Journal of the Society of Korea Industrial and Systems Engineering
Vol. 33, No. 1, pp.114－120, March 2010.

다중 형태 데이터를 위한 요소선택 방법

양재경 · 이태한†

전북대학교 산업정보시스템공학과

Feature Selection for Mixed Type of Data

Jaekyung Yang · Taehan Lee†

Dept. of Industrial and Information Systems Eng, Chonbuk National University

데이터마이닝의 사전 단계에서 데이터의 차원(Dimensionality)을 줄이기 위한 단계로서 많은 요소선택(Feature Selection) 방법들이 개발되었다. 이 방법은 결과를 예측하거나 데이터를 설명하고자 할 때 어떤 요소들이 관련이 있는지를 결정하는 과정을 포함한다. 또한 이 방법은 데이터의 크기에 대한 확장성(Scalability)를 향상시키며 학습 모델을 더욱 이해하기 쉽도록 줄 수 있다. 이 논문에서는 NP(Nested Partition) 방법을 사용한 최적화 기반의 새로운 요소선택 방법을 NP 구조의 기본적인 이론 근거와 함께 제안한다. 또 한 편으로 많은 요소선택 방법들이 다중 형태의 데이터를 처리하는데 한계를 가지고 있는데, NP 기반의 요소선택 방법에 다중 형태의 데이터를 처리할 수 있도록 하는 요소 성능 평가도구(Evaluators)를 도입하여 이를 극복하고자 한다. 또한 어떤 평가도구가 특정 데이터 형태에서 더욱 좋은 결과를 보이는지를 실험 결과와 함께 제시하였다.

Keywords : Data Mining, Feature Selection, Nested Partition, Mixed Type of Data

1. Introduction

Feature selection is an important data mining problem for numerous reasons [9]. It can be used to eliminate redundant and irrelevant features from a data set, resulting in a dimensionality reduction that reduces the learning time needed for induction algorithms that are applied to the data set, and in many cases also results in better (that is, more accurate) predictive models. Careful feature selection can improve the scalability of a data mining system as the induction is usually much faster with fewer features. The feature selection problem involves selecting a best subset of features from a finite subset and is therefore a discrete optimization problem. As such, any number of well known optimization approaches can be applied to this problem and previous work has for example used mathematical programming [2], branch-and-bound [10], and evolutionary search[6]. These feature selection algorithms have been used in various applications. Clark et al. [3] used a feature selection algorithm for land mine detection. Evgeniou et al. [4] employed feature selection in multimedia database search.

These and other feature selection methods can be typically classified as either filtering methods that produce a ranking of all features before the learning algorithm is applied or wrapper methods that use the learning algorithm to evaluate subsets of features. As a general rule, filtering methods are faster whereas wrapper methods usually produce subsets that results in more accurate models. We note that wrapper methods always fall into the latter category. In this paper, we propose the new feature selection methodology, which applies an optimization method called the nested partitions (NP) method [11] and its capability

논문접수일 : 2009년 12월 28일　　논문수정일 : 2009년 12월 29일　　게재확정일 : 2010년 01월 27일
† 교신저자 myuth0789@jbnu.ac.kr

for dealing with the mixed type of data. The new NP based feature selection method can be implemented as both filter and wrapper. However, a filter version of the NP based feature selection algorithm, namely NP-Filter is presented in this paper.

This feature selection algorithm uses the information gain to determine a partitioning order of features by evaluating quality of a feature. However, if we have a data set containing continuous valued features, we have to apply an appropriate discretization method to be able to make it nominal and then apply the information gain. Even though many discretization methods have been introduced, it is difficult to know where boundaries should be drawn. Thus in order to overcome this limitation and show that the new feature selection method is very well adapted to the mixed type of data, we assume that we do not use any discretization method first for continuous valued features, which means we use continuous values directly to evaluate the quality of features. We apply two different methods in the NP-Filter which are correlation based subset evaluator [5] and ReliefF feature evaluator [8] in order to deal with a mixed type of features, nominal and numerical data type. These methods will be evaluated and compared with information gain feature evaluator enabled by discretization using the NP-Filter for the selected data sets. Further, we investigate whether partitioning orders by the different feature quality evaluators may affect performances in the NP-Filter with Naïve Bayes, C4.5 and k-nearest neighbor learners.

The remainder of the paper is organized as follows. In section 2 we introduce the new methodology. Section 3 addresses how the NP-Filter deals with the mixed type of data. Section 4 shows our numerical results in applying the method to the various data domain on well known classification problems. Finally, section 5 contains some concluding remarks.

2. The models

The following notation will be used:

T : Training data (instances).

m : Number of instances ($m = |T|$).

$A^{(ALL)}$: Set of all features.

n : Number of features ($n = |A^{(ALL)}|$).

a : A specific feature ($a \in A^{(ALL)}$).

f : Performance measure.

f^* : Optimal Performance.

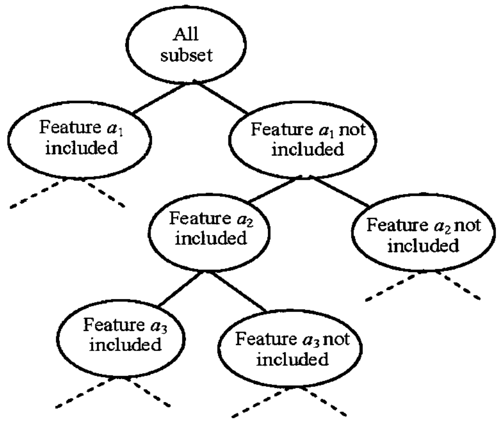
The main component in formulating the feature selection problem is selecting a performance measure. Depending on how this is done, feature selection methods may be divided into two categories : wrappers and filters. Wrapper methods use the accuracy of the resulting predictive model. This is an expensive evaluation and only applies for supervised learning. Filtering methods, on the other hand, select features before any other learning algorithm is applied. Thus, a different performance measure must be specified. The NP-framework can be implemented as either a wrapper or filter, resulting in the NP-Wrapper and NP-Filter algorithm, respectively [11]. In this paper, we focus a filter employing the following correlation based measure [5]:

$$f_{correlation}(A) = \frac{k\overline{\rho_{ca}}}{\sqrt{k + k(k-1)\overline{\rho_{aa}}}} \tag{1}$$

where k is the number of features in the set A, $\overline{\rho_{ca}}$ is the average correlation between the features in this set and the classification feature, and $\overline{\rho_{aa}}$ is the average correlation between features in the set A.

The nested partitions method is a general optimization methodology that can be applied to any combinatorial optimization problems [11]. The main idea of the method is to use iterative partitioning of the feasible region, that then creates a partitioning tree as shown in <Figure 1>. Thus, in each iteration a subset of the feasible region is determined to be the most promising or most likely to contain the global optimum. This subset is then partitioned into further subsets and what remains of the feasible region aggregated into one subset called the surrounding regions. Each of these subsets is randomly sampled and based on those samples a new subset is selected. If the surrounding region is selected, the algorithm backtracks, that is, simply moves to what was previously considered the most promising region. This approach can be effectively applied to feature selection as originally described by [11], either as a filter or a wrapper, depending on how feature subsets are evaluated.

The key to the convergence of the NP method is the probability by which a region is selected correctly in each iteration. A sufficient condition for asymptotic convergence is that this probability of correct selection is bigger than one half, and to guarantee that a minimum probability is obtained, and we can use a two-stage sampling procedure that determines how much random sampling effort, $N(\psi, \delta)$[11], is needed from each region to guarantee correct selection with probability ψ within an indifference zone $\delta > 0$. The two-stage sampling also allows



<Figure 1> Partitioning Tree

us to further analyze the convergence of the algorithm and develop statements concerning the quality of the solution once maximum depth is reached. In particular, an expression can be derived for the probability of having found sufficiently good solution the first time maximum depth is reached :

$$\Pr\{|f(A(k))-f^*|\}\leq \delta \geq \Psi \tag{2}$$

where $\delta > 0$ is an indifference zone, that is a performance value difference that is considered insignificant, and

$$\Psi=\frac{\psi^n}{(1-\psi)^n+\psi^n} \tag{3}$$

where ψ is the user selected minimum probability by which a correct selection is made in each iteration, and n is as before the total number of features. Sometimes it may be beneficial to stop the algorithm early, that is, we can specify a stopping depth $d_{stop}(n) \leq n$, define the objective function on sets of feature subsets as

$$f(A(k))=\max_{a\in A(k)} f(a) \tag{4}$$

and equation (3) holds with Ψ replaced with

$$\Psi'=\frac{\psi^{d_{stop}(n)}}{(1-\psi)^{d_{stop}(n)}+\psi^{d_{stop}(n)}} \tag{5}$$

Partitioning for the feature selection problem reduces to determine an order for the features and then the subregions correspond to either including a feature or not including a feature. Thus, assuming that the current most promising region is some

Given $d_{stop}(n)$, δ, Ψ and an order $a[1]$, $a[2]$,..., $a[n]$ of features
Initialize $A(0) \leftarrow A$, $k \leftarrow 0$, $A^*=\{\ \}$ and
Loop
  $A_1(k) \leftarrow \{A\in A(k);\, a_{d(k)}\in A\}$,
  $A_2(k) \leftarrow \{A\in A(k);\, a_{d(k)} \not\in A\}$,
  $A_3(k) = A\setminus A(k)$,
  for everyset $A_j(k)$
   $A^j_{best}(k) \leftarrow \{\ \}$, $f^j_{best}(k) \leftarrow \infty$, $i \leftarrow 1$
   loop
    $A_{ji}(k) \leftarrow$ Randomly select a feature subset
    if $f_{ji}(k) < f^j_{best}(k)$
    then $f^j_{best}(k) \leftarrow f_{ji}(k)$,
      $A^j_{best}(k) \leftarrow A_{ji}(k)$
    $i \leftarrow i+1$
   until enough feature subset samples given δ and Ψ
    $j \leftarrow \arg\min_j f^j_{best}(k)$
   if $j^*= 3$ then $A(k+1) \leftarrow A(k-1)$
   else $A(k+1) \leftarrow A_{j^*}(k)$
   $k \leftarrow k+1$
  end
until $d(A(k)) = d_{stop}(n)$

<Figure 2> NP Feature Selection Pseudocode

subset $A(k) \subset A$ of the entire feasible region, then this subset is partitioned by fixing the next feature a in the order, that is, the subsets are

$$A_1(k)=\{A\in A(k)\ :\ a\in A\} \tag{6}$$

$$A_2(k)=\{A\in A(k)\ :\ a\not\in A\} \tag{7}$$

The surrounding region is simply $A_3(k)=A\setminus A(k)$. Each of these three regions is then sampled as discussed above and based on these samples the next most promising region is selected. In theory, the features can be selected in an arbitrary order, but an intelligent partitioning where features are ordered according to their information gain performs significantly better, and this partitioning is used in all of the numerical experiments below. A complete description of the NP-Filteris shown in <Figure 2>. Note that it uses a fixed number of n_0 samples to evaluate each region, starts with the set A of all possible feature subsets as the most promising region, and terminates when the depth of the most promising region has reached maximum, that is, it is a singleton. We also let A^* be the best feature subset found and f^* be the corresponding performance value, which is calculated according to equation (1) above.

3. Feature Selection for Mixed Type of Data

The new NP based feature selection method uses information gain filter to determine a partitioning order of features. Therefore, it must discretize the data if a feature consists of continuous data. Although many discretization methods have been shown to have good performances, this approach has a disadvantage in that it does not use characteristics of continuous values itself. Thus we will use two feature quality evaluators, correlation based evaluator and ReliefF as stated in the introduction for determining a partition order in the NP-Filter. For a correlation based feature evaluator, we modify equation (1) which was originally developed for evaluating feature subsets [5]. It simply calculates the correlation between feature and class, which can be regarded as a Pearson's correlation. ReliefF is an extended version of RELIEF developed by [7] for estimating the quality of features that consider interdependency between features. Since ReliefF was developed to solve such problems, it will be one of the feature quality evaluators adapted to handle mixed type of features. Searching for its two nearest neighbors, one from the same class (nearest hit) and the other from a different class (nearest miss) given an instance, RELIEF presented in <Figure 3> estimates qualities of features based on how well each can separate neighbor instances from different classes by having different values and have the same values for neighbor instances from the same class [8]. The RELIEF randomly selects m training instances, where mis the user-defined parameter and usually set to 10 that has been believed to perform well in many cases.

```
for i = 1 to n
Q[A_i] = 0.0
for j = 1 to m
randomly select an instance r
find nearest hit h and nearest miss t
for  i = 1 to n
Q[A_i] = Q[A_i]-diff(A_i, r, h)/m
        +diff(A_i, r, t)/m
```

<Figure 3> Pseudocode of RELIEF Algorithm[8]

In the pseudo code of RELIEF, for a discrete feature, $diff(Feature, Instance1, Instance2) = 0$, if the values are equal, otherwise it is 1, while for a continuous feature the difference is the actual difference normalized to the interval [0, 1]. A High value of $Q[A_i]$ has small amount of difference value for same class instances and large amount of difference for different class. As stated previously, ReliefF can handle incomplete and multi-class data. ReliefF finds one near miss $t(C)$ for each different class C and calculates weighted average with the prior probability of each class. Thus, it can estimate the capability of features to distinguish each pair of classes regardless of which two classes are closer to each other. Since the portion of dealing with incomplete data in the ReliefF is somewhat related to a feature filter, it is discussed further here. Using correlation and ReliefF as a feature evaluator, we compare performances with information gain evaluator in perspective of the NP-Filter with several classifiers. The results are reported in the next section.

4. Numerical Results

In this section we present numerical results for tests whether the NP-Filter can be adapted to handle the mixed type of data. We used 15 data sets from the *UCI* Repository of machine learning databases [1]. The characteristics of these data sets are shown in <Table 1>.

<Table 1> Characteristics of the test datasets

Data Set	Instance	Features	Type
lymph	148	18	nominal values
vote	435	16	
audiology	226	69	
cancer	286	9	
kr-vs-kp	3196	36	
anneal	898	38	mixed values
hepatitis	155	19	
credit-g	1000	20	
hypothyroid	3772	29	
labor	57	16	
vehicle	946	18	continuous values
glass	214	9	
ionosphere	351	34	
segment	2310	19	
diabetes	768	8	

The data sets can be categorized into three different domains in terms of data type of features, namely discrete, mixed (discrete and continuous) and continuous. However, all the data sets have a nominal class feature. This experiment addresses the effectiveness of a feature evaluator on performances in terms of accuracy and size using the NP-Filter for the data sets as stated in <Table 1>, and evaluates if the evaluators perform differently according to each domain.

Three classifiers, Naïve Bayes, C4.5, and 5-nearest neighbor are used to induce classification models with the selected features. For simplicity, the parameter k in the k-nearest neighbor learner is arbitrarily set to 5 since the number of selected features may vary for each run. Each case is repeated 5 times, and 10-fold cross validation averages and standard deviations are reported. First we consider the accuracy of the models induced after feature selection with three different feature evaluators, and second consider simplicity, that is a size of the models after feature selection is employed. Strictly speaking, the accuracy of the Naïve Bayes learner and size of selected features reported in <Table 2> do not show a significant difference between feature evaluators. However, if we admit even small differences, the correlation evaluator has better performances in the discrete data set domain where 3 out of 5 data sets have higher accuracies. Here the better performance implies higher averaged accuracy and size.

If a tie in average of accuracy occurs, lower standard deviation is better. Smaller size can be also regarded as better performance with a premise where there is significant difference in the number of selected features (size), but it does not present any prominent pattern on results we can notice. In mixed type of data domain, the ReliefF shows better performances in 3 out of 5 data sets, and correlation and ReliefF report better performances in the mixed and continuous data type domain, which implies that capability handling a continuous value itself may affect the results. In <Table 2>, First for the accuracy we note that it actually improves or is no worse when we use feature selection except just 4 data sets such as 'audiology', 'lymph', 'anneal', and 'labor'. Such improvement in accuracy may or may not occur as discussed previously, but feature dimension reduction, real contribution of feature selection resulting in simpler and easier to explain models, was accomplished.

The accuracy results for C4.5 are reported in <Table 3>. These information gain evaluators applying entropy discretization performs better in the discrete data domain. In that the discrete domain does not need to employ discretization and C4.5 uses the information gain as a feature selection order criterion in a tree composition procedure, it is not surprising that the information gain evaluator may perform better in this domain. In the mixed data type domain, the correlation and ReliefF methods perform better as expected in terms of accuracy. The ReliefF performs clearly better in 4 out of 5 data sets in the continuous data domain. For the number of selected features, even though the ReliefF evaluator in the NP-Filter performs pretty well in the discrete data domain, it does poorly in the mixed

<Table 2> Comparison of Naive Bayes with Feature Evaluators

Data Set	Info. Gain		Correlation		ReliefF	
	Accuracy	Size	Accuracy	Size	Accuracy	Size
vote	93.7* ± 0.9	5.8* ± 0.8	93.3 ± 1.3	5.8 ± 2.4	93.0 ± 0.7	7.0 ± 1.0
audiology	70.1 ± 2.3	27.0 ± 4.0	70.1* ± 0.8	26.8 ± 3.3	69.2 ± 1.1	21.2* ± 4.4
cancer	73.9 ± 0.4	5.4 ± 0.5	74.0* ± 0.3	5.4* ± 0.5	72.8 ± 1.6	5.4 ± 0.9
kr-vs-kp	86.2 ± 6.0	11.4 ± 1.9	85.7 ± 4.8	11.2 ± 1.3	89.0* ± 8.3	8.8* ± 3.0
lymph	83.4 ± 1.7	11.0 ± 1.0	84.5* ± 1.9	9.8* ± 1.8	83.5 ± 1.9	10.8 ± 1.9
anneal	85.1 ± 1.7	12.0 ± 2.0	83.4 ± 8.0	12.0±1.4	86.2* ± 2.8	10.2* ± 2.9
hepatitis	85.3* ± 1.1	10.4 ± 1.1	85.0 ± 0.8	10.0* ± 0.7	83.7 ± 1.2	11.4 ± 1.1
credit-g	73.6 ± 0.8	6.8* ± 0.8	74.8* ± 0.8	8.0 ± 1.0	73.5 ± 1.7	9.0 ± 1.7
hypothyroid	94.4 ± 0.3	7.6 ± 1.5	94.4 ± 0.4	7.4 ± 2.9	94.4* ± 0.3	4.6* ± 1.3
labor	91.2 ± 0.0	5.8 ± 1.5	92.3 ± 1.0	5.6* ± 1.5	93.0* ± 0.0	7.2 ± 2.6
vehicle	46.6 ± 2.0	9.8* ± 2.6	46.9* ± 1.1	10.0 ± 0.8	46.6 ± 0.7	11.8 ± 1.6
glass	48.6 ± 0.0	7.0 ± 0.0	48.6 ± 0.0	7.0 ± 0.0	48.8* ± 1.8	5.4* ± 1.1
ionosphere	88.7 ± 1.6	16.6* ± 1.5	86.4 ± 2.1	19.0 ± 1.9	88.7* ± 1.4	18.0 ± 3.2
segment	85.1 ± 2.2	6.4* ± 1.1	86.2* ± 4.0	7.2 ± 1.3	81.9 ± 4.2	8.2 ± 1.9
diabetes	76.5* ± 0.6	4.0 ± 0.7	75.9 ± 1.1	3.8* ± 0.8	75.9 ± 0.9	4.6 ± 0.9

Note : * implies the best performance among three evaluators.

<Table 3> Accuracy Comparison of C4.5 with Feature Evaluators

Data Set	Info. Gain		Correlation		ReliefF	
	Accuracy	Size	Accuracy	Size	Accuracy	Size
vote	95.6 ± 0.1	6.0 ± 0.7	95.5 ± 0.2	5.8 ± 0.8	95.6* ± 0.0	5.2* ± 1.3
audiology	75.3 ± 4.0	27.4 ± 1.7	73.0 ± 4.7	25.2 ± 2.2	76.2* ± 1.3	23.0* ± 3.6
cancer	74.6* ± 1.1	5.8 ± 1.1	74.3 ± 0.9	5.8 ± 1.3	73.2 ± 2.8	5.2* ± 0.8
kr-vs-kp	93.3* ± 1.4	11.0 ± 3.2	91.2 ± 4.2	11.8 ± 2.8	92.0 ± 4.2	10.4* ± 2.5
lymph	79.2* ± 0.5	10.8 ± 1.9	76.8 ± 2.9	10.6 ± 1.8	76.9 ± 3.7	8.6* ± 1.9
anneal	96.6± 2.0	12.4* ± 1.8	97.1 ± 1.0	13.0 ± 2.1	97.7* ± 0.4	12.8 ± 1.1
hepatitis	79.6* ± 0.9	10.6 ± 1.1	79.2 ± 1.0	9.4* ± 0.5	78.9 ± 1.9	11.6 ± 0.9
credit-g	74.2± 1.2	6.2* ± 1.3	74.2* ± 0.9	6.8 ± 1.3	72.7 ± 0.8	8.0 ± 2.6
hypothyroid	97.0 ± 0.4	5.8 ± 2.2	97.3* ± 0.6	5.3 ± 1.9	97.2 ± 0.7	4.8* ± 1.9
labor	84.9 ± 2.3	6.2* ± 1.5	83.9 ± 1.5	7.4 ± 1.1	84.9* ± 1.6	6.6 ± 1.9
vehicle	69.1 ± 1.8	10.0 ± 1.0	69.8 ± 1.6	9.8* ± 1.3	71.7* ± 2.3	11.0 ± 2.4
glass	67.3 ± 0.0	7.0 ± 0.0	67.8 ± 1.0	6.8 ± 0.4	67.8* ± 1.3	6.6* ± 0.5
ionosphere	89.3 ± 1.6	17.8 ± 2.3	89.2 ± 1.4	16.6* ± 1.5	90.7* ± 1.0	16.8 ± 1.1
segment	96.6 ± 0.3	8.2 ± 1.3	96.6* ± 0.1	9.4 ± 0.5	96.4 ± 0.2	4.8* ± 0.8
diabetes	75.1 ± 0.2	3.4* ± 0.5	75.3 ± 0.6	4.4 ± 0.5	75.3* ± 0.1	3.4 ± 0.9

<Table 4> Accuracy Comparison of 5-Nearest Neighbor with Feature Evaluators

Data Set	Info. Gain		Correlation		ReliefF	
	Accuracy	Size	Accuracy	Size	Accuracy	Size
vote	94.3 ± 1.4	6.2* ± 1.5	94.7* ± 1.0	7.0 ± 2.5	94.6 ± 0.9	7.2 ± 0.8
audiology	68.3* ± 2.1	24.8 ± 3.8	64.3 ± 1.4	24.2 ± 8.7	66.7 ± 4.4	20.0* ± 1.7
cancer	71.3 ± 0.0	5.4 ± 0.5	72.7 ± 1.5	5.0* ± 0.7	74.8* ± 1.0	5.0 ± 1.9
kr-vs-kp	89.9 ± 6.5	11.0 ± 2.9	89.4 ± 5.7	11.0* ± 2.4	91.9* ± 5.8	11.2 ± 2.6
lymph	83.5* ± 0.8	10.8 ± 1.8	81.1 ± 1.7	9.2* ± 1.5	81.9 ± 2.3	10.2 ± 1.5
anneal	97.0* ± 0.9	12.6 ± 0.9	96.2 ± 1.0	15.0 ± 4.0	95.1 ± 1.5	11.6* ± 3.0
hepatitis	83.6* ± 1.1	10.2 ± 1.3	83.0 ± 1.9	10.0* ± 0.7	83.4 ± 1.4	10.6 ± 2.1
credit-g	72.3 ± 0.6	7.8 ± 1.3	73.3* ± 1.0	5.8* ± 1.3	71.8 ± 1.2	8.8 ± 2.1
hypothyroid	96.4* ± 1.5	5.4* ± 2.5	95.5 ± 0.8	6.4 ± 2.5	95.2 ± 0.9	6.4 ± 1.7
labor	88.8 ± 1.6	7.8 ± 1.1	89.8 ± 2.3	6.6 ± 1.9	90.5* ± 1.6	6.6* ± 1.1
vehicle	62.1 ± 3.7	10.0* ± 1.4	63.0 ± 2.3	10.8 ± 0.8	65.4* ± 3.3	11.6 ± 1.1
glass	72.2* ± 0.4	6.8 ± 0.4	72.0 ± 0.0	7.0 ± 0.0	71.2 ± 0.8	6.4* ± 0.5
ionosphere	87.3* ± 1.0	17.6 ± 3.0	86.8 ± 1.9	16.8* ± 1.8	86.3 ± 1.0	18.0 ± 1.0
segment	95.4 ± 0.2	7.2 ± 0.4	95.7 ± 0.4	8.0 ± 1.4	95.8* ± 0.2	6.0* ± 1.0
diabetes	73.2 ± 0.4	3.6* ± 0.5	73.9 ± 0.5	4.2 ± 0.8	74.4* ± 0.9	4.2 ± 1.1

data domain. Accuracies of 5-nearest neighbor in <Table 4> have similar pattern as those of the previous two results.

5 Conclusions

The major contribution of this paper is that the new approach with an optimization framework can guarantee a near optimal solution within a certain distance and a given probability after a finite time stopping criterion is satisfied. Further we showed that the NP-Filter is capable of handling the mixed type of data. With regard to the accuracy, the information gain evaluator with entropy discretization performs a little better in the discrete data domain. On the other hand, the ReliefF presents better performances in the continuous data domain. In the mixed type data domain, any evaluator did not show conspicuously better results. However we achieved significant reduction on feature dimension.

References

[1] Blake, C. L. and Merz, C. J.; UCI Repository of machine learning databases <http://www.ics.uci.edu/mlearn/MLRepository.html>, University of California, Irvine, CA, 2006.

[2] Bradley, P. S., Mangasarian, O. L., and Street, W. N.; "Feature selection via mathematical programming," *INFORMS Journal on Computing*, 10(2) : 209-217, 1998.

[3] Clark, G., Sengupta, S., Aimonentti, W. D., Roeske, F., and Donetti, J. G.; Multispectral image feature selection for land mine detection, IEEE Transactions on Geoscience and remote sensing," 38(1) : 304-311, 2000.

[4] Evgeniou, T., Pontil, M., Papageorgiou, C., and Poggio, T.; "Image representations and feature selection for multimedia database search," IEEE Transactions on Knowledge and Data Engineering, 15(4) : 911-920, 2003.

[5] Hall, M. A.; "Correlation-based feature selection for discrete and numeric class machine learning, in Proceedings of the Seventeenth International Conference on Machine Learning," Stanford University, CA. Morgan Kaufmann, 1998.

[6] Kim, Y. S., Street, W. N., and Menczer, F.; "Feature slection in unsupervised learning via evolutionary search," in Proceedings of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000.

[7] Kira, K. and Rendell, L.; "A practical approach to feature selection, in Proceedings of International Conference on Machine Learning," ICML-92 : 249-256, 1992.

[8] Kononenko, I.; "Estimating features : analysis and extensions of RELIEF," in Proc. European Conf. on Machine Learning, 1994.

[9] Liu, H. and Motoda, H.; "Feature Extraction, Construction and Selection; A Data Mining Perspective," Kluwer Academic Publishers, 1998.

[10] Narendra, P. M. and Fukunaga, K.; "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, 26(9) : 917-922, 1977.

[11] Olafsson, S. and Yang, J.; "Intelligent partitioning for feature selection," *INFORMS Journal on Computing*, 17(3) : 339-355, 2005.