Expert Systems with Applications 37 (2010) 602–609

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A decision rule-based method for feature selection in predictive data mining

Patricia E.N. Lutu [a,*], Andries P. Engelbrecht [b]

[a] Department of Informatics, University of Pretoria, South Africa
[b] Department of Computer Science, University of Pretoria, South Africa

ARTICLE INFO

*Keywords:*
Feature selection
Feature subset search
Predictive data mining

ABSTRACT

Algorithms for feature selection in predictive data mining for classification problems attempt to select those features that are relevant, and are not redundant for the classification task. A relevant feature is defined as one which is highly correlated with the target function. One problem with the definition of feature relevance is that there is no universally accepted definition of what it means for a feature to be *'highly correlated with the target function or highly correlated with the other features'*. A new feature selection algorithm which incorporates domain specific definitions of high, medium and low correlations is proposed in this paper. The proposed algorithm conducts a heuristic search for the most relevant features for the prediction task.

## 1. Introduction

Algorithms for feature selection in predictive data mining for classification problems attempt to select those features that are relevant, and are not redundant for the classification task. A relevant feature is defined as one which is highly correlated with the target function (Blum & Langley, 1997; Hall, 1999). A redundant feature is defined as one which is highly correlated with other features (Hall, 1999; Ooi, Chetty, & Teng, 2007). One problem with the definition of feature relevance is that there is no universally accepted definition of what it means for a feature to be *'highly correlated with the target function or highly correlated with the other features'*. Different fields of enquiry use different thresholds for correlation values to distinguish between high and low correlations (Cohen, 1988).

The correlation based feature selection algorithms that have been reported in the literature, such as correlation-base feature selection (CFS) (Hall, 1999) and differential prioritisation (DP) (Ooi et al., 2007), employ heuristic search procedures which use mathematical functions to compute measures of merit which provide a high value for relevant feature subsets and low values for non-relevant feature subsets. The problem with these algorithms is that, the mathematical functions they use are lacking in flexibility and precision. Since these algorithms do not use precise definitions of high and low correlation, they cannot be used to perform feature selection based on domain meanings of high and low correlations. Secondly, as will be demonstrated in this paper, they can make bad decisions on which features are most relevant.

It is desirable to use a feature subset selection algorithm which will first of all select features based on precise definitions of high correlation and low correlation. Secondly, the algorithm should never select pure noise or prefer pure noise over features which have a high or medium correlation to the target function. The required level of precision can be achieved by determining the merit of a feature subset using logic that is implemented as a programmed function. In order to provide flexibility, parameters are used in the function, so that the user can provide threshold values to distinguish between a high, medium and low correlation.

In this paper, an algorithm which does precisely this, is proposed. The algorithm incorporates user specified thresholds as well as decision rules, in order to select feature subsets. Experimental results are presented to demonstrate the weaknesses of the CFS and DP feature selection procedures, and to demonstrate how the proposed algorithm eliminates these problems. The rest of the paper is organised as follows. Section 2 provides a review of some of the literature on feature selection. Section 3 provides a description of the proposed feature selection algorithm, based on decision rules. Section 4 provides a description of the experiments that were conducted, as well as the experimental results. Section 5 concludes the paper.

## 2. Background

Feature relevance and selection for classification, is a well studied problem (Aha & Bankert, 1996; Blum & Langley, 1997; Hall & Holmes, 2003; Koller & Sahami, 1996; Langley & Iba, 1993). Feature selection is concerned with the identification of a subset of features which are most relevant to the prediction problem and are not redundant. For some classification algorithms, irrelevant and redundant features can have a devastating effect on sample

* Corresponding author. Tel.: +27 124203373; fax: +27 123625287.
*E-mail addresses:* Patricia.Lutu@up.ac.za (P.E.N. Lutu), engel@cs.up.ac.za (A.P. Engelbrecht).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.06.031

complexity. Langley and Iba (1993) have conducted an average-case analysis and concluded that, for nearest-neighbour classification, the number of examples needed to obtain a given level of accuracy grows exponentially with the number of irrelevant features. The curse of dimensionality is also a well known problem in machine learning and statistical pattern recognition.

There are two categories of feature selection methods namely: filters and wrappers. The research reported here is concerned with filtering methods. Wrapper methods incorporate model construction with feature selection (Kohavi, 1995; Blum & Langley, 1997). These methods will select that subset of features which provides the highest level of predictive accuracy. Filtering methods on the other hand, select feature subsets without constructing predictive models for these feature subsets (Hall, 1999; Ooi et al., 2007). The filtering methods may further be categorised as pure ranking methods or search methods. Ranking methods simply produce a ranking of the most relevant features and then select say, the top k features. Search methods use heuristic search algorithms which attempt to establish that subset of features that provide the highest level of relevance, and the lowest level of redundancy. Correlation-based feature selection is a commonly used method for feature selection (Hall, 1999; Ooi et al., 2007). For this approach, feature relevance is measured in terms of the strength of the correlation between a feature and the class variable. Feature redundancy is measured in terms of the strength of the correlation between a given feature and the other features. Pearson's correlation coefficient is commonly used as a measure for correlation.

### 2.1. Merit measures for heuristic search

Given a subset F of features, with $F = \{f_1 \ldots, f_k\}$, a mathematical function is typically used to compute a measure of merit which guides the heuristic search. The correlation-based feature selection (CFS) method proposed by Hall (1999, 2000), uses the merit measure defined in equation (1).

$$Merit_{CFS} = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{1}$$

In Eq. (1), $\bar{r}_{cf}$ is the mean correlation between each feature and the class attribute, and, $\bar{r}_{ff}$ is the mean correlation between the features in subset F. The numerator on the RHS of equation (1) measures of the level of relevance of the feature subset, while the denominator measures the level of redundancy of the feature subset. The differential Prioritisation (DP) method, proposed by Ooi et al. (2007) uses the merit measure defined in Eqs. (2) and (3).

$$Merit_{DP} = (\bar{r}_{cf})^{\alpha}.(RD)^{1-\alpha} \tag{2}$$

$$RD = \frac{1}{k^2} \sum_{f_i, f_j \in F, i \neq j} 1 - |r(f_i, f_j)| \tag{3}$$

In Eq. (2), the first term on the RHS measures the level of relevance, while the second term measures the level of redundancy of the feature subset. The parameter $\alpha$ is used to control the levels of relevance and redundancy based on the user's preference. In Eq. (3), the term $r(f_i, f_j)$ represents the correlation coefficient between two features $f_i$ and $f_j$.

The correlation coefficients in Eqs. (1)–(3) are computed using either Pearson's correlation coefficient for quantitative features or the symmetrical uncertainty coefficient for qualitative features. For two quantitative features X and Y, the correlation is measured using Pearson's correlation coefficient, which is defined in Eq. (4).

$$r_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_X S_Y} \tag{4}$$

In Eq. (4), $\bar{x}$ and $\bar{y}$ are the sample means for X and Y respectively, $S_X$ and $S_Y$ are the sample standard deviations for X and Y, and n is the size of the sample used to compute the correlation coefficient.

To compute a correlation coefficient between two qualitative features X and Y, the symmetrical uncertainty coefficient $SU$, defined in Eq. (5), is used. In Eq. (5), $E(X)$ is the entropy in $X$ and $E(X|Y)$ is the entropy of $X$ conditioned on $Y$.

$$SU = 2.0\left[\frac{E(X) - E(X|Y)}{E(X) + E(Y)}\right] \tag{5}$$

When one feature $(X)$ is qualitative and the other feature $(Y)$ is quantitative, a weighted Pearson's correlation is used. For the qualitative feature $X$, if $X$ has $v$ levels, $L_1 \ldots L_V$, then $V$ binary features $B_1 \ldots B_V$ are created through a process called binarisation, and then each of the binary features is correlated with the quantitative feature, $Y$. The binary feature $B_i$ is assigned the value 1 when $X$ has

**Table 1**
Common definition of feature relevance and redundancy.

| Situation | Class–feature correlation for *f* | Mean feature–feature correlation for *f* | Interpretation |
|---|---|---|---|
| 1 | Not high | Not high | Irrelevant |
| 2 | Not high | High | Redundant |
| 3 | High | Not high | Relevant |
| 4 | High | High | Redundant |

**Table 2**
A possible definition of feature relevance and redundancy based on user specified levels.

| Situation | Category | Class–feature correlation for *f* | Mean feature–feature correlation for *f* | Proposed new interpretation |
|---|---|---|---|---|
| 1 | F | Insignificant | Any level | Irrelevant |
| 2 | C | Low | Insignificant | Weakly relevant |
| 3 | C | Low | Low | Weakly relevant |
| 4 | C | Low | Medium | Weakly relevant |
| 5 | F | Low | High | Irrelevant |
| 6 | B | Medium | Insignificant | Relevant |
| 7 | B | Medium | Low | Relevant |
| 8 | D | Medium | Medium | Weakly redundant |
| 9 | E | Medium | High | Redundant |
| 10 | A | High | Insignificant | Strongly relevant |
| 11 | A | High | Low | Strongly relevant |
| 12 | D | High | Medium | Weakly redundant |
| 13 | E | High | High | Redundant |