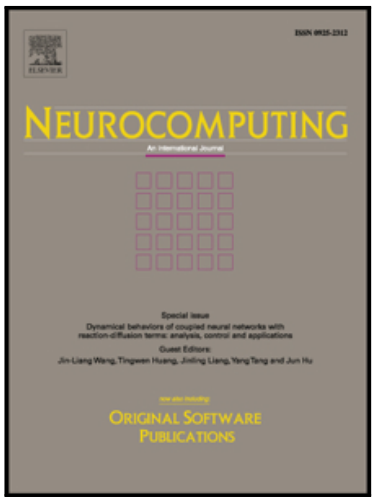

Feature selection in machine learning: a new perspective

Jie Cai , Jiawei Luo , Shulin Wang , Sheng Yang

PII: S0925-2312(18)30291-1
DOI: 10.1016/j.neucom.2017.11.077
Reference: NEUCOM 19407

To appear in: *Neurocomputing*

Received date: 6 August 2017
Revised date: 15 October 2017
Accepted date: 17 November 2017

Please cite this article as: Jie Cai , Jiawei Luo , Shulin Wang , Sheng Yang , Feature selection in machine learning: a new perspective, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2017.11.077

Feature selection in machine learning: a new perspective

Jie Cai, Jiawei Luo, Shulin Wang, Sheng Yang*

College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China

***Abstract*: High-dimensional data analysis is a challenge for researchers and engineers in the fields of machine learning and data mining. Feature selection provides an effective way to solve this problem by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding for the learning model or data. In this study, we discuss several frequently-used evaluation measures for feature selection, and then survey supervised, unsupervised, and semi-supervised feature selection methods, which are widely applied in machine learning problems, such as classification and clustering. Lastly, future challenges about feature selection are discussed.**

***Keywords*: Feature Selection; Dimensionality reduction; Machine learning; Data mining**

1. Introduction

With the rapid development of modern technology, tremendous new computer and internet applications have generated large amounts of data at an unprecedented speed, such as video, photo, text, voice, and data obtained from social relations and the rise of the Internet of things and cloud computing. These data often have the characteristics of high dimensions, which poses a high challenge for data analysis and decision-making. Feature selection has been proven in both theory and practice effective in processing high-dimensional data and in enhancing learning efficiency [1-3].

Feature selection is referred to the process of obtaining a subset from an original feature set according to certain feature selection criterion, which selects the relevant features of the dataset. It plays a role in compressing the data processing scale, where the redundant and irrelevant features are removed. Feature selection technique can pre-process learning algorithms, and good feature selection results can improve learning accuracy, reduce learning time, and simplify learning results [4-6]. Notably, feature selection and feature extraction [7-9] are two ways to dimensionality reduction. Unlike feature selection, feature extraction usually needs to transform the original data to features with strong pattern recognition ability, where the original data can be regarded as features with weak recognition ability.

Feature selection, which has been a research topic in methodology and practice for decades, is used in many fields, such as image recognition [10-14], image retrieval [15-17], text mining [18-20], intrusion detection [21-23], bioinformatic data analysis [24-31], fault diagnosis [32-34], and so on. According to the theoretical principle, feature selection methods can be based on statistics [35-39], information theory [40-45], manifold [46-48], and rough set [49-53], and can be categorized according to various standards.

a) According to the utilized training data (labeled, unlabeled, or partially labeled), feature selection methods can be divided into supervised, unsupervised, and semi-supervised models. A unified framework for supervised, unsupervised and semi-supervised feature selection is shown in Fig. 1.

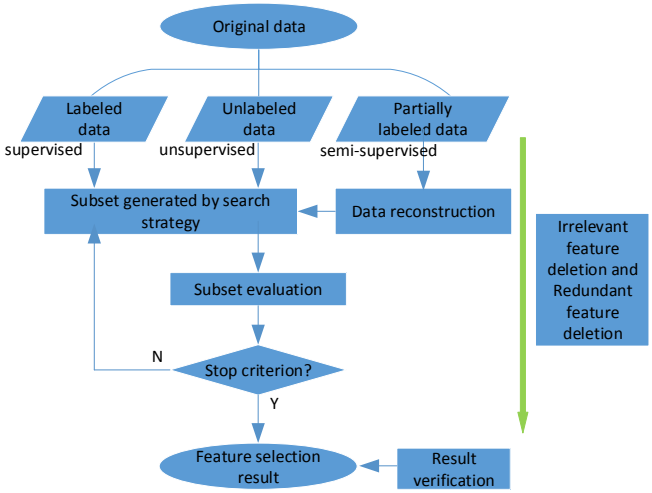

Fig. 1. A framework for feature selection

b) According to their relationship with learning methods, feature selection methods can be classified into filter, wrapper, and embedded models.

c) According to the evaluation criterion, feature selection methods can be derived from correlation, Euclidean distance, consistency, dependence, and information measure.

d) According to the search strategies, feature selection

* Corresponding author.
E-mail addresses: yangsh0506@sina.com