Knowledge-Based Systems 121 (2017) 83–98

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A new validity index of feature subset for evaluating the dimensionality reduction algorithms

Chuan Liu[a], Wenyong Wang[a,*], Martin Konan[a], Siyang Wang[c], Lisheng Huang[a], Yong Tang[a], Xiang Zhang[b]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[c] Mathematics Department, University of California San Diego, La Jolla 92093, USA

ARTICLE INFO

Article history:
Received 12 August 2016
Revised 13 December 2016
Accepted 12 January 2017
Available online 12 January 2017

Keywords:
Machine learning
Validity index
Cross validation
Dimensionality reduction algorithm

ABSTRACT

A critical aspect of dimensionality reduction is to assess the quality of selected (or produced) feature subsets properly. Feature subset assessment in machine learning refers to split a given feature subset into a training set, which is used to estimate the parameters of a classification model, and a test set used to estimate the predictive performance of the model. Then, averaging the results of multiple splitting (i.e., Cross-Validation, CV) is commonly used to decrease the variance of the estimator. *But in practice, CV scheme is very computationally expensive.* In this paper, we propose a new statistics index method called LW-index for evaluation of feature subset and dimensionality reduction algorithms in general. The proposed method is a type of “classical statistics” approach that uses the feature subset to compute an empirical estimate of the quality of feature subset. A large number of performance comparisons with the machine learning approach conducted on fourteen benchmark collections show that the proposed LW-index is highly correlated with the external indices (i.e., $MacroF_1$, $MicroF_1$) of SVM and Centroid-Based Classifier (CBC) trained by five-fold CV scheme. Furthermore, the experimental results indicate that LW-index has the same performance as the traditional CV scheme for evaluating the dimensionality reduction algorithms and it is more efficient than the traditional methodology. *Therefore, one contribution of this paper is to present an alternative methodology, based on an internal index typically used in the unsupervised learning context, that is computationally cheaper than the traditional CV methodology. Another contribution is to propose a new internal index that behaves better than other similar indices widely used in clustering and shows high correlation with the results obtained by the traditional methodology.*

1. Introduction

Data objects are typically described by a large number of features which negatively affect the performance of the underlying learning algorithms. To improve the prediction and learning efficiency to understand the underlying process better, reducing the dimensionality of data is a critical preprocessing step in machine learning. Therefore, Dimensionality Reduction (DR) algorithm [1,2] selects (or produces) the most important features from the whole feature set through a data mining process. Doing so, it removes irrelevant and redundant features in the original data, which brings many advantages [3] such as facilitating data visualization, reducing storage requirements, avoiding over-fitting, and reducing training time. Thus, an effective DR algorithm will dramatically improve the learning efficiency and efficacy for a given model. It is especially suitable for high dimensional data applications such as text classification [4,5], WEB data mining [6,7] and DNA biological information classification [8,9].

Throughout the past years, DR algorithms have been proposed in many literatures [9–16]. However, few literatures were able to successfully address the evaluation challenge of feature subsets returned by DR algorithms, which is no doubt an important and challenging issue. To our best knowledge, the state-of-the-art methods to evaluate feature subset are Hold-Out (HO) and Cross-Validation (CV) [17,18] schemes using a particular classifier as the measure of importance (or significance) for a candidate feature subset. However, HO method is unreliable since the selection of

[*] Corresponding author.
E-mail addresses: liuchuan@uestc.edu.cn (C. Liu), wangwy@uestc.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.knosys.2017.01.017
0950-7051/

testing set has a direct impact on the whole performance, while CV method is accompanied with huge resource consumption.

Under the CV scheme, each feature subset obtained from DR algorithm in supervised classification problem can be evaluated using external indices, such as *Entropy* [19], *Purity* [20], *F-measure* [21], since we train and test the classification model. Hence, the obtained data partition by the classification model can be compared with the original data partition provided by the class label. Alternatively, we can also evaluate the feature subset with an internal index (*see Section 4.1*), if we use the class label to obtain the data partition (if the data partition given by the class label represents well-separated groups, it will be easier to be modeled by a classification method). Thus, there is no need to train and test the classifier, which is, in general, time-consuming. If the internal index computation is more efficient than the classifier training and testing along with obtained partition evaluation process and, additionally, if the proposed internal index is highly correlated with external indices, then the proposed methodology can be a good alternative to evaluate the feature subsets and further DR algorithms in supervised classification problem.

In this paper, to solve the issue mentioned above, we propose a new internal validity index which directly measures the quality of candidate feature subset without incorporating a classification (or clustering) model in supervised classification problem. In contrast to CV scheme, the proposed index based on a novel conception of *freedom degree* is less computationally expensive due to its linear complexity. Therefore, we replace the time-consuming CV scheme by the proposed index to evaluate DR algorithms. The experimental results indicate that LWI is an efficient and effective approach that outperforms the traditional CV scheme.

The rest of this paper is structured as follows: Section 2 presents DR algorithms and the evaluation methods of DR algorithms. Section 3 proposes a new index method, and the experimental results and analyses are illustrated in Section 4. The concluding remarks are presented in Section 5, where the proposed approach is summarized and its key features are identified. Finally, the paper ends with the Acknowledgements.

2. Related work

Two approaches have been proposed for the purpose of DR [22]: feature selection and feature extraction. The feature selection methods search a relevant subset from a host of existing features [23,24], which can be transformed into a combinatorial optimization problem. Most of feature selection algorithms depend on heuristic methods to obtain a subset of relevant features in a flexible time. In contrast, the feature extraction (also known as feature transformation) methods learn a new set of features which are different from the existing features [25]. Feature extraction algorithms usually produce a set of continuous vectors that represent data objects in the extracted feature space. However, the problem is that feature extraction algorithms produce features that are difficult to interpret, therefore they are not competitive and suitable especially in applications where understanding the meaning of features is the major need for data analysis.

Feature selection algorithms can be broadly grouped into approaches [10,26] that are classifier-independent (filter methods [13,14]), and classifier-dependent (wrapper and embedded methods [9,12]). Filter methods evaluate the importance of features independently of any particular classifier, thereby leading to a faster-learning pipeline of features that are generic and less likely to overfitting than wrapper and embedded methods. Usually, they choose the sorted features according to a heuristic scoring criterion to act as a proxy measure of the classification accuracy. Many hand-designed heuristic filter criteria, such as Information Gain (IG) [27], Mutual Information (MI) [27], Chi-Square (CS) [28], Cross Entropy (CE) [29], Latent Semantic Analysis (LSA), and GSS Coefficient [30], have been proposed in text categorization research. Recently, Lu et al.[31] proposed a text feature selection method based on category-distribution divergence. In literature [32], a novel integration approach called modified union was proposed, which applies union on selected top-ranked features and intersection on the remaining sub-lists features. Also, a survey on feature selection methods was done in literature [1] that presents some feature selection techniques to provide a comparative analysis on standard datasets. In summary, the methods used to evaluate feature subsets (or feature selection algorithms) in these literatures are still HO and CV schemes. However, the authors in literature [33] show that CV scheme would contribute to overestimating the model performance. Moreover, they demonstrate that intensive search techniques, such as Sequential Forward Floating Selection (SFFS) algorithm, do not necessarily outperform a simpler and faster method such as Sequential Forward Selection (SFS). Therefore, they recommend the use of bias for the simplest search strategies that are less prone to overfitting. However, they cannot solve the overfitting problem of CV scheme completely.

Wrapper methods search the space of feature subsets, using the training and testing of a particular classifier as the measure of utility for a candidate subset. A growing number of machine learning techniques have been applied to the wrapper methods as induction algorithms, which include Naive Bayes (NB) [34], k-Nearest Neighbor (kNN) [34,35], Neural Network (NN) [36], Decision Tree (DT) [34,37], and Support Vector Machines (SVM) [10,12,34,38]. Though these methods may guarantee good results, they also suffer from the disadvantage of being computationally expensive and more unfeasible (computationally) as the number of features increases. Additionally, these methods may produce feature subsets that are overly specific to the used classifier. To overcome the drawback of wrapper that needs more computational cost and the weakness of filter that is insufficiently reliable for classification, the hybrid methods such as SAGA [15] and GAMIFS [16] have been recently proposed. However, the performance evaluation of the methods mentioned above remains a thorny problem.

Wrapper methods measure the quality of feature subsets without incorporating knowledge about the specific structure of the classifier, therefore they can be combined with any learning methods. On the contrary, embedded methods exploit the structure of specific classes of learning classifiers to guide the feature selection process, thus the defining component of an embedded method is a criterion derived from the fundamental knowledge of a specific class of regression or classification function [26]. For instance, Weston et al.[39] presented a method that selects features to minimize a generalization bound held by SVM. Perkins et al.[40] suggested to minimize a function that defines the family of regression or classification, and solve it in a greedy forward way.

Each of these feature selection methods has its advantages and disadvantages [26,41]. In general, in terms of accuracy, wrapper methods have high learning capacity. Thus, they usually obtain higher accuracy than embedded methods, which in turn are better than filter methods. However, in terms of speed, filters are the fastest among all the methods as they need not incorporate learning, while wrappers are the slowest since they typically need to evaluate the CV procedure at each iteration. Also, embedded methods are faster than wrappers since the function that measures the quality of a scaling factor can be evaluated faster than a CV estimation procedure. Hence, embedded methods take the advantage of wrapper methods while avoiding their computational complexity to improve the performance. However, using embedded methods, we still have to compute the criterion derived from classification function several times (at each iteration, we have to compute the criterion as many times as the number of features).