [23] iteratively picks features that maximize their mutual information with the target class to predict, conditionally to the response of any feature already picked. Based on the idea of feature selection, the most optimal feature subset can be found if all of the irrelevant and redundant features are identified and removed. In the best situation, N features can be reduced to M features (assuming that the number of features in the optimal subset is equal to M). Although, this optimal feature subset can be more reduced by feature combination idea. Despite removing all irrelevant and redundant features, they can be combined to make more powerful features. Quite

Please cite this article in press as: A. Jalilvand, N. Salim, Feature unionization: A novel approach for dimension reduction, Appl. Soft Comput. J. (2016), http://dx.doi.org/10.1016/j.asoc.2016.08.031