IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8923

Chapter IV

Feature Selection in Data Mining

YongSeog Kim
University of Iowa, USA

W. Nick Street
University of Iowa, USA

Filippo Menczer
University of Iowa, USA

ABSTRACT

Feature subset selection is an important problem in knowledge discovery, not only for the insight gained from determining relevant modeling variables, but also for the improved understandability, scalability, and, possibly, accuracy of the resulting models. The purpose of this chapter is to provide a comprehensive analysis of feature selection via evolutionary search in supervised and unsupervised learning. To achieve this purpose, we first discuss a general framework for feature selection based on a new search algorithm, Evolutionary Local Selection Algorithm (ELSA). The search is formulated as a multi-objective optimization problem to examine the trade-off between the complexity of the generated solutions against their quality. ELSA considers multiple objectives efficiently while avoiding computationally expensive global comparison. We combine ELSA with Artificial Neural Networks (ANNs) and Expectation-Maximization (EM) algorithms for feature selection in supervised and unsupervised learning respectively. Further, we provide a new two-level evolutionary algorithm, Meta-Evolutionary Ensembles (MEE), where feature selection is used to promote the diversity among classifiers in the same ensemble.

This chapter appears in the book, *Data Mining: Opportunities and Challenges,* edited by John Wang. Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often builds a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, based on the selected features, a physician may decide whether a risky surgical procedure is necessary for treatment or not.

Feature selection in supervised learning where the main goal is to find a feature subset that produces higher classification accuracy has been well studied. Recently, several researchers (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998; Devaney & Ram, 1997; Dy &Brodley, 2000b) have studied feature selection and clustering together with a single or unified criterion. For feature selection in unsupervised learning, learning algorithms are designed to find natural grouping of the examples in the feature space. Thus, feature selection in unsupervised learning aims to find a good subset of features that forms high quality clusters for a given number of clusters.

However, the traditional approaches to feature selection with a single evaluation criterion have shown limited capability in terms of knowledge discovery and decision support. This is because decision-makers should take into account multiple, conflicted objectives simultaneously. No single criterion for unsupervised feature selection is best for every application (Dy & Brodley, 2000a), and only the decision-maker can determine the relative weights of criteria for her application. In order to provide a clear picture of the possibly nonlinear trade-offs among the various objectives, feature selection has been formulated as a *multi-objective* or *Pareto* optimization problem.

In this framework, we evaluate each feature subset in terms of multiple objectives. Each solution s_i is associated with an evaluation vector $F = F_1(s_i),..., F_C(s_i)$ where C is the number of quality criteria. One solution s_1 is said to *dominate* another solution s_2 if $\forall c: F_C(s_1) \geq F_C(s_2)$ and $\exists c: F_C(s_1) > F_C(s_2)$, where F_C is the c-th criterion, $c \in \{1,...,C\}$. Neither solution dominates the other if $\exists c_1, c2: F_{C1}(s_1) > F_{C2}(s_2), F_{C2}(s_2) > F_{C2}(s_1)$. We define the *Pareto front* as the set of nondominated solutions. In feature selection such as a Pareto optimization, the goal is to approximate the *Pareto front* as best as possible, presenting the decision-maker with a set of high-quality solutions from which to choose.

We use Evolutionary Algorithms (EAs) to intelligently search the space of possible feature subsets. A number of multi-objective extensions of EAs have been proposed (VanVeldhuizen, 1999) to consider multiple fitness criteria effectively. However, most of them employ computationally expensive selection mechanisms to favor dominating solutions and to maintain diversity, such as Pareto domination tournaments (Horn, 1997) and fitness sharing (Goldberg & Richardson, 1987). We propose a new algorithm, Evolutionary Local Selection Algorithms (ELSA), where an individual solution is allocated to a local environment based on its criteria values and competes with others to consume shared resources only if they are located in the same environment.

The remainder of the chapter is organized as follows. We first introduce our search algorithm, ELSA. Then we discuss the feature selection in supervised and unsupervised learning, respectively. Finally, we present a new two-level evolutionary environment, Meta-Evolutionary Ensembles (MEE), that uses feature selection as the mechanism for boosting diversity of a classifier in an ensemble.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

24 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/feature-selection-data-mining/7597

Related Content

Incremental Algorithm for Discovering Frequent Subsequences in Multiple Data Streams

Reem Al-Mullaand Zaher Al Aghbari (2013). *Developments in Data Extraction, Management, and Analysis (pp. 259-279).*

www.irma-international.org/chapter/incremental-algorithm-discovering-frequent-subsequences/70801

On the Advancement of Using Data Mining for Crime Situation Recognition: A Comparative Review

Omowunmi E. Isafiade, Antoine Bagulaand Sonia Berman (2016). *Data Mining Trends and Applications in Criminal Science and Investigations (pp. 1-31).*

www.irma-international.org/chapter/on-the-advancement-of-using-data-mining-for-crime-situation-recognition/157451

Application of Improved Chameleon Swarm Algorithm and Improved Convolution Neural Network in Diagnosis of Skin Cancer

Wu Beibeiand Nikolaj Jade (2023). *International Journal of Data Warehousing and Mining (pp. 1-16).*

www.irma-international.org/article/application-of-improved-chameleon-swarm-algorithm-and-improved-convolution-neural-network-in-diagnosis-of-skin-cancer/325059

Identification of Genomic Islands by Pattern Discovery

Nita Parekh (2012). *Pattern Discovery Using Sequence Data Mining: Applications and Studies (pp. 166-181).*

www.irma-international.org/chapter/identification-genomic-islands-pattern-discovery/58679

Mass Media Strategies: Hybrid Approach using a Bioinspired Algorithm and Social Data Mining

Carlos Alberto Ochoa Ortiz Zezzatti, Darwin Young, Camelia Chira, Daniel Azpeitiaand Alán Calvillo (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications (pp. 1163-1188).*

www.irma-international.org/chapter/mass-media-strategies/73490