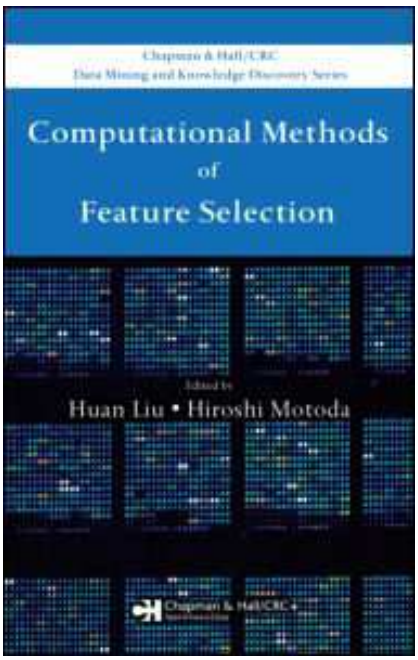
Computational Methods of Feature Selection

HUAN LIU AND HIROSHI MOTODA



REVIEWED BY LONGBING CAO AND DAVID TANIAR

Feature selection selects a subset of relevant features, and also removes irrelevant and redundant features from the data to build robust learning models. Feature selection is very important, not only because of the curse of dimensionality, but also due to emerging data complexities and quantities faced by multiple disciplines, such as machine learning, data mining, pattern recognition, statistics, bioinformatics, and text mining.

In recent years, we have seen extensive and productive efforts on feature selection. The research has been expanding from simple to complex feature types, from supervised to unsupervised and semi-supervised feature selection, and from simple to more advanced techniques, both in depth and in breadth.

"Computational Methods of Feature Selection", edited by H. Liu and H. Motoda, two leading experts in the field, collects recent research works from various disciplines on computational methods in feature selection and extraction. The collection reflects the advancements in recent years, following the editors' pioneer book on feature selection published in 1998. Consequently, these publications provide a comprehensive roadmap of feature selection research, and this is certainly very helpful to a very wide audience from beginners to professionals, and from practitioners to researchers.

The collection features a state-of-the-art survey, technique advancement, practical guides, promising directions, and case studies. It ranges from presenting background and fundamentals relevant to feature selection, to recent results in extending feature selection, weighting and local methods for feature selection, as well as feature selection progress in text mining and bioinformatics, organized in five independent parts. The content, carefully selected from invited contributors, relevant workshops and tutorials, covers areas such as text classification, web mining, bioinformatics and high-dimensional data.

The book starts with an introduction and background material, which consists of four chapters. A very insightful and enjoyable overview on background and the basics of feature selection are presented in Chapter 1. An evolutionary picture is drawn on feature selection development from supervised to unsupervised and semi-supervised learning in order to handle the increasing mixture of labeled, unlabeled and partially labeled data. An overview of unsupervised feature selection is presented in Chapter 2, which highlights the identification of the smallest feature subset that best uncovers interesting and natural clusters based on certain criteria. Filter and wrapper methods are introduced to select features in unlabeled data, while subspace clustering and co-clustering/bi-clustering are discussed from the local unsupervised feature selection perspective.

Randomization is widely used in feature selection when appropriate choices can be managed. A survey on randomized feature selection is presented in Chapter 3. Two types of randomization, namely Las Vegas and Monte Carlo algorithms, are introduced, followed by an overview of three complexity classes defining the probabilistic requirements in analyzing randomized algorithms. The work features six illustrations of randomized features and prototype algorithms for feature selection problems. Another factor that may facilitate feature selection is causal relationship discovery for cutting down dimensionality and deep understanding of the underlying mechanism. Chapter 4 addresses non-causal and causal feature selection. With a definition of probabilistic causality, the causal Bayesian network is used to analyze feature relevance and further to design a causal discovery algorithm by finding the Markov blanket.

Recent advancements in feature selection are highlighted in Part II in a number of strategies. Firstly, Chapter 5 describes how an active feature value is acquired to estimate feature relevance in domains where feature values are expensive to measure. A sampling benefit function is derived from a statistical formulation of the problem, followed by an active sampling algorithm, which is shown to outperform random sampling by a mixture model for the joint class-feature distribution to reduce the number of feature samples. Secondly, from the feature extraction perspective, Chapter 6 presents the notion of the decision border, in which a labeled vector quantizer, that can efficiently be trained by the Bayes risk weighted vector quantization (BVQ) algorithm, is devised to extract the best linear approximation. It is shown that the approach gives comparable results to the SVM-based decision boundary and performs better than the multi-layer perceptron-based method.

Chapter 7 further explains how independent probe variables are used in the same distribution in generating a probe. Feature relevance is compared with the relevance of its randomly per-

muted probes for classification using random forests. The approach is promising in terms of data types and quantity, performance, and computational complexity. Finally, in Chapter 8, an incremental ranked usefulness is used to decide whether or not a feature is relevant in massive data, and then to select the best non-consecutive features from the ranking. The approach chooses a small subset of features with similar predictive performance to others in dealing with high-dimensional data.

Strategies and methods related to weighting and local methods are addressed in Part III. Firstly, the Relief family algorithms are described in Chapter 9. Relief is extended to a more realistic variant ReliefF to deal with incomplete data for classification, and is further extended to the Regressional ReliefF for regression problems. The variety of the Relief family shows its general applicability as a non-myopic feature quality measure. Feature selection in K-means is usually not automated. Chapter 10 proposes techniques to automatically determine the important features in K-means clustering. This is done through calculating the sum of the within-cluster dispersions of the feature, and renewing the weights in an iterative process.

In contrast to maximum benefit-based active feature sampling, Chapter 11 focuses on local feature relevance and weighting by designing adaptive metrics or parameter estimates that are local in an input space. Chapter 12 presents a mathematical interpretation of the Relief algorithms. It is proven to be equivalent to solving an online convex optimization problem with a margin-based objective function. New feature weighting algorithms are then proposed to find the nearest neighbor classifier.

In Part IV, text feature selection is addressed by a survey, a new feature selection score, and constraint-guided and aggressive feature selection approaches. Firstly, Chapter 13 presents a comprehensive overview of feature selection for text classification, including feature generation, representation, and selection, with illustrative examples, from a pragmatic viewpoint. Text feature generators, such as word merging, word phrases, character N-grams, and multi-field records are introduced. An introduction to classification feature filtering is also provided. Secondly, Chapter 14 introduces a new feature selection score, namely posterior inclusion probability under Bernoulli and Poisson distributions. The score is defined as the posterior probability of including a given feature over all possible models, in which each model corresponds to a different set of features that includes the given feature. The advantage of the score is that the selected features are easy to interpret while maintaining comparable performance to other typical score metrics, such as information gain.

Two different pairwise constraint-guided dimensionality reduction approaches, through projecting data into a lower space and co-clustering of features and data, are introduced in Chapter 15. Investigations are also conducted on improving semi-supervised clustering performance in high-dimensional data. In Chapter 16, an aggressive feature selection method is proposed, which can filter more than 95% features for text mining. To handle feature redundancy, information gain-based ranking for text classification is also proposed using a mutual information measure and inclusion index.

The last section covers feature selection in bioinformatic data, which may not be effectively handled by general feature selection approaches. This part consists of four chapters. Chapter 17 introduces the challenges of micro-array data analysis and presents a redundancy-based feature selection algorithm. A Markov blanket based filter method is proposed to approximate the selection of discriminative and non-redundant genes. In Chapter 18, a scalable method based on sequence components and domain knowledge is developed to generate automatic features on biological sequence data. The algorithm can construct features, explore the space of possible features, and identify the most useful ones. Chapter 19 proposes an ensemble-based method to find robust features for biomarker discovery. Ensembles are obtained by choosing different alternatives at each stage of data mining from normalization to binning, feature selection, and classification. Finally, a penalty-based feature selection method is proposed in Chapter 20 to produce a sparse model by utilizing the grouping effect. As a generalization of a penalized least squares method, lasso, the proposed approach is promising in handling high-dimensional data for various purposes, such as regression and classification problems.

Overall, we enjoyed reading this book. It presents state-of-the-art guidance and tutorials on methodologies and algorithms in computational methods in feature selection. Enhanced by the editors insights, and based on previous work by these leading experts in the field, the book forms another milestone of relevant research and development in feature selection. The selected chapters also present interesting open issues and promising directions for further exploration of feature selection in the next decade. With such a research roadmap, it is highly exciting to foresee the next generation of feature selection methodologies and techniques inspired by this collection.

About the reviewers:

LONGBING CAO: Data Sciences and Knowledge Discovery Laboratory, Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia. Contact him at lbcao@it.uts.edu.au, and http://www-staff.it.uts.edu.au/∼lbcao/ and datamining.it.uts.edu.au for more information.

DAVID TANIAR: Clayton School of Information Technology, Monash University, Australia. Contact him at David.Taniar@infotech.monash.edu.au, and http://users.monash.edu.au/∼dtaniar/ for more information.