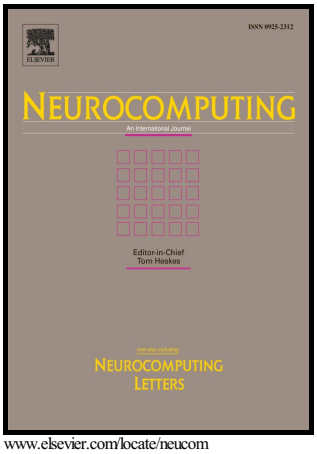

www.elsevier.com/locate/neucom

An Efficient Feature Selection Algorithm for Hybrid Data

Feng Wang, Jiye Liang

PII: S0925-2312(16)00155-7
DOI: http://dx.doi.org/10.1016/j.neucom.2016.01.056
Reference: NEUCOM16712

To appear in: *Neurocomputing*

Received date: 10 July 2015
Revised date: 5 January 2016
Accepted date: 15 January 2016

Cite this article as: Feng Wang and Jiye Liang, An Efficient Feature Selection Algorithm for Hybrid Data, *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2016.01.056

An Efficient Feature Selection Algorithm for Hybrid Data

Feng Wang, Jiye Liang*

[a]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China*

Abstract

Feature selection for large-scale data sets has been conceived as a very important data preprocessing step in the area of machine learning. Data sets in real databases usually take on hybrid forms, i.e., the coexistence of categorical and numerical data. In this paper, based on the idea of decomposition and fusion, an efficient feature selection approach for large-scale hybrid data sets is studied. According to this approach, one can get an effective feature subset in a much shorter time. By employing two common classifiers as the evaluation function, experiments have been carried out on twelve UCI data sets. The experimental results show that the proposed approach is effective and efficient.

Keywords: Feature selection, Hybrid data, Rough set theory, Large-scale data sets

1. Introduction

With the rapid development of information technologies including internet and databases, a very large number of data are acquired in many areas or industries, such as text and bioinformatics data. Both the size and the dimension of these data increase at an unprecedented rate, which has resulted in large-scale data with high dimension. Feature selection is an important technique used in dimensional reduction [3, 4, 6, 18, 19, 22]. It aims to improve the accuracy and performance of classifiers through removing redundant features and selecting informative features from the data. Feature selection has been successfully used in many areas and has attracted much attention in recent years [13, 14, 32, 38]. The rapid growth of data brings new challenges for traditional feature selection, and exploring efficient feature selection approaches has quickly become a key issue in machine learning [20, 39, 43, 47, 54].

In the process of feature selection, feature evaluation criteria are used to evaluate the quality of the candidate subsets. For a feature subset, different evaluation criteria may give different results. Roughly speaking, there are five kinds of evaluation criteria [8, 23]. They are distance measures, information measures, dependency measures, consistency measures, and classification error rate measures. The first four evaluation criteria are used to evaluate feature subsets according to inherent characteristics of the data. The last one relies on a classification algorithm to evaluate and select useful features [1, 25]. Obviously, compared with the first four evaluation criteria, using the last evaluation criteria can usually improve classification performance, but is also time-consuming. Because of different evaluation criteria being used in feature selection algorithms, existing feature selection algorithms are divided into three categories: the filter model, the wrapper model, and the hybrid model [3, 18, 25]. The "filter" algorithm model relies on the aforementioned first four evaluation criteria to select features. The "wrapper" algorithm model uses classification algorithms to evaluate candidate features. The hybrid model combines the advantages of "filter" and "wrapper" models by employing different evaluation criteria in a feature selection algorithm. With the development of feature selection and its deep research, the algorithms which can be used to deal with labeled data are called supervised feature selection algorithms [3]. The algorithms used to deal with unlabeled data are called unsupervised feature selection algorithms [5]. In addition, with the rise of big data, semi-supervised feature selection algorithms are gradually introduced to handle the small-labeled-sample problem in which unlabeled data is much more than labeled data [2]. For given data sets, feature types include numeric and nominal. To deal with nominal data, Wang

*Corresponding author. Tel./Fax: +86 0351 7018176.
Email addresses: sxuwangfeng@126.com (Feng Wang), ljy@sxu.edu.cn (Jiye Liang)