Wrapper Feature Selection

Kyriacos Chrysostomou
Brunel University, UK

Manwai Lee
Brunel University, UK

Sherry Y. Chen
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

It is well known that the performance of most data mining algorithms can be deteriorated by features that do not add any value to learning tasks. Feature selection can be used to limit the effects of such features by seeking only the relevant subset from the original features (de Souza et al., 2006). This subset of the relevant features is discovered by removing those that are considered as irrelevant or redundant. By reducing the number of features in this way, the time taken to perform classification is significantly reduced; the reduced dataset is easier to handle as fewer training instances are needed (because fewer features are present), subsequently resulting in simpler classifiers which are often more accurate.

Due to the abovementioned benefits, feature selection has been widely applied to reduce the number of features in many data mining applications where data have hundreds or even thousands of features. A large number of approaches exist for performing feature selection including filters (Kira & Rendell, 1992), wrappers (Kohavi & John, 1997), and embedded methods (Quinlan, 1993). Among these approaches, the wrapper appears to be the most popularly used approach. Wrappers have proven popular in many research areas, including Bioinformatics (Ni & Liu, 2004), image classification (Puig & Garcia, 2006) and web page classification (Piramuthu, 2003). One of the reasons for the popularity of wrappers is that they make use of a classifier to help in the selection of the most relevant feature subset (John et al., 1994). On the other hand, the remaining methods, especially filters, evaluate the merit of a feature subset based on the characteristics of the data and statistical measures, e.g., chi-square, rather than the classifiers intended for use (Huang et al., 2007). Discarding the classifier when performing feature selection can subsequently result in poor classification performance. This is because the relevant feature subset will not reflect the classifier's specific characteristics. In this way, the resulting subset may not contain those features that are most relevant to the classifier and learning task. The wrapper is therefore superior to other feature selection methods like filters since it finds feature subsets that are more suited to the data mining problem.

These differences between wrappers and other existing feature selection techniques have been reviewed by a number of studies (e.g. Huan & Lei, 2005). However, such studies have primarily focussed on providing a holistic view of all the different types of techniques. Although these works provide comprehensive information, there has yet to appear a deep review that solely focuses on the most popular feature selection technique; the wrapper. To this end, this paper aims to present an in-depth survey of the wrapper. In particular, attention will be given to improvements made to the wrapper since they are known for being much slower than other existing feature selection techniques. This is primarily because wrappers are required to repeatedly run the classifier when determining feature accuracy and perform feature selection again each time a different classifier is used. To overcome such problems, researchers in this area have spent considerable effort in improving the performance of wrappers (Yu & Cho, 2006). Basically, improvements can be divided into two trends. One focuses on reducing the time taken to do feature selection and the other emphasises on improv-

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

ing the accuracy of the selected subset of features. In fact, there are close relationships between these two trends because one can potentially influence the other. In other words, decreasing the time taken to perform feature selection with wrappers may potentially affect the accuracy of the final output. This relationship has been investigated by several studies and will be reviewed in this paper.

The paper will first formally define the feature selection process, with an emphasis on the wrapper approach. Subsequently, improvements made to the wrapper for reducing the time taken to do feature selection and increasing the overall accuracy of the selected subset of features will be discussed. It then moves to discuss future directions for wrapper feature selection approaches. Finally, conclusions are drawn at the end of the paper.

BACKGROUND

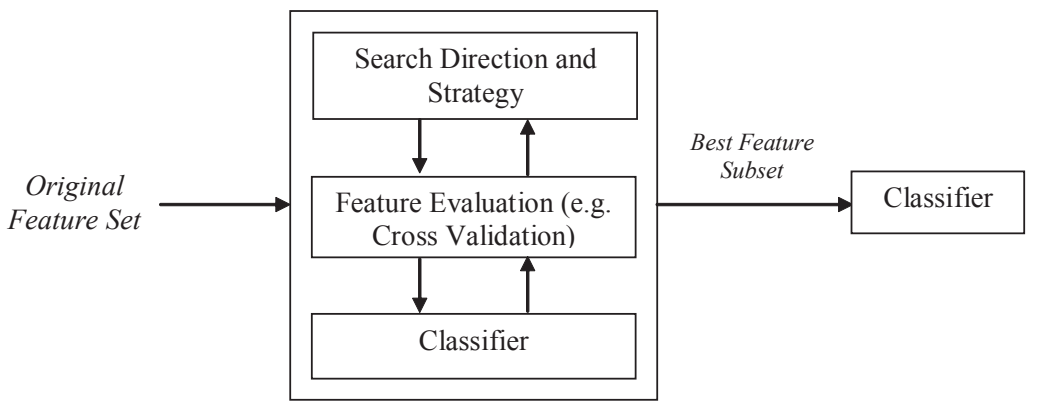
Typically, feature selection can be formally defined in the following manner. Suppose F is the given set of original features with cardinality n (where n symbolises the number of features in set F), and $\overline{F}$ is the selected feature subset with cardinality $\overline{n}$ (where $\overline{n}$ symbolises the number of features in set $\overline{F}$), then $\overline{F} \subseteq F$. Also, let $J(\overline{F})$ be the selection criterion for selecting feature set $\overline{F}$. We assume that a higher value of J indicates a better feature subset. Thus, the goal is to maximise $J()$. The problem of feature selection is to find a subset of features $\overline{F} \subseteq F$ such that,

$$J(\overline{F}) = \max_{Z \subseteq F, |Z|=n} J(Z)$$

Deriving a feature subset that maximises $J(\ )$ typically consists of four key steps namely, search direction, search strategy, subset evaluation and stopping criterion (Huan & Lei, 2005). Search direction defines the point at which the search for the most relevant subset will begin. Complimentary to the direction of the search is the search strategy. The search strategy outlines the way in which feature subsets are searched within the feature space. Each of the feature subsets found is then evaluated according to some evaluation criteria. Finally, a stopping criterion is used for halting the search through feature subsets.

As indicated in the Introduction, this paper will focus on the wrapper approach, where a classifier is used as the evaluation criterion for maximising $J(\ )$. Basically, the wrapper uses the classifier as a black box. The classifier is repeatedly run on the dataset using various subsets of the original features. These feature subsets are found through the use of a search strategy. The classifier's performance and some accuracy estimation method like cross validation are then used to evaluate the accuracy of each subset (John et al., 1994). The feature subset with the highest accuracy is chosen as the final set on which to run the classifier.

Figure 1. Wrapper feature selection



4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/wrapper-feature-selection/11110

Related Content

Data Mining for Lifetime Value Estimation
Silvia Figini (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 431-437).*
www.irma-international.org/chapter/data-mining-lifetime-value-estimation/10856

Quantization of Continuous Data for Pattern Based Rule Extraction
Andrew Hamilton-Wrightand Daniel W. Stashuk (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1646-1652).*
www.irma-international.org/chapter/quantization-continuous-data-pattern-based/11039

Process Mining to Analyze the Behaviour of Specific Users
Laura Maruster (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1589-1597).*
www.irma-international.org/chapter/process-mining-analyze-behaviour-specific/11031

Data Mining for Fraud Detection System
Roberto Marmo (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 411-416).*
www.irma-international.org/chapter/data-mining-fraud-detection-system/10853

Learning from Data Streams
João Gamaand Pedro Pereira Rodrigues (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1137-1141).*
www.irma-international.org/chapter/learning-data-streams/10964