

Chapter 16

Population-Based Feature Selection for Biomedical Data Classification

Seyed Jalaleddin Mousavirad
University of Kashan, Iran

Hossein Ebrahimpour-Komleh
University of Kashan, Iran

ABSTRACT

Classification of biomedical data plays a significant role in prediction and diagnosis of disease. The existence of redundant and irrelevant features is one of the major problems in biomedical data classification. Excluding these features can improve the performance of classification algorithm. Feature selection is the problem of selecting a subset of features without reducing the accuracy of the original set of features. These algorithms are divided into three categories: wrapper, filter, and embedded methods. Wrapper methods use the learning algorithm for selection of features while filter methods use statistical characteristics of data. In the embedded methods, feature selection process combines with the learning process. Population-based metaheuristics can be applied for wrapper feature selection. In these algorithms, a population of candidate solutions is created. Then, they try to improve the objective function using some operators. This chapter presents the application of population-based feature selection to deal with issues of high dimensionality in the biomedical data classification. The result shows that population-based feature selection has presented acceptable performance in biomedical data classification.

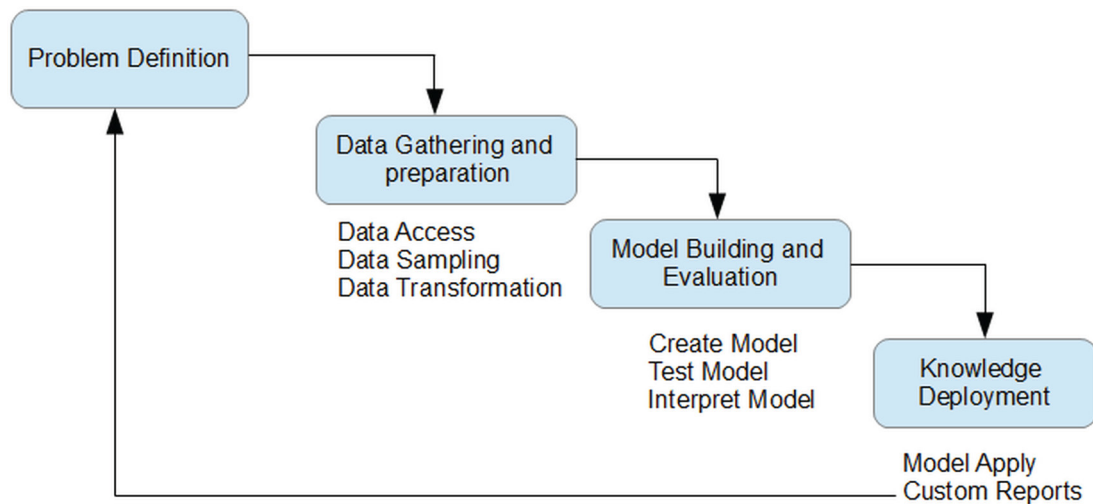
INTRODUCTION

Data mining or knowledge discovery is a computational process of extracting hidden knowledge in large databases. The goal of data mining process is to extract useful information from a dataset. Figure 1 illustrates the phases of a data mining

process. The first step in data mining process is to understanding of the problem. In the next step, data collect and prepare. In this step, data is cleaned from outlier instances or missing data and dataset reduces to only variables that are useful in a given data mining process. In the third step, a mining model or model is built. The quality of a model

DOI: 10.4018/978-1-4666-6086-1.ch016

Figure 1. The data mining process



can evaluate using a number of the techniques. The last step in the data mining process is to deploy the models to a real environment.

Data mining techniques have been successfully used in various biomedical domains, for example the detection of tumors, the diagnosis of cancers and other diseases. One of the main challenge in biomedical data mining and analysis is the so called “curse of dimensionality”. Especially the biomedical data are presented by relatively few instances and exhibited in a high dimensional feature space(Peng, Wu, & Jiang, 2010). Feature selection, a process in data transformation phase, reduces the number of features, removes irrelevant, redundant and misleading features, which leads to expediting learning algorithm and improves predictive performance. Feature selection algorithms are divided into three categories: wrapper methods that uses the learning algorithms to evaluate the usefulness of features, filter methods that evaluate features according to the statistical characteristics of the data, and embedded methods that feature selection embed in the learning algorithm. Population based metaheuristics such as genetic algorithm, particle swarm optimization, Imperialist competitive algorithm, artificial bee

algorithm, Ant colony optimization, and leap frog optimization have been considered as effective wrapper feature selection approach. These meta-heuristics are based on a population of solutions and an iterative procedure. At each iteration, they try to find a better solution than previous solutions using some operators. Feature selection algorithms have been successfully applied in various biomedical domains. A. Antoniadis et al, (2003) presented a statistical feature reduction approach for the classification of tumors. I. Guyan et al, (2002) address the problem of selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays. Using available training examples from cancer and normal patients, they build a classifier suitable for genetic diagnosis, as well as drug discovery. In another work, wrapper approaches was applied for gene selection(Blanco, Larrañaga, Inza, & Sierra, 2004). Y. Peng et al. (2010) presents a novel feature selection approach to deal with issues of high dimensionality in biomedical data classification. The approach proposed in this paper integrated filter and wrapper methods into a sequential search procedure with the aim to improve the classification performance of the

29 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/population-based-feature-selection-for-biomedical-data-classification/109988

Related Content

SOM-Based Clustering of Multilingual Documents Using an Ontology

Minh Hai Pham, Delphine Bernhard, Gayo Diallo, Radja Messai and Michel Simonet (2008). *Data Mining with Ontologies: Implementations, Findings, and Frameworks* (pp. 65-82).

www.irma-international.org/chapter/som-based-clustering-multilingual-documents/7572

Virtual Telemedicine and Virtual Telehealth: A Natural Language Based Implementation to Address Time Constraint Problem

Shazia Kareem and Imran Sarwar Bajwa (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 881-892).

www.irma-international.org/chapter/virtual-telemedicine-virtual-telehealth/73475

A Methodology Supporting the Design and Evaluating the Final Quality of Data Warehouses

Maurizio Pighin and Lucio Ieronutti (2008). *International Journal of Data Warehousing and Mining* (pp. 15-34).

www.irma-international.org/article/methodology-supporting-design-evaluating-final/1811

Query Interaction Based Approach for Horizontal Data Partitioning

Ladjel Bellatreche and Amira Kerkad (2015). *International Journal of Data Warehousing and Mining* (pp. 44-61).

www.irma-international.org/article/query-interaction-based-approach-for-horizontal-data-partitioning/125650

TBSGM: A Fast Subgraph Matching Method on Large Scale Graphs

Fusheng Jin, Yifeng Yang, Shuliang Wang, Ye Xue and Zhen Yan (2018). *International Journal of Data Warehousing and Mining* (pp. 67-89).

www.irma-international.org/article/tbsgm/215006