Chapter 10
A Promising Direction towards Automatic Construction of Relevance Measures

Lucianne Varn
Independent Researcher, New Zealand

Kourosh Neshatian
University of Canterbury, New Zealand

ABSTRACT

A relevance measure is a measure over the space of features of a learning problem that quantifies the degree of relatedness of a single feature or a subset of features to a target variable. The measure can be used to both detect relevant features (when the target variable is the response variable) and detect redundant features (when the target variable is another input feature). Measuring relevance and redundancy is a central concept in feature selection. In this chapter, the authors show that there is a lack of generality in the features selected based on heuristic relevance measures. Through some counter-examples, the authors show that regardless of the type of heuristic measure and search strategy, heuristic methods cannot optimise the performance of all learning algorithms. They show how different measures may have different notions of relevance between features and how this could lead to not detecting important features in certain situations. The authors then propose a hyper-heuristic method that through an evolutionary process automatically generates an appropriate relevance measure for a given problem. The new approach can detect relevant features in difficult scenarios.

INTRODUCTION

High dimensionality is not usually a desired situation in the context of machine learning and data mining. Often the need for more training examples grows exponentially with respect to the number of dimensions in a problem—an effect known as *the curse of dimensionality*. High dimensionality makes the hypothesis space bigger (again often exponentially), which makes finding a good hypothesis computationally more challenging. Feature selection, a practice usually carried out at the preprocessing stage, deals with the high dimensionality issue. While feature selection is

DOI: 10.4018/978-1-4666-6078-6.ch010

not formally defined in the literature, it informally refers to the process of finding a minimal subset of features that is sufficient to solve a learning problem. The sufficiency criterion may refer to improving learning performance (with some definition of performance), maintaining performance at some acceptable level, or even other criteria regarding model complexity, intelligibility, etc.

Feature selection algorithms have, in abstract terms, two main components. The first component is a search mechanism that searches the space of power sets of features which grow exponentially ($O(2^n)$) with respect to the number of features in problems. The second component is an evaluation mechanism which measures the goodness of (candidate) subsets of features. There are two major approaches for evaluation: wrapper and filter (or non-wrapper) (Kohavi & John, 1997). In the wrapper approach, the performance of a learning algorithm (e.g. a decision tree inducer) is used to guide the search. The wrapper approach is computationally intensive; every evaluation involves training and testing a model. In the filter approach, instead of using a learner's performance as a measure of the utility of a candidate subset of features, computationally-cheap heuristics are incorporated. The most common measure of utility in the filter approach is relevance. Relevance quantifies the degree of relatedness between a subset of features and another feature (that does not exist in the subset). Features with a significant degree of relevance to target concepts (such as class labels) are desired, while features with a considerable degree of relevance to each other are considered redundant and thus unwanted. Examples of commonly-used heuristic relevance measures are those based on information theory such as Information Gain (IG) and Information Gain Ratio (IGR) (Last, K, & Maimon, 2001), and those based on statistical methods such as $\chi 2$ (Chi-square) ranking (Liu & Setiono, 1995) and Logistic Regression (Cheng, Varshney, & Arora, 2006).

Filter-based feature selection methods are known to be computationally efficient in comparison with methods taking the wrapper approach. Since filter methods do not use any learning algorithms directly, they are usually described as being "independent of any learning algorithms" (Kohavi & John, 1997). However, it is unclear whether the importance (utility) of features can be determined independently from any learning algorithms. Clearly, filter methods have improved the performance of some learning algorithms over some problems, but a question that remains to be answered is whether "independence from learning algorithms" implies that a highly relevant subset of features found by a filter method is expected to optimise the learning performance of any arbitrary learning algorithm. If the answer is 'no', then what can be done? This chapter investigates these issues and proposes a solution.

PROBLEM STATEMENT

Let $\mathcal{D}$ represent the set of all possible observations in a classification domain; for example $\mathcal{D}$ could be the population of patients receiving a medical diagnosis. A feature (or attribute) is a mapping from $\mathcal{D}$ to a *co-domain*; for example, *height* and *gender* as features can be mappings of the form $height : \mathcal{D} \rightarrow \mathbb{R}^+$ and

$$gender : \mathcal{D} \rightarrow \{male, female\}.$$

If d is a member of the population (a data item), then *height(d)* and *gender(d)* represent the value of the two features for the given data item.

We use $\mathcal{F}$ to represent the set of all features that are available (defined or measurable) for all members of the population. In a supervised learning context $\mathcal{F}$ is partitioned into two sets $\mathcal{X} = \{X_1, X_2, \ldots, X_{|\mathcal{X}|}\}$ and $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_{|\mathcal{Y}|}\}$ such that $\mathcal{X} \cap \mathcal{Y} = \varnothing$, $\mathcal{X} \cup \mathcal{Y} = \mathcal{F}$ and thus