Comp. Appl. Math. (2013) 32:145–161
DOI 10.1007/s40314-013-0019-5

A hybrid feature selection scheme for mixed attributes data

Haitao Liu · Ruxiang Wei · Guoping Jiang

Received: 2 August 2012 / Accepted: 28 September 2012 / Published online: 15 March 2013

Abstract Feature selection aims at reducing the number of features in many applications. Existing feature selection approaches mainly deals with classification problems with continuous or discrete attributes. However, data usually come with mixed attributes in real-world applications. In this paper, a hybrid feature selection (HFS) scheme is proposed to deal with mixed attributes data. Firstly, a new correlation measure between mixed attributes is defined by giving a model for calculating mutual information between continuous and discrete attributes; secondly, the features are evaluated by a filter model with the new correlation measure; finally, feature selection is done by optimizing the parameter in the filter model with estimation accuracy criterion. Experimental results show that HFS acquires better stability and estimation accuracy.

Keywords Feature selection · Mixed attributes · Mutual information · Filter · wrapper · Case-based reasoning

1 Introduction

Feature selection (also known as variable selection or attribute selection) plays an important role in machine learning and pattern recognition (Hu et al. 2010; Guyon and Elisseeff 2003). It is to select some most effective features from the original feature set to reduce the dimension of the feature space according to certain criteria (Sheng 2000). By feature selection, some

Communicated by José Mario Martinez.

H. Liu (✉) · G. Jiang
Department of Equipment E&M, Naval University of Engineering,
Wuhan 430033, People's Republic of China
e-mail: liuhaitao0211@163.com

R. Wei
College of Science, Naval University of Engineering,
Wuhan 430033, People's Republic of China

irrelevant or redundant features are removed, thereby reducing the computational complexity, improving the estimation accuracy of the learning model and facilitating the intelligibility of the model (Amiri et al. 2011; Cakır et al. 2011).

A great number of feature selection approaches have been developed in recent years. Two key issues in constructing a feature selection approach are the search strategy and the evaluating criteria (Yao et al. 2012; Mao et al. 2007). According to the search strategy, global (Somol et al. 2004), heuristic (Dash and Liu 2003) and random (Oh et al. 2004) strategies were introduced in the literatures. An overall review on this issue is presented in Monirul Kabir et al. (2011). With respect to the evaluation criteria, feature selection approaches can be classified into three categories (Monirul Kabir et al. 2011): the filter, the wrapper and the hybrid approach. The wrapper approach (Hsu et al. 2002; Verikas and Bacauskiene 2002; Wang et al. 2008; Zhu et al. 2007) assesses feature subset with the training accuracy of the learning model. The filter approach (Ke et al. 2008; Sun 2007;Fleuret 2004) assesses features with statistical properties of the training data, and is independent from the learning model. In the hybrid approach (Hu et al. 2006; Hsu et al. 2011; Yang et al. 2011), features are first filtered, and then determined by the wrapper model. It is often found that, the hybrid approach is capable of locating a good solution, while a single technique often traps into an immature solution.

In another view of point, feature selection can also be classified into discrete and continuous approaches. Discrete approaches consider that all features take values in a finite set, while continous approaches assume that samples are described with a set of numerical variables. In a whole, existing feature selection approaches are mainly designed for classification problems with discrete or continuous attributes (Liu and Yu 2005; Dash and Liu 1997). However, actual data usually tend to have mixed attributes. For example in Software cost estimation, the data collected include both discrete and continuous attributes.

For mixed attributes data, existing approaches consist of two categories. One approach is to perform a discretization for continuous attribute (Ferreira and Figueiredo 2011), but the discretization brings an inevitable loss of information. Another approach is to do the granulation of mixed attributes (Hu et al. 2008a). In Hu et al. (2008b), the granulation approach is summarized, and neighborhood rough set is used to handle mixed attributes. Farther, the neighborhood mutual information is defined in (Hua et al. 2011) to do feature selection for high-dimensional mixed attributes data. However, the shortcoming of the granulation approach lies in that its scale parameter is not easy to determine, which leads to instability.

To deal with mixed attributes data, a hybrid feature selection scheme is constructed in this work. The rest of this paper is organized as follows. In Sect. 2, related works on feature selection for mixed attributes data are studied. In Sect. 3, a new correlation measure which is to be used in the filter model is defined based on mutual information, by solving the calculation of mutual information between mixed attributes. Section 4 gives a hybrid feature selection scheme: the features are first filtered with a filter model, and then the final feature subset is determined by a wrapper model. Section 5 describes the evaluation metrics and datasets used in our study. Experiments and results are presented in Sect. 6. Finally, conclusions and future works are given in Sect. 7.

2 Related works

In this section, related works on the search strategy and evaluating criteria in feature selection are discussed with respect to the processing of mixed attributes.

According to the search strategy, feature selection can be categorized into three categories: the global search, the random search and the heuristic search (Sun et al. 2004; Muni et al.

2006). Global search strategy can find the best feature subset corresponding to its evaluating criteria, with the number of features to be selected known in advance. However, the difficulty is that the number is hard to be determined in advance, and the computing complexity is too high with a searching space $O(2^N)$(where N is the number of original features) (Somol et al. 2004; Liu and Sun 2007). Random search strategies, such as genetic algorithm (Ooi and Tan 2003) and ant colony optimization (Ke et al. 2008), find an approximation of the optimum solution in the whole search space. But they bear high uncertainty, and their parameters take a great influence on the results. Heuristic search strategies include sequential forward selection (SFS) (Guan et al. 2004), sequential backward selection (SBS) (Abe 2005) and so on. They obtain high computing efficiency at the cost of global optimal solution. In Peng et al. (2005), a fast feature selection scheme is designed, where the minimal redundancy and maximal relevance (mRMR) criteria are used to improve the feature selection performance. In our work, the idea of mRMR criteria is adopted and improved to match with the characteristic of mixed attributes data.

According to the evaluating criteria, feature selection can be categorized into three categories: the filter, the wrapper and the hybrid approach. The filter model gets higher calculating speed by assessing features with statistical properties of the training data, without any learning model assumed between outputs and inputs of the data. In the framework of feature selection given by Yu and Liu (2004), the target is to maximize the correlation between selected features and the decision variable, and minimize the correlation between selected features. Therefore, how to measure the correlation between features is a crucial point in a filter model. It relies on various measures of the general characteristics of the training data, such as distance, information, dependency, and consistency (Liu and Motoda 1998). Among these measures, mutual information (Kwak and Choi 2002) is mostly used because it does not require to assume knowing the sample distribution, does not need to transform the data, and can measure the degree of uncertainty between features in a quantified form. However, the mutual information-based correlation measure can only be defined between continuous variables or discrete variables. Kwak and Choi (2002) proposed a method for calculating mutual information between mixed attributes, with an assumption that all samples having the same probability of occurrence. In our work, a new method for calculating mutual information between mixed attributes is proposed with this assumption removed. Then, a new correlation measure is defined based on mutual information.

The wrapper approach assesses feature subset with the training accuracy of the learning model, and usually yields high fitting accuracy at the cost of high computational complexity. In Hsu (2004), the genetic algorithm is used to find a feature subset with the smallest classifying error rate of decision tree. In Chiang and Pell (2004), the Fisher discriminant analysis is combined with the genetic algorithm to identify the pivotal variables in the failure process of chemical engineering. In Guyon et al. (2002), the importance of features is measured by the classification performance of support vector machine, based on which a classifier is constituted. In Michalak and Kwasnicka (2006), a wrapper model is constituted based on a two-pronged correlation strategy. In Monirul Kabir et al. (2010), the neural networks is used to present a wrapper feature selection algorithm. All the above models are focused on the classification problem, in which the target is to improve the classification accuracy. However, when the decision variable is continuous, the wrapper model is to be re-designed. To deal with mixed attributes, the case-based reasoning (CBR) approach is adopted in this paper. CBR is a well-established methodology with broad applications and is good at dealing with mixed attributes data. The fundamental principle of CBR is when provided a new project, the most similar historical projects are selected to estimate the new project using similarity measure.

The hybrid approach attempts to take advantage of the filter and wrapper approaches, while a single technique often traps into an immature solution. Therefore, a hybrid feature selection scheme is proposed in this paper.

3 The correlation measure based on mutual information

3.1 The calculation of mutual information between mixed attributes

3.1.1 Entropy and mutual information

In information theory, entropy is a measure to describe the uncertainty of a random variable. Let X be a discrete random variable with a range of Φ, and the probability distribution function is $p(x) = P\{X = x\}$, then the entropy of X is defined as:

$$H(X) = -\sum_{X\in\Phi} p(x)\log p(x) \tag{1}$$

For two discrete random variables X and Y (the range of Y being Ω), let their joint probability density function be $p(x, y)$, then the joint entropy of X and Y is defined as:

$$H(X, Y) = -\sum_{x\in\Phi}\sum_{y\in\Omega} p(x, y)\log p(x, y) \tag{2}$$

When X is known, the conditional entropy of Y is defined as:

$$H(Y|X) = \sum_{x\in\Phi} p(x)H(Y|x) = -\sum_{x\in\Phi}\sum_{y\in\Omega} p(x, y)\log p(y|x) \tag{3}$$

Therefore, the relationship between joint entropy and conditional entropy is:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{4}$$

Mutual information defines the shared information between two random variables:

$$I(X, Y) = \sum_{x\in\Phi}\sum_{y\in\Omega} p(x, y)\log\frac{p(x, y)}{p(x)p(y)} \tag{5}$$

in which $p(x) = \sum_{y\in\Omega} p(x, y)$, $p(y) = \sum_{x\in\Phi} p(x, y)$. The more information shared between X and Y, the larger is $I(X, Y)$. When $I(X, Y) = 0$, X and $Y$are independent.

By the above definition, the relationship between entropy, conditional entropy, joint entropy and mutual information is:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y) \tag{6}$$

The relationship between the above four is illustrated in Fig. 1.

When X and Y are continuous, the entropy, conditional entropy, joint entropy and mutual information are, respectively, defined as:

$$H(X) = -\int p(x)\log p(x)\mathrm{d}x \tag{7}$$

$$H(Y|X) = -\int\int p(x, y)\log p(y|x)\mathrm{d}x\mathrm{d}y \tag{8}$$

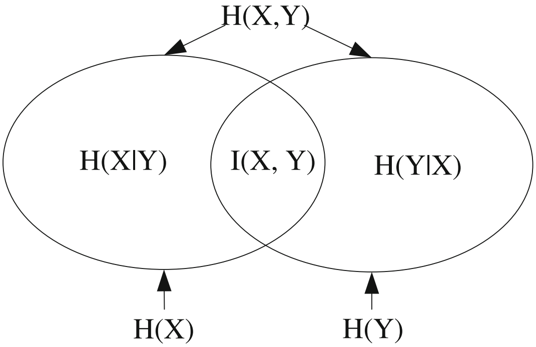


Fig. 1 Relationships between entropy, mutual information, joint entropy, and conditional entropy

$$H(X, Y) = -\int\int p(x, y)\log p(x, y)\mathrm{d}x\mathrm{d}y \tag{9}$$

$$I(X, Y) = \int\int p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \tag{10}$$

in which $p(x) = \int_y p(x, y)\mathrm{d}y$, $p(y) = \int_x p(x, y)\mathrm{d}x$

3.1.2 The calculation of mutual information between continuous attributes

For two discrete variables, the mutual information can be calculated with Eq. (5), after their joint distribution and marginal distribution are estimated. However, when X and Y are continuous, it is practically impossible to find exact integration in Eq. (10). Therefore, the approximation estimators are proposed. Existing methods include the histogram method, the kernel density method and neighbor method (Schaffernicht et al. 2010). Paper (Schaffernicht et al. 2010) came up with the Gaussian kernel density method by comparing the above methods on different datasets. Therefore, the mutual information is estimated with the kernel density method in this paper for continuous variables.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a dataset with n d-dimensional samples. The approximate of the density function has the following form:

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{x} - \mathbf{x}_i, h), \tag{11}$$

where $\delta(\cdot)$ is the Parzen window function, h is the window width. Parzen has proven that with proper chosen $\delta(\cdot)$ and h, the estimation $\hat{p}(\mathbf{x})$ can converge to the true density $p(\mathbf{x})$ when n tends to infinity. Usually, $\delta(\cdot)$is chosen as the Gaussian window:

$$\delta(z) = \frac{1}{(2\pi)^{d/2}h^d|\sum|^{1/2}}\exp\left(-\frac{z\sum^{-1}z}{2h^2}\right), \tag{12}$$

where $\mathbf{z} = \mathbf{x} - \mathbf{x}_i$, $\sum$ is the covariance of $\mathbf{z}$. The window width is practically set as $h = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-\frac{1}{d+4}}$.

The mutual information between X and Y can be estimated with Eqs. (10–12) and $d = 2$.

3.1.3 The calculation of mutual information between mixed attributes

Let X be a continuous variable with a range of Φ, and Y be a discrete variable with a range of $\{y_1, y_2, \ldots, y_m\}$, By Eq. (6), the mutual information is represented as follows:

$$I(X, Y) = H(X) - H(X|Y) = H(X) - \sum_{i=1}^{m} p(y_i) H(X|y_i) \tag{13}$$

In Eq. (13), we need to calculate $H(X)$ and $H(X|y_i)$. To calculate $H(X)$, the density function $p(x)$ of X is estimated using Eq. (11):

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i, h_Y), \tag{14}$$

where $h_Y = \left(\frac{4}{3}\right)^{1/(4)} n^{-\frac{1}{4}}$.

Replace the integration with a summation of the sample points, and the estimation of $H(X)$ is received as:

$$\hat{H}(X) = -\sum_{i=1}^{n} \hat{p}(x_i) \log \hat{p}(x_i) \tag{15}$$

To calculate $H(X|y_i)$, we have

$$H(X|y_i) = \int_x p(x|y_i) \log p(x|y_i) \mathrm{d}x \tag{16}$$

Let n_k be the number of examples with $Y = y_k$and I_k be the set of indices of the samples with $Y = y_k$, then the estimation of $p(x|y_i)$ is

$$\hat{p}(x|y_i) = \frac{1}{n_k} \sum_{i \in I_k} \delta(x - x_i, h_k), \tag{17}$$

where $h_k = (\frac{4}{3})^{1/(4)} n_k^{-\frac{1}{4}}$. Replace the integration with a summation of the sample points in Eq. (16), and the estimation of $H(X)$ is received as

$$\hat{H}(X|y_i) = -\sum_{i \in I_k} \hat{p}(x|y_i) \log \hat{p}(x|y_i) \tag{18}$$

Let the estimation of $p(y_i)$ be $\hat{p}(y_i) = n_i/n$, then the mutual information is:

$$\hat{I}(X, Y) = -\sum_{i=1}^{n} \hat{p}(x_i) \log \hat{p}(x_i) + \frac{1}{n} \sum_{i=1}^{m} \left[n_i \sum_{i \in I_k} \hat{p}(x|y_i) \log \hat{p}(x|y_i) \right] \tag{19}$$

3.2 The correlation measure

For two random variables X and Y, the correlation measure is defined as the mutual information in paper (Ooi and Tan 2003), based on which a maximum correlation minimum redundancy algorithm is given. However, experiments show that this correlation measure tends to choose features with more values. Therefore, by normalizing the correlation measure into

[0, 1], a new correlation measure $C(X, Y)$ is defined as:

$$C(X, Y) = \frac{1}{2}\left[\frac{I(X, Y)}{H(X)} + \frac{I(X, Y)}{H(Y)}\right] \tag{20}$$

Obviously, the above definition meets the symmetry, with the range of $C(X, Y)$ being [0, 1] where $C(X, Y) = 1$ means knowing any of X or Y, the other is determined, and $C(X, Y) = 0$ means X and Y are independent from each other.

4 A hybrid feature selection scheme for mixed attributes data

In this section, a hybrid feature selection scheme is proposed taking advantages of both the filter and the wrapper model. In this scheme, N features are firstly filtered, and then the number N is optimized by minimizing the estimation accuracy of the CBR.

4.1 Filter feature selection for mixed attributes data

In filter feature selection based on information criteria, the primary issue is to find a feature subset as more correlative as possible with the decision variable, and meanwhile the correlation between features in the subset is as small as possible. However, in high dimensional space, to estimate the probability density is difficult and slow. Therefore, the algorithm of mRMR algorithm gave an evaluation criteria based on mutual information (Liu and Sun 2007) to select N features from the original feature set. The algorithm is as follows:

(1) (Initialization) Set $F \leftarrow'$ whole feature set$'$, $S \leftarrow'$ empty set$'$, $y \leftarrow'$ decision variable’.
(2) $\forall f_i \in F$, compute $I(f_i, y)$.
(3) Find the feature f_i that maximizes $I(f_i, y)$, set $F \leftarrow F\backslash\{f_i\}$, $S \leftarrow \{f_i\}$.
(4) Repeat until desired number N of features is selected.

   (a) $\forall f_i \in F$, $f_s \in S$, compute $I(f_i, f_s)$, if it is not yet available.
   (b) Choose the feature $f_i \in F$ that maximizes$J(f_i) = I(f_i, y) - \frac{1}{|S|}\sum_{s\in S} I(f_i, f_s)$; set $F \leftarrow F\backslash\{f_i\}$, $S \leftarrow S \cup \{f_i\}$.

(5) Output the subset S containing N selected features.

However, mRMR based on $J(f_i)$ tends to select features with more values. Therefore, we replace $J(f_i) = I(f_i, y) - \frac{1}{|S|}\sum_{s\in S} I(f_i, f_s)$ in (b) as $J(f_i) = I(f_i, y) - \frac{1}{|S|}\sum_{s\in S} C(f_i, f_s)$, and denote the new algorithm as C-mRMR.

4.2 Determination of the filter’s parameter based on CBR

In the C-mRMR algorithm, the parameter N is to be determined. In this study, it is determined by optimizing the estimation accuracy of the CBR. The CBR estimates the target case by similar historical cases, and usually consists of three sub-problems (Li et al. 2009): similarity measure, number of analogies and analogy adaptation.

4.2.1 Similarity measure

Similarity measure describes the level of similarity between different samples. Several similarity functions have been proposed, however, the measures used in this study are the Euclidean distance, and the Manhattan distance, since they have been reported with good results in software cost estimation studies (Chiu and Huang 2007).

The Euclidean distance measures the Euclidean distance $d(p, p')$ between two samples after the continuous features have been normalized:

$$d(p, p') = \sqrt{\sum_{i=1}^{d} w_i \mathrm{Dis}(f_i, f_i')} \tag{21}$$

$$\mathrm{Dis}\left(f_i, f_i'\right) = \begin{cases} \left(f_i - f'\right)^2, & f_i \text{ and } f' \text{ are numeric or ordinal} \\ 1, & f_i \text{ and } f' \text{ are nominal and } f_i = f' \\ 0, & f_i \text{ and } f' \text{ are nominal and } f_i \neq f' \end{cases} \tag{22}$$

The Manhattan distance is the sum of the absolute distances for each pair of features:

$$d(p, p') = \sum_{i=1}^{d} w_i \mathrm{Dis}(f_i, f_i') \tag{23}$$

$$\mathrm{Dis}\left(f_i, f_i'\right) = \begin{cases} \left|f_i - f'\right|, & f_i \text{and } f' \text{ are numeric or ordinal} \\ 1, & f_i \text{ and } f' \text{ are nominal and } f_i = f' \;, \\ 0, & f_i \text{ and} f' \text{ are nominal and } f_i \neq f' \end{cases} \tag{24}$$

where p_1 and p' denote the samples, f_i and f_i' denote the $i$thfeature value of p_1 and p', $w_i = \{0, 1\}$ is the weight of the $i$thfeature, where $w_i = 1$ means the $i$thfeature is selected and $w_i = 0$ means the $i$thfeature is not selected, d is the total number of features.

4.2.2 Number of analogies

The number of analogies refers to the number of most similar samples that will be used to generate the estimation. $K = 1$ means the closest analogy. However, in this study $K = \{1, 2, 3, 4, 5\}$ are considered since it could cover most of the suggested numbers (Jørgensen et al. 2003).

4.2.3 Analogy adaptation

After the analogies are selected, the final estimation for the new sample is determined by computing certain statistic based on the selected samples. The adaptation techniques used in this study are the closet analogy, the mean of closet analogies and the inverse distance weighted mean.

The mean is the average of the costs of $K(K > 1)$ analogies. It is a classical measure of central tendency and treats all analogies as being equally influential on the cost estimates. The median is the median of the costs of $K(K > 1)$ analogies. It is another measure of central tendency and a more robust statistic when the number of analogies increases.

The inverse distance weighted mean allows more similar analogies to have more influence than less similar ones. The formula for weighed mean is shown in (25):

$$\omega_k = \frac{1/(\delta + d(p, p_k))}{\sum_{i=1}^{K} 1/(\delta + d(p, p_i))}, \tag{25}$$

where K is the number of analogies, p_k represents the kth closet analogy with the new sample p, $d(p, p_k)$ is the distance measure between p_k and p, δ is a small constant and in our study δ is set to 0.001.

5 Evaluation criteria and data sets

5.1 Evaluation criteria

To evaluate the performance of the method in this study, it is compared with existing methods on the feature selection results and the estimation accuracies. On the estimation accuracy, three evaluation criteria are used out of the majorities of existing studies, which are the mean magnitude of relative error (MMRE), the median magnitude of relative error (MdMRE) and the PRED(0.25). The MMRE is defined as below:

$$\text{MMRE} = \frac{1}{n} \times \sum_{i=1}^{n} \text{MRE}_i \tag{26}$$

$$MRE_i = \frac{\left|E_i - \hat{E}_i\right|}{E_i}, \tag{27}$$

where n denotes the number of samples, E_i denotes the actual effort of the ith sample, $\hat{E}_i$ denotes the estimated effort of the ith sample. Small MMRE value indicates the low level of estimation error. However, this metric is unbalanced and penalizes overestimation more than underestimation.

The MdMRE is the median of all MREs:

$$\text{MdMRE} = \text{median (MRE)} \tag{28}$$

MdMRE is an aggregate measure which is less sensitive extreme values. It exhibits a similar pattern to MMRE, but is more likely to select the true model especially in the underestimation cases.

The PRED(0.25) is the percentage of estimations that fall within 25% of actual value:

$$\text{PRED}(0.25) = \frac{1}{n} \times \sum_{i=1}^{n} I\left\{MRE_i \leq 0.25\right\} \tag{29}$$

5.2 Data sets description

To conveniently compare with other methods, two representative datasets (the Desharnais dataset and the Maxwell dataset) are used for experiments, which have been used by many recent research works, such as Li et al. (2009), Mair et al. (2000), Maxwell (2002), Sentas et al. (2005).

The Desharnais dataset contains two discrete variables (‘YearEnd’ and ‘Language’) and nine continuous variables (the rest variables), while the discrete variables can be further classified into one nominal variable and one ordinal variable. The decision variable ‘Effort’ is continuous. This dataset contains a total of 81 samples, and 4 out of 81 samples are excluded due to the missing of feature values. A more detailed description of all features is shown in Table 1.

The Maxwell dataset with 62 samples from one of the biggest commercial banks in Finland is a relative new software projects dataset. The features are described in Table 2. There are 3 continuous variables (‘Duration’, ‘Size’ and ‘Effort’) and 24 discrete variables (the rest variables), while the discrete variables can be further classified into 5 nominal variables and 19 ordinal variables. The decision variable ‘Effort’ is continuous. The variable ‘Time’ is

Table 1 Feature definition in Desharnais dataset

ID	Features	Full name	Description
1	TeamExp	Team experience	Numerical in years
2	ManagerExp	Manager’s experience	Numerical in years
3	YearEnd	Year of end	Nominal in year
4	Length	Actual project schedule	Numerical in months
5	Language	programming languages	Ordinal: {1, 2, 3}
6	Transactions	Transactions	Numerical in function points
7	Entities	Entities	Numerical in function points
8	PointsNonAdjust	Unadjusted function points	Numerical in function points
9	Adjustment	Adjustment factor	Numerical
10	PointsAjust	Adjusted function points	Numerical in function points
11	Effort	Development effort	Numerical in person-hours

Table 2 Feature definition in Maxwell dataset

ID	Features	Full name	Description
1	Syear	Starting year	Ordinal in year
2	App	Application type	Nominal in {1, 2, 3, 4, 5}
3	Har	Hardware platform	Nominal in {1, 2, 3, 4, 5}
4	Dba	Database	Nominal in {1, 2, 3, 4,}
5	Ifc	User interface	Nominal in {1, 2}
6	Source	Where developed	Nominal in {1, 2}
7	Telonuse	Telon use	Nominal in {0, 1}
8	Nlan	Number of languages	Ordinal in {1, 2, 3, 4,}
9	T01	Customer participation	Ordinal:
10	T02	Development environment adequacy	1 = Very low
11	T03	Staff availability	2 = Low
12	T04	Standards use	3 = Nominal
13	T05	Methods use	4 = High
14	T06	Tools use	5 = Very high
15	T07	Software’s logical complexity	
16	T08	Requirements volatility	
17	T09	Quality requirements	
18	T10	Efficiency requirements	
19	T11	Installation requirements	
20	T12	Staff analysis skills	
21	T13	Staff application knowledge	
22	T14	Staff tool skills	
23	T15	Staff team skills	
24	Duration	Duration	Numerical in months
25	Size	Application size	Numerical in function points
26	Time	Time	Ordinal, Time = Syear-1985 +1
27	Effort	Effort	Numerical in hours

eliminated due to its same meaning as the variable ‘Syear’. As the dataset only contains one sample with the variables ‘subapp’ and ‘subhar’ at the value ‘4’, respectively, by following (Maxwell 2002), two new variables ‘subapp’ and ‘subhar’ are used instead of the variables ‘app’ and ‘har’, while new variables are subsets of the original ones. That is, subapp = {1,2,3,5}, subhar = {1,2,3,5}.

6 Experiments

To validate the proposed HFS, the feature selection results and the estimation accuracy of HFS are compared with published works based on the above two datasets.

6.1 The results on Desharnais dataset

6.1.1 Analysis of feature selection

To conduct the feature selection, the parameters (similarity measure, number of analogies and analogy adaptation) of HFS are firstly to be determined. Consulting Mair et al. (2000), 87% samples are selected as the training set, and the rest 13% as the testing set. Table 3 summarizes the results with considerations of different parameter configurations: two distance measures (Euclidean distance and Manhattan distance), five K values (1, 2, 3, 4 and 5), and four adaptation techniques [closest analogy (CA), mean, inverse distance weighted mean (IWM), and median].

The results show that, in general, the choice of different distance measures has an insignificant influence on the estimation accuracy. As to the adaptation, the ‘Median’ is steadier than the others, and gets slightly better results than ‘Mean’ and ‘IWM’ when $K = 4$ and $K = 5$. The choice of K values has some influence on the accuracies. The best configuration on the training set is ‘the Manhattan distance’, ‘$K = 4$’ and ‘Median’.

Figure 1 shows a histogram for the mutual information values of each feature f and the decision variable y. When using ‘the Manhattan distance’, ‘$K = 4$’ and ‘Median’, the selected features in turn are ‘PointsAjust’, ‘Entities’, ‘Transactions’, ‘PointsNonAdjust’ and ‘Language’, which are the 10th, 7th, 6th, 8th and 5th variable in Fig. 2 (Table 1). From the meaning of variables, knowing any two of ‘PointsAjust’, ‘PointsNonAdjust’ and ‘Adjustment’, the third can be determined, therefore C-mRMR algorithm choose ‘PointsAjust’ and ‘PointsNonAdjust’ for their greater mutual information. The duration of the project is strongly correlative with ‘PointsAjust’ and ‘PointsNonAdjust’, so C-mRMR algorithm did not choose ‘Length’. ‘Transactions’ and ‘Entities’ are important informations on the software size, so it is retained by C-mRMR algorithm. In terms of the same development unit, team and manager experience are relatively fixed, so C-mRMR algorithm did not select ‘TeamExp’ and ‘ManagerExp’. Though the mutual information of ‘language’ with the decision variable is relatively small, it is relatively independent of other variables, therefore, it is retained.

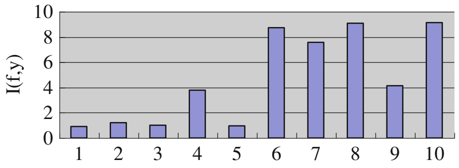
Next, HFS (using ‘the Manhattan distance’, ‘$K = 4$’ and ‘Median’) is compared with NMI in Hua et al. (2011), mRMR in Peng et al. (2005) and MICBR in Li et al. (2009). In Li et al. (2009), the three-folder cross-validation is used to test the performance of candidate methods, in which the Desharnais dataset is randomly divided into three different training splits and three testing splits. By following it, the 87% split (87% in the training set and 13% in the validating set) is used in this paper.

Results are shown in Table 4, where the scale parameter of NMI is taken as 0.15 as suggested in Hua et al. (2011), the number of selected features of mRMR in Peng et al. (2005)

Table 3 Results of different parameters on Desharnais dataset

Distance	*K* value	Adaptation	Training			Testing		
			MMRE	PRED (0.25)	MdMMRE	MMRE	PRED (0.25)	MdMMRE
Euclidean	$K=1$	CA	0.50	0.31	0.44	0.55	0.20	0.46
	$K=2$	Mean	0.46	0.33	0.42	0.48	0.30	0.43
		IWM	0.45	0.32	0.42	0.55	0.30	0.44
	$K=3$	Mean	0.46	0.35	0.39	0.43	0.30	0.38
		IWM	0.45	0.36	0.41	0.44	0.30	0.40
		Median	0.44	0.38	0.39	0.48	0.30	0.41
	$K=4$	Mean	0.43	0.31	0.43	0.40	0.30	0.42
		IWM	0.44	0.30	0.44	0.41	0.30	0.40
		Median	0.41	0.36	0.35	0.35	0.40	0.34
	$K=5$	Mean	0.37	0.33	0.45	0.37	0.30	0.42
		IWM	0.40	0.29	0.42	0.39	0.40	0.39
		Median	0.38	0.31	0.40	0.36	0.40	0.37
Manhattan	$K=1$	CA	0.51	0.28	0.43	0.55	0.10	0.44
	$K=2$	Mean	0.47	0.32	0.41	0.52	0.20	0.41
		IWM	0.48	0.31	0.42	0.47	0.30	0.40
	$K=3$	Mean	0.47	0.34	0.43	0.48	0.30	0.39
		IWM	0.45	0.32	0.41	0.45	0.30	0.38
		Median	0.45	0.34	0.38	0.43	0.30	0.37
	$K=4$	Mean	0.42	0.34	0.37	0.40	0.40	0.34
		IWM	0.41	0.34	0.36	0.38	0.40	0.31
		Median	0.37	0.37	0.34	0.34	0.41	0.31
	$K=5$	Mean	0.40	0.31	0.38	0.37	0.30	0.36
		IWM	0.39	0.32	0.41	0.35	0.30	0.40
		Median	0.39	0.34	0.39	0.38	0.34	0.37

Fig. 2 MI between features and decision variable



is determined by optimizing the estimation accuracy of CBR, and the result of MICBR is extracted from Li et al. (2009). The symbol ‘1’ denotes that the feature in its corresponding row is selected by the feature selection method in its corresponding column.

It can be seen from Table 4 that, though the neighborhood mutual information is able to handle mixed attributes, its scale parameter is not easy to be determined, which leads to unstable results. The mRMR algorithm is more stable, but its direct use of mutual information values to measure the correlation leads to the selected features being all continuous, and all three variables ‘PointsAjust’, ‘PointsNonAdjust’ and ‘Adjustment’ being selected indicate that it is not good at eliminating redundancy for mixed attributes data. The MICBR in Li et al. (2009) excluded the variable ‘Language’, so its result is less

Table 4 Selected features in three data subsets

Datasets:	Training set 1				Training set 2				Training set 3			
Variables	NMI	mRMR	MICBR	HFS	NMI	mRMR	MICBR	HFS	NMI	mRMR	MICBR	HFS
TeamExp												
ManagerExp			1								1	
YearEnd												
Length	1								1			
Language	1			1	1			1	1			1
Transactions	1	1	1	1	1	1		1	1	1	1	1
Entities		1	1	1	1	1	1	1		1	1	1
PointsNonAdjust		1	1	1	1	1		1	1	1	1	1
Adjustment	1	1	1		1	1	1		1	1	1	
PointsAjust	1	1	1	1	1	1	1	1	1	1	1	1

Table 5 Comparison of estimation accuracy on the Desharnais dataset

	Training set			Testing set		
	MMRE	PRED (0.25)	MdMMRE	MMRE	PRED (0.25)	MdMMRE
NMI	0.39	0.35	0.40	0.38	0.39	0.37
mRMR	0.41	0.33	0.40	0.40	0.36	0.38
MICBR	0.68	0.32	0.39	0.36	0.40	0.33
HFS	0.37	0.37	0.34	0.34	0.41	0.31

interpretable. In comparison, the proposed HFS in this paper is more stable, better able to remove redundancy and stronger interpretability.

6.1.2 Analysis of estimation accuracy

In Table 5, the estimation accuracy of HFS is compared with the above three methods with 87% samples being selected as training set and the rest 13% as testing set. The result of HFS is extracted from Table 3, and the result of MICBR is extracted from Table 6 in Li et al. (2009) with ‘the Manhattan distance’, ‘$K = 4$’ and ‘Median’.

It can be seen from Table 5 that the HFS obtains the best results on MMRE, PRED(0.25) and MdMMRE than the other three methods.

6.2 The results on Maxwell dataset

6.2.1 Analysis of feature selection

Consulting Maxwell (2002) and Sentas et al. (2005), the 50 projects finished before 1992 are used as training set, and the 12 projects finished from 1992 to 1993 are used as testing set. Feature selection is conducted on the training set with the same configuration used in Desharnais dataset: ‘the Manhattan distance’, ‘$K = 4$’ and ‘Median’.

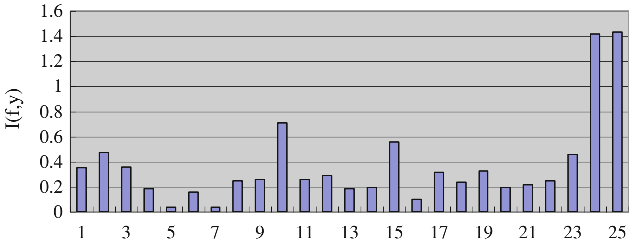


Fig. 3 MI between features and decision variable

Figure 3 shows a histogram for the mutual information values of each feature f and the decision variable y. With HFS, the selected features in turn are 'Size', 'Dba', 'T12', 'Source' and 'T15' and 'T02', which are the 25th, 4th, 20th, 6th, 23rd and 10th variable in Fig. 3 (Table 2).

The result of HFS is compared with NMI in 27, mRMR in 35 and MICBR in 46, where the scale parameter of NMI is taken as 0.1, 0.15 and 0.2, the number of selected features of mRMR is determined by optimizing the estimation accuracy of CBR, and the result of MICBR is extracted from Li et al. (2009). The symbol '1' denotes that the feature in its corresponding row is selected by the feature selection method in its corresponding column.

It can be seen from Table 6 that the scale parameter of NMI in Hua et al. (2011) is not easy to be determined, which leads to unstable results. The mRMR algorithm in Peng et al. (2005) is more stable, but it selects both 'Duration' and 'Size', while the two variables are strongly correlative with each other. This again indicates that the mRMR is not good at eliminating redundancy for mixed attributes data. In Li et al. (2009), 'Time' and 'Duration' are treated as numerical variables. This leads to the mutual information value between 'Time' and 'Effort' being different from the mutual information value between 'Duration' and 'Effort', while 'Time' and 'Duration' have exactly the same meanings. This indicates that the results of Li et al. (2009) are less interpretable. In comparison, the proposed HFS in this paper is more stable, better able to remove redundancy and stronger interpretability.

6.2.2 Analysis of estimation accuracy

In Table 7, the estimation of HFS is compared with the above three methods with the 50 projects finished before 1992 being selected as training set and the 12 projects finished from 1992 to 1993 as testing set. And the result of MICBR is extracted from Table 11 in Li et al. (2009) with 'the Euclidean distance', '$K = 4$' and 'Mean'.

It can be seen from Table 5 that the HFS obtains the best MMRE and PRED(0.25) on training set and the best MMRE and MdMMRE on testing set than the other three methods.

7 Conclusions

Feature selection plays an important role in pattern recognition and machine learning. Traditional feature selection methods are mainly designed for the handling classification problems with discrete or continuous features. However, in many practical problems (such as software cost estimation problem), the collected data often have mixed attributes, with the decision variable being continuous. To deal with these problems, a hybrid feature selection scheme for mixed attributes data is proposed which takes advantages of both the filters and the wrappers.

Table 6 Selected features for training set of Maxwell dataset

Variables	NMI			mRMR	MICBR	HFS
	0.1	0.15	0.2			
Syear						
App	1			1		
Har			1			
Dba	1	1	1			1
Ifc						
Source		1				1
Telonuse						
Nlan		1				
T01	1	1				
T02	1		1	1		1
T03						
T04			1			
T05						
T06						
T07				1		
T08						
T09						
T10						
T11						
T12						1
T13						
T14					1	
T15	1	1	1	1		1
Duration			1	1	1	
Size	1	1		1	1	1

Table 7 Comparison of estimation accuracy on the Maxwell dataset

	Training set			Testing set		
	MMRE	PRED (0.25)	MdMMRE	MMRE	PRED (0.25)	MdMMRE
NMI	0.46	0.33	0.36	0.29	0.44	0.28
mRMR	0.49	0.34	0.37	0.30	0.41	0.31
MICBR	0.51	0.48	0.29	0.28	0.67	0.19
HFS	0.44	0.49	0.33	0.26	0.66	0.18

To do feature selection, a proper correlation measure for features is essential. In this paper, we first give a method for calculating mutual information between discrete and continuous variables. Then, we use the mutual information to define a new correlation measure suitable for mixed attributes data. With this correlation measure, features are filtered with a undefined parameter N. Finally, a CBR-based wrapper model is proposed to determine the parameter

N. Examples show that this method is applicable for feature selection of the mixed attributes data, being more stable, interpretable, and with better estimation accuracy.

However, only the Desharnais dataset is used for experiments in the study, and the future work could include the application on other datasets such as the ISBSG database.

References

Abe S (2005) Modified backward feature selection by cross validation. In: Proceedings of the European symposium on artificial neural networks, pp 163–168

Amiri F, Yousefi MR, Lucas C, Shakery A, Yazdani N (2011) Mutual information-based feature selection for intrusion detection systems. J Netw Comput Appl 34:1184–1199

Cakır S, Aytac T, Yilm A(2011) Classifier-based offline feature selection and evaluation for visual tracking of sea-surface and aerial targets. Opt Eng 50(10):1–13

Chiang LH, Pell RJ (2004) Genetic algorithms combined with discriminant analysis for key variable identification. J Process Control 14(2):143–155

Chiu NH, Huang SJ (2007) The adjusted analogy-based software effort estimation based on similarity distances. J Syst Softw 80:628–640

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156

Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151:155–176

Ferreira A, Figueiredo M (2011) Unsupervised joint feature discretization and selection. Lect Notes Comput Sci 6669:200–207

Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555

Guan S, Liu J, Qi Y (2004) An incremental approach to contribution-based feature selection. J Intell Syst 13(1)

Guyon I, Weston J, Barnhill S et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hsu C, Huang H, Schuschel D (2002) The ANNIGMA-wrapper approach to fast feature selection for neural nets. IEEE Trans Syst Man Cybern Part B Cybern 32(2):207–212

Hsu WH (2004) Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning. Inf Sci 163(17):103–122

Hsu HH, Hsieh CW, Lu MD (2011) Hybrid feature selection by combining filters and wrappers. Expert Syst Appl 38:8144–8150

Hu QH, Zhao H, YU DR (2008b) Efficient symbolic and numerical attribute reduction with neighborhood rough sets. PR & AI 21(6):732–738 (In Chinese)

Hu QH, Yu DR, Xie ZX (2006) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn Lett 27(5):414–423

Hu QH, Liu JF, Yu DR (2008a) Mixed feature selection based on granulation and approximation. Knowl Based Syst 21:294–304

Hu QH, Che XJ, Zhang L, Yu DR (2010) Feature evaluation and selection based on neighborhood soft margin. Neurocomputing 73:2114–2124

Hua QH, Zhang L, Zhang D, Pan W, An S, Pedrycz W (2011) Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Syst Appl 38:10737–10750

Jørgensen M, Indahl U, Sjøberg D (2003) Software effort estimation by analogy and regression toward the mean. J Syst Softw 68:253–262

Ke L, Feng Z, Ren Z (2008) An efficient ant colony optimization approach to attribute reduction in rough set theory. Pattern Recogn Lett 29:1351–1357

Kwak N, Choi CH (2002) Feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159

Kwak N, Choi CH (2002) Input feature selection by mutual information based on Parzen window. IEEE Trans Pattern Anal Mach Intell 24(12):1667–1671

Li YF, Xie M, Goh TN (2009) A study of mutual information based feature selection for case based reasoning in software cost estimation. Expert Syst Appl 36:5921–5931

Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Kluwer Academic Publishers, Boston

Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502

Liu JK, Sun FC (2007) A novel dynamic terminal sliding mode control of uncertain nonlinear systems. J Control Theory Appl 5(2):189–193

Mair C, Kadoda G, Lefley M, Phalp K, Schofield C (2000) An investigation of machine learning based prediction systems. J Syst Softw 53:23–29

Mao Y, Zhou XB, Xia Z, Yin Z, Sun YX (2007) A Survey of study of feature selection algorithms. PR & AI 20(2):211–218 (In Chinese)

Maxwell K (2002) Applied statistics for software managers. Prentice-Hall, Englewood Cliffs

Michalak K, Kwasnicka H (2006) Correlation-based feature selection strategy in classification problems. Int J Appl Math Comput Sci 16(4):503–511

Monirul Kabir Md., Monirul Islam Md., Murase K (2010) A new wrapper feature selection approach using neural network. Neurocomputing 73:3273–3283

Monirul Kabir Md., Shahjahan Md., Murase K (2011) A new local search based hybrid genetic algorithm for feature selection. Neurocomputing 74:2914–2928

Muni DP, Pal NR, Das J (2006) Genetic programming for simultaneous feature selection and classifier design. IEEE Trans Syst Man Cybern Part B Cybern 36(1):106–117

Oh IS, Lee JS, Moon BR (2004) Hybrid genetic algorithms for feature selection. IEEE Trans Pattern Anal Mach Intell 26(11):1424–1437

Ooi CH, Tan P (2003) Genetic algorithm applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19(1):37–44

Peng HC, Long FH, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Schaffernicht E, Kaltenhaeuser R, Verma SS, Gross HM (2010) On estimating mutual information for feature selection. Lect Notes Comput Sci 6325:362–367

Sentas P, Angelis L, Stamelos I, Bleris G (2005) Software productivity and effort prediction with ordinal regression. Inf Softw Technol 47:17–29

Sheng LD (2000) Introduction to pattern recognition. Beijing University of Posts and Telecommunications Press, Beijing (In Chinese)

Somol P, Pudil P, Kittler J (2004) Fast branch & bound algorithms for optimal feature selection. IEEE Trans Pattern Anal Mach Intell 26(7):900–912

Sun ZH, Bebis G, Miller R (2004) Object detection using feature subset selection. Pattern Recogn 37(11):2165–2176

Sun Y (2007) Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell 29(6):1035–1051

Verikas A, Bacauskiene M (2002) Feature selection with neural networks. Pattern Recogn Lett 23:1323–1335

Wang L, Zhou N, Chu F (2008) A general wrapper approach to selection of class-dependent features. IEEE Trans Neural Netw 19(7):1267–1278

Yang Y, Liao YX, Meng G, Lee J (2011) A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. Expert Syst Appl 38:11311–11320

Yao X, Wang XD, Zhang YX, Quan W (2012)Summary of feature selection algorithms. Control Decision 27(2):161–166 (In Chinese)

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5(1):1205–1224

Zhu Z, Ong YS, Dash M (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recogn 49(11):3236–3248