International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.4 (2016), pp.67-78
http://dx.doi.org/10.14257/ijmue.2016.11.4.08

An Approach to Feature Selection for Continuous Features of Objects

Wang Hong-Wei[1,2], Li Guo-He[1] and Li Xue[3]

[1]*College of Geophysics and Information Engineering, China University of Petroleum, Beijing, 102249, China*
[2]*College of Information Science and Technology, Bohai University, Jinzhou 121013, China*
[3]*School of Information Technology and Electric Engineering, University of Queensland, Brisbane 4072, Australia)*

Abstract

A novel approach to feature selection is proposed for data space defined over continuous features. This approach can obtain a subset of features, such that the subset features can discriminate class labels of objects and the discriminant ability is prior or equivalent to that of the original features, so to effectively improve the learning performance and intelligibility of the classification model. According to the spatial distribution of objects and their classification labels, a data space is partitioned into subspaces, each with a clear edge and a single classification label. Then these labelled subspaces are projected to each continuous feature. The measurement of each feature is estimated for a subspace against all other subspace-projected features by means of statistical significance. Through the construction of a matrix of the measurements of the subspaces by all features, the subspace-projected features are ranked in a descending order based on the discriminant ability of each feature in the matrix. After evaluating a gain function of the discriminant ability defined by the best-so-far feature subset, the resulting feature subset can be incrementally determined. Our comprehensive experiments on the UCI Repository data sets have demonstrated that the approach of the subspace-based feature ranking and feature selection has greatly improved the effectiveness and efficiency of classifications on continuous features.

Keywords*: Continuous Features, Feature Ranking, Data Reduction, Feature Selection*

1. Introduction

Feature selection is one of the methods used in dimensionality reduction [1] to find a subset of original features. Feature selection has wide applications, such as pre-processing of knowledge modeling processes, automatic parameter optimization, *etc.*

Feature selection can be regarded as a process of searching through the subset feature space. However, the exhaustive search in order to obtain a global optimal feature subset is a NP-Hard problem [2]. So a heuristic approach should always be considered. The solutions of heuristic searches are divided into two categories [3]: *Wrapper* and *Filter*.

Feature selection with Wrapper [4] is tightly bound to a given classifier. A feature subset by a Wrapper approach is only suitable for the given classifier, not common for any other classifiers. In order to maintain classification accuracies, this method has to learn and classify repeatedly with different feature subsets, resulting in the inefficiency of feature selection. For this reason, Wrapper has higher computational cost.

Feature selection with Filter [3] is independent from a classifier. This approach directly filters out irrelevant features by user-defined criterion for obtaining a feature subset. Since it is independent from induction learning methods, the feature selection with Filter usually

ISSN: 1975-0080 IJMUE

has excellent performance for getting a feature subset, which is independent from any classifiers. The good behaviours of Filter lead to being widely used in feature selections.

Relief-series are the classic methods of feature selection with Filter. Relief series includes *Relief, ReliefF, RReliefF* and their ameliorations. ReliefF extends Relief from binary classification to multi-classification, and applies KNN (K Nearest Neighbours) method to measure features for the robustness of feature selection. Stanczyk U. [5] used Relief algorithm to rank conditional attributes, and generate decision rules on examples within Dominance-Based Rough Set Approach. In order to effectively use the rich features of the image objects, Jia J. [6] *et al*. also used Relief algorithms to improve in terms of aspects of randomly drawing samples, the influence of sample quantity variance, and iteration times to evaluate the features. Wang P. [7] *et al*. combined DDNA and SOEKS with feature selection learning algorithm RELIEF-F to improve the quality of predictions.

Above algorithms use Euclidian distance function for the projection of objects on each feature to measure the discriminant ability of features in classifications. We identify their drawbacks as that:

(1) Those methods only considered the objects and their local spatial distributions in the data space (*e.g*. Nearest Neighbour), ignoring the overall spatial distributions of objects. In this case, the overall ability of discriminating objects of a feature cannot be expressed;

(2) Those computations only used distance functions amongst objects as a measurement to calibrate the feature discriminant ability. However, distances from different feature domains, whether long or short, should have the equal ability in classification as long as they discriminate objects for their corresponding class labels;

In order to overcome the drawbacks identified above, we propose a novel method namely, Feature Selection for Continuous Features based on the Distribution of Objects (FSFSF).

This paper is organized as follows. Section 2 introduces the basic concepts. Section 3 presents the ideas and the algorithms of our work. Section 4 describes the experiments and their significant results. Section 5 gives the conclusions.

2. Basic Concepts

2.1. Information Model and Feature Selection

An information model K is an abstract presentation of a data set[8], denoted as $K=(U, A, V)$, where $U=\{u_1, u_2, \ldots, u_{|U|}\}$ is a set of object identifiers ($|U|$ is the cardinality of objects in U); $A=\{a_i|i=1, 2, \ldots, k\}$ is a set of features of objects. $V=\{Va_i|i=1, 2, \ldots, k\}$, Va_i is a value domain set for feature a_i, denoting object $u \in U$ as the projection of a_i, namely $a_i : U \rightarrow Va_i$, $a_i(u) \in Va_i$. If feature a_i assumes real numbers, it is called a continuous feature and the value domain set Va_i has a continuous range. If $A=C \cup D$, and $C \neq \Phi$, $D \neq \Phi$, $C \cap D = \Phi$, where C is a condition-feature set, and D is a decision-feature set, then a decision table (DT) can be defined over K.

If the condition-feature subset $Fs \subseteq C$ exists, and the ability of Fs classification is not inferior to that of C classification, then Fs is called a solution of feature selection.

This paper presents the feature selection FSFSF for continuous condition-feature set and discretized decision-feature set. All the features below imply the continuous condition features if not mentioned especially.

2.2. Distribution Center and Radius

Given a subset $A' \subseteq A$ and any *subset* $S \subseteq U$, the center and radius of S with respect to A' are defined as follows:

$$Center\ (S, A') = \{\mu_{(S,a)} \mid \forall a \in A', \mu_{(S,a)} = \frac{\sum_{x \in S} a(x)}{|S|}\}. \quad (1)$$

$$Radius\ (S, A') = \max_{obj \in S}(\{dist\ (obj, Center\ (S, A'))\}). \quad (2)$$

Where *dist* is distance function (such as Euclidean distance etc).

3. Feature Selection on Continuous Features

3.1. Covers and Its Optimization

Given a decision table $DT=(U, C \cup D, V)$, the process of clustering with condition-feature set C by Nearest Neighbour (*NN*) algorithm generates a set of clusters, denoted as $Clus(U, C)$, which satisfies: (1) $|S_i| \geq 2$, $\forall u,v \in S_i$, $D(u)=D(v)$ for $\forall S_i \in Clus(U, C)$, and (2) if $\forall w \in U$, $Radius(S_i \cup \{w\},C) > Radius(S_i,C)$, then at least $\exists x,y \in U$, $D(x) \neq D(y)$. The hypersphere with radius $Radius(S_i,C)$, including S_i, is called a Cover of *DT*.

Decision table *DT* corresponds to a set of Covers, denoted as $CoverSet(DT)$, defined over continuous feature space. For any *DT*, $CoverSet(DT)$ is relatively steady. Obviously, $|CoverSet(DT)| \leq |U|$, and this inequation is helpful to select feature subset. If $|S_i|=1$, then S_i may be regarded as an outlier or a noisy set, and deleted further.

3.2. Matrix of Feature Discriminant Ability

The discriminant ability of a feature is denoted as *DAF*. Firstly, the range of any feature $c \in C$ is determined by the projections of all Covers on this feature, $\forall Co \in CoverSet(DT)$, expressed as follows:

$$Range_c(Co, \alpha) = [\mu_{(Co,a)} - \sigma_{(Co,c)} \times Z_{\alpha/2}, \mu_{(Co,a)} + \sigma_{(Co,c)} \times Z_{\alpha/2}]. \quad (3)$$

Where $\mu_{(Co, a)}$ is mean, $\sigma_{(Co, c)}$ is standard variance, $Z_{\alpha/2}$ is double-sided values of *Normal Distribution Standard* under probability α, namely:

$$P(\frac{|X(c) - \mu_{(Co,c)}|}{\sigma_{(Co,c)}} \leq Z_{\alpha/2}) = \alpha \quad (4)$$

And $Z_{\alpha/2} \in [0,3.09]$, $\alpha \in [0,1]$.

For Co_1, $Co_2 \in CoverSet(DT)$, and $D(Co_1) \neq D(Co_2)$, the discriminant ability of feature *c* between Co_1 and Co_2 is defined as:

$$DP_c(Co_1, Co_2, \alpha) = 1 - \frac{|Range_c(Co_1, \alpha) \cap Range_c(Co_2, \alpha)|}{|Range_c(Co_1, \alpha) \cup Range_c(Co_2, \alpha)|}. \quad (5)$$

As we can see from Formula 5, $DP_c(Co_1,Co_2, \alpha) \in [0,1]$, implies that if the objects in each Cover are large in number and much centralized, and the distance between two Covers is much long on feature *c*, then the discriminant ability of feature *c* is much strong. $DP_c(Co_1,Co_2, \alpha)=1$, denoting Co_1 and Co_2 with different labels is able to be discriminated by *c* under probability α. Certainly a Cover defines an equivalent class satisfying equivalent relationship *(viz.* $\sigma_{(Co,c)}=0$*)*.

On the basis of Formula 5, a Feature-Importance Matrix (FIM), which presents all the discriminant abilities of all the features by all the Covers of *DT*, is defined below:

$$FIM\ (DT, \alpha) = (dp_{(Co,c)})_{\substack{Co \in CoverSet\ (DT) \\ c \in C}}. \quad (6)$$

Where

$$dp_{(Co,c)} = \frac{\sum_{c \in C} DP_c(Co, Other, \alpha)}{\substack{\forall Other \in CoverSet\ (DT) \\ D(Other) \neq D(Co)} \over |CoverSet\ (DT)|}. \quad (7)$$

Formula 7 is the discriminant ability of feature *c* which distinguishes Cover *Co* from all the other Covers with the labels different from that of *Co*. Obviously, $dp_{(Co,c)} \in [0,1]$. So one can see that every row in FIM corresponds to a Cover, and every column in FIM corresponds to a feature.

3.3. Feature Ranking

```
ALGORITHM 1   FeatureRanking (EFIM(DT, α), C);
(* Input:   EFIM, Probability, and condition-feature set *)
(* Output: Features in descending order *)
If C ≠ Φ Then   (*the most powerful column (feature) *)
   c_max = arg  max   {dp_(Co,c) ∈ EFIM (DT, α)};
                Co ∈ CoverSet (DT)
                Selected (Co) = False
                c ∈ C − RankedFeatureSet
  For ∀r∈[1..|CoverSet(DT)|]   (* a row(viz. Cover)*)
  Begin
    For ∀c∈C-RankedFeatureSet   (* all columns(viz. all features) *)
      If c(r) ≤ cmax(r) Then Selected(r)=True; (* update tag *)
      If ∀r∈[1..|CoverSet(DT)|] And Selected(r)=True Then   (* all rows (viz. all Covers) *)
        Selected(r)=False;   (* if all rows is tagged, then clear all tags *)
  End;
  RankedFeatureSet={cmax} ∪ FeatureRanking (FIM(DT,α), C-{cmax});
  (* append a feature recursively *)
Else
  Return RankedFeatureSet;
EndIf;
```

Figure 1. The Ranking Algorithm for Subspace-Projected Features

In order to measure the discriminant abilities of all features through the Covers, and then rank all features in descending order by the discriminant abilities, a column namely, *Selected* is added to FIM, called Extended FIM (EFIM), which is helpful to design the *Feature Ranking* algorithm (Figure 1). The *Selected* value is either true or false, designating whether the raw is used or not in current feature ranking. The algorithm of *Feature Ranking* based on EFIM is described in detail in Figure 1.

From Algorithm 1 (Figure 1), it can be seen that a feature that has the most powerful discriminant ability of a certain Cover in EFIM is selected first. Furthermore, if the discriminant abilities of the feature for all other Covers are the most powerful too, those Covers do not participate in feature ranking process, because those Covers would not be able to make any difference in assessing the discriminant ability of a feature. Through a recursive process, the descending order of all features is finally obtained on the basis of the discriminant abilities of all the features that possess global superiority.

3.4. Gains of Discriminant Ability of Feature Subset

For $\forall Co_1, Co_2 \in CoverSet(DT)$, and $D(Co_1) \neq D(Co_2)$, namely different labels between Co_1 and Co_2, the discriminant ability of feature subset $C' \subseteq C$ is defined as below:

$$DP_{C'}(Co_1, Co_2, \beta) = 1 - \frac{|Aera_{C'}(Co_1, \beta) \cap Aera_{C'}(Co_2, \beta)|}{|Aera_{C'}(Co_1, \beta) \cup Aera_{C'}(Co_2, \beta)|}. \quad (8)$$

Where $Aera_{C'}(Co, \beta)$ is a bounded hyper-geometrical body in C' space with the statistical significance under the probability β, and simplified a bounded sphere with *Center(Co, C')* as center and radius.

$$Radius_{C'}(Co, \beta) = \underset{c \in C'}{Max} \{\sigma_{(Co,c)} \times Z_{\beta/2}\} \quad (9)$$

Generally, suppose $Radius_{C'}(Co_2, \beta) \geq Radius_{C'}(Co_1, \beta)$ and the connected line between two centers as a projection axis, then Formula 8 is simplified as:

$$DP_{C'}(Co_1, Co_2, \beta) = \begin{cases} 1 & d > Radius_{C'}(Co_2, \beta) + Radius_{C'}(Co_1, \beta) \\ \dfrac{Radius_{C'}(Co_1, \beta)}{Radius_{C'}(Co_2, \beta)} & d < Radius_{C'}(Co_2, \beta) - Radius_{C'}(Co_1, \beta) \\ \dfrac{2d(Co_1, Co_2, C')}{Radius_{C'}(Co_1, \beta) + Radius_{C'}(Co_2, \beta) + d(Co_1, Co_2, C')} & others \end{cases} \quad (10)$$

Where d is a distance function with feature set C' between Co_1 and Co_2. Obviously $DP_{C'}(Co_1, Co_2, \beta) \in [0,1]$, denoting the discriminant ability of C' to distinguish Co_1 and Co_2 with different labels with the statistical significance under probability β. If $DP_{C'}(Co_1, Co_2, \beta)=1$, then two Covers can be discriminated clearly. If $DP_{C'}(Co_1, Co_2, \beta)=0$, then two Covers cannot be discriminated clearly.

Aiming at all the Covers, the discriminant ability of feature set C' is defined:

$$DP_{C'}(CoverSet(DT), \beta) = \sum_{\substack{\forall Co_1, Co_2 \in CoverSet(DT') \\ D(Co_1) \neq D(Co_2)}} \frac{DP_{C'}(Co_1, Co_2, \beta)}{2 \times |Co_1| \times |Co_2|} \quad (11)$$

Obviously $DP_{C'}(CoverSet(DT), \beta) \in [0,1]$. When the feature set changes with features in it, the discriminant ability of feature set maybe change too. Thereby the gain $DPG(CoverSet(DT), C_1 \cup C_2, \beta)$ of discriminant ability of feature set is defined below:

$$DPG(CoverSet(DT), C_1 \cup C_2, \beta) = DP_{C_1 \cup C_2}(CoverSet(DT), \beta)) - DP_{C_1}(CoverSet(DT), \beta)) \quad (12)$$

Where $C_1, C_2 \subseteq C$, denoting the change of discriminant ability of feature set with the extension from C_1 to $C_1 \cup C_2$ with the statistical significance under probability β. For given $\forall \varepsilon$ $(\varepsilon \geq 0)$, if $DPG(CoverSet(DT), C_1 \cup C_2, \beta) > \varepsilon$, then it shows the ascent of the discriminant ability of feature set by C_2. Otherwise if $DPG(CoverSet(DT), C_1 \cup C_2, \beta) \leq 0$, it shows the descent of the discriminant ability of feature set by C_2. So that C_2 with powerful discriminant ability can be determined according to the gain $DPG(CoverSet(DT), C_1 \cup C_2, \beta)$.

3.5. FSFSF Algorithm

```
ALGORITHM 2      FSFSF (DT,α,β, ε);
(* Input: Decision table, Probability, Threshold *)
(* Output: Selected feature subset *)
CoverSet(DT) ;              (* form Cover and optimization *)
FIM(DT, α);                 (* form Feature Importance Matrix*)
RankedFeatureSet=FeatureRanking(FIM(DT,α),C);   (*obtain ranked features by discriminant ability *)
FeatureSubSet=RankedFeatureSet(0);      (*select feature by the gain of discriminant ability *)
For i=1 to |C|-1
  If DPG(CoverSet(DT), FeatureSubSet ∪ {RankedFeatureSet(i)}, β) > ε Then
     FeatureSubSet = FeatureSubSet ∪ {RankedFeatureSet(i)};
Return FeatureSubSet;
```

Figure 2. The Main Algorithm of the Proposed Approach

Upon the introduction of above basic concepts, the algorithm of Feature Selection for Continuous Features Based on the Distribution of Objects (FSFSF) is described in detail in Figure 2.

It can be seen from Algorithm 2 that the selected feature subset consists of features added in one-by-one from ranked features, sorted in descending order by the discriminant abilities. If a feature from ranked features improves the gain *DGP* and the ranking is greater than other selected features, the feature is appended to the selected feature subset. Otherwise this feature is dropped.

4. Experiments

Our experiments are conducted on a laptop Lenovo E255 computer with Windows XP platform. FSFSF algorithm and *CoverSet(DT)* are implemented in Visual Basic 6.0 and MS Access 2007.

4.1. Using CoverSet as Classifier

Procedure *CoverSet(DT)* defines a classifier. This subsection demonstrates that *CoverSet(DT)* as a classifier can perform as good as those well-known classifiers such as C4.5, SVM, and CLIP3.

After *CoverSet(DT)* is constructed, it may include some Covers with only one object, so that these Covers may be regarded as noise. Deleting these Covers often results in the combination of Covers or the readjustment of the boundaries of Covers till there is no Cover with a single object in the *CoverSet(DT)*. This process is optimal, which forms the Covers with much more objects with same label and much less number of Covers in *CoverSet(DT)*, to improve the robust and accuracy of classifier *CoverSet(DT)*. For any object *obj*, the classification rules are following:

$$\begin{aligned} &D(obj) = D(Co) \\ &if \exists Co \in CoverSet(DT), \forall C_{other} \in CoverSet(DT) - \{Co\}, obj \in Co, obj \notin C_{other}, or \\ &if \exists Co \in CoverSet(DT), Co = \arg \min_{\forall x \in CoverSet(DT)} \{dist(\mu(x, C), obj) - Radius(x,1)\} \end{aligned} \quad (13)$$

The experimental data sets are downloaded from UCI Repository[9]. The data sets are randomly divided into 10 groups according to *10-Fold Cross Validation*. Every experiment is implemented 10 times (viz. 100-time solutions) and the classification results are regarded as accuracy averages with α*=78%,* β*=100% (viz.* $Z_{\alpha/2}$*=1,* $Z_{\beta/2}$*=3.09).* In Formula 13 *dist* is a Euclidean distance function. The accuracies of other classifiers come from the relevant literatures as shown in Table 1. Accuracies of *CoverSet* for given data sets are better than those of other classifiers except for the accuracy of *Iris*. The results illuminate that *CoverSet* is a good classifier with high classification accuracy.

Table 1. Comparison of Classification Accuracies with Different Classifiers

Data Set	Classifiers	Accuracy
Ionosphere	CoverSet	89.4%
	SVM	61.9%[10]
Sonar	CoverSet	81.4%
	C4.5	60.7±7.2%[11]
Iris	CoverSet	94.4%
	C4.5	94.7%[12]
Spectf	CoverSet	79.7%
	CLIP3	77.0%[9]

4.2. Effectiveness of Feature Ranking

All experiments below adopt our proposed *CoverSet* as classifier to validate the accuracies with different feature sets of every data set.

Parkinson data set consists of 195 records with 23 features, including 22 continuous condition features and 1 decision feature, and is divided into 2 categories. The parameters are set as $\alpha=78\%$, $\beta=100\%$, $\varepsilon=0$ (namely $Z_{\alpha/2}=1$, $Z_{\beta/2}=3.09$). The given 22 condition features are ranked by Algorithm 1: *FeatureRanking*.

In order to illustrate the efficiency of feature ranking, the experiments with Parkinson data set are implemented by 10 times of *10-Fold Cross Validation* (viz. 100-time classifications). The performance is evaluated with the averages of time costs in learning and classification accuracies.

Due to the limited paper length, only a part of total experiment results are shown in Figures 3-7. The axis of "*Data File Number*" in figures denotes for the *i*th (*i=1..10*) file of 10-Fold data files, namely *i*th experiment. The experiment titled "*First Feature Set*" denotes the process that always selects the current best feature first and *vice versa* for the "*Last Feature Set*". The experiment titled "*Random Feature Set*" denotes a set of features randomly selected from ranked feature set. The experiment results are the averages of experiment results (viz. many feature sets with the same size). It is expected that the performance of "*Random Feature Set*" should be better than the "*Last Feature Set*" but worse than the "*First Feature Set*".

Figures 3-6 show that (1) the experiment "*First Feature Set*" has the highest classification accuracy, while the experiment "*Last Feature Set*" has the lowest. The experiment "*Random Feature Set*" has the accuracy between the two formers. This confirms with our hypothesis that our feature selection algorithm does make sense by selecting best features first. (2) The experiment "*First Feature Set*" has the lowest time cost, while the experiment "*Last Feature Set*" has the highest time cost. The experiment "*Random Feature Set*" is between the two formers. (3) Along with the augmentation of features from 4 to 22 ranked features, the differences amongst "*First Feature Set*", "*Last Feature Set*" and "*Random Feature Set*" become smaller, because they include more and more the same features in their feature subsets. When all features are included, these three subsets eventually become the same. Then their classification accuracies are all the same by all the features (see Figure 6). But their time costs are slightly different mainly because the logic order of features is different from that of the physical order in database, bringing up the difference when data is accessed through a DBMS (MS Access 2007) database.

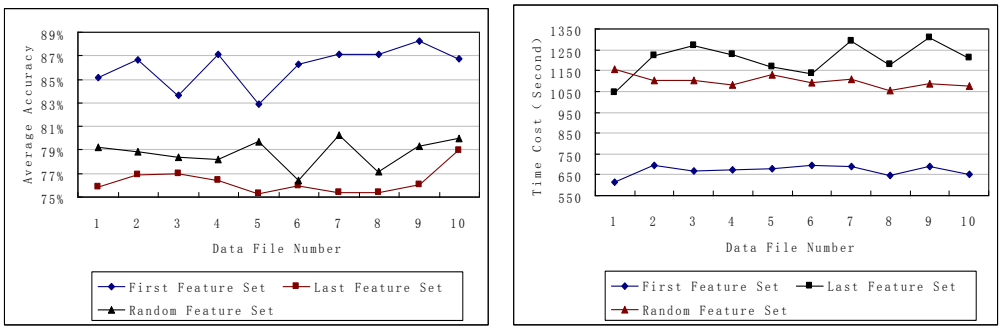


Figure 3. Experiment Results of 4 Features

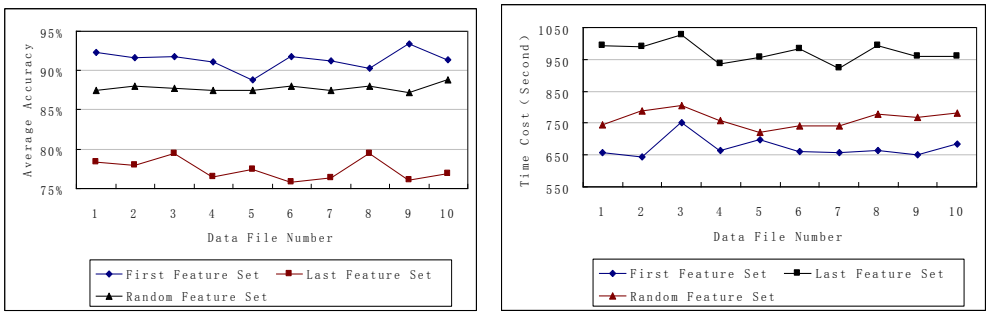


Figure 4. Experiment Results of 10 Features

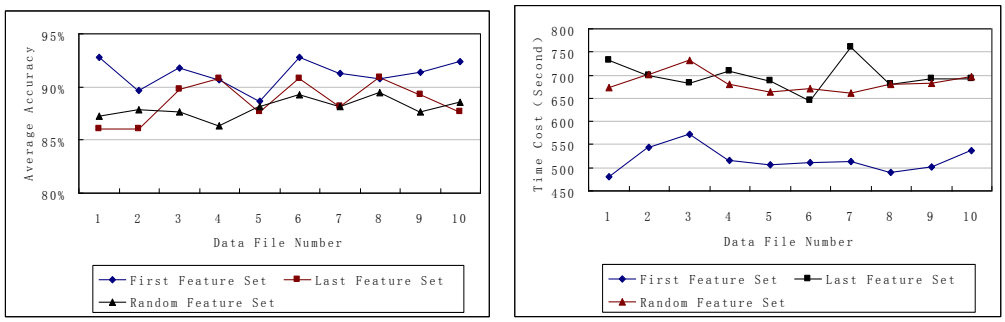


Figure 5. Experiment Results of 15 Features

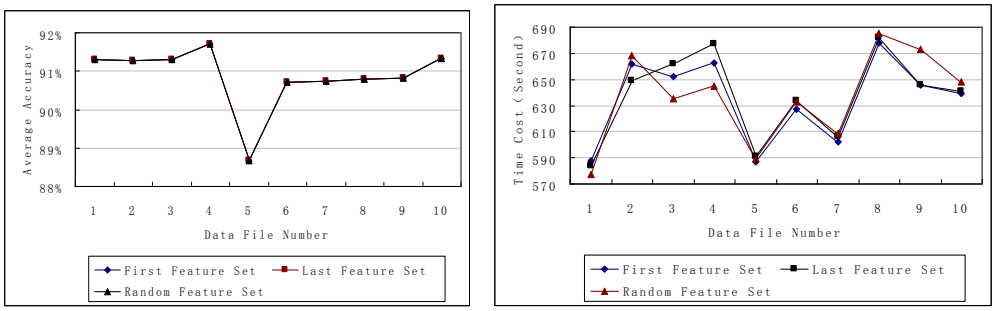


Figure 6. Experiment Results of 22 Features

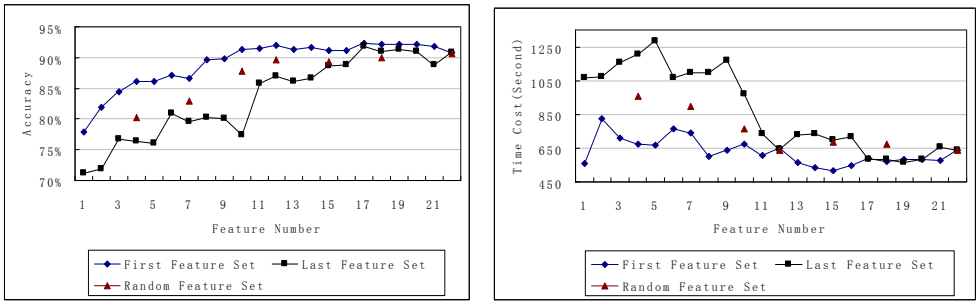


Figure 7. Average Values of 10 Experiments

Figure 7 shows the averages of classification accuracies and time efficiency by the three-feature sets with different feature number. It can be seen from Figure 7 that the classification accuracies gradually improve in tendency along with the number of features increasing. There are some peak points of classification accuracies associated with different feature subsets, implying that the classification accuracies can be improved by finding the feature subsets with low time costs.

The conclusions from experiment results with other data sets from UCI Repository are similar to that with Parkinson data set. These conclusions illuminate that the feature ranking is very effective by the means of the discriminant abilities of condition features determined by Covers of decision table.

4.3. Effectiveness of Feature Selection

The experiments on *Spectf* data set are conducted on the training data set and testing data set downloaded from UCI Repository. The training data set is used for feature selection by FSFSF. The testing data set is used for classification accuracy by *CoverSet*. Other original data sets from UCI Repository are adopted for feature

selection by FSFSF, and then tested for the accuracies of classifier *CoverSet* by 10 *10-Fold Cross Validations* (viz. 100 solutions) with the original and selected feature set respectively. The resulting accuracies and time costs are averaged for every data set for both original and selected feature sets.

Table 2. Comparison of Classification Accuracies with Feature Subset to Original Features

Data Set	Number of Features	Classification Accuracy	Time Cost [Second]
Parkinson	22	90.9%	635.4
	19	92.2%	583.4
Ionosphere	33	89.4%	911.25
	26	90.7%	886
Sonar	60	81.4%	102
	30	81.6%	76
Iris	4	94.4%	102.7
	2	96.7%	85
Spectf	44	79.7%	15
	29	84.0%	9

Table 2 shows for every data set that the former and latter rows are the results by original and selected feature set respectively. It can be seen that the accuracy of the selected feature set is consistently higher than that of the original feature set, whereas the time cost is lower significantly. Although the accuracy for *Sonar* data set with selected features is only slightly higher than that of the original features, the number of features and time cost are significantly reduced. Comparing the experiments on *Iris* data set between Tables 1 and 2, especially note that the accuracy for *Iris* data set with the selected features in Table 2 is greater than that of the original data set using classifier C4.5 shown in Table 1.

Our experiment results illustrate that FSFSF is able to automatically select the feature subset from the original feature set. The discriminant ability of the selected subset features is superior to that of original feature set.

5. Conclusions

Continuous features are common in the data sets collected from sensor networks, land surveys, medical, astronomical and meteorological applications. In those applications, very large and high dimensional data sets are mostly the case. This paper has presented a novel feature selection approach that is based on the projection of subspaces on the features in terms of spatial distribution of data points. Then the features are ranked according to their statistical significance in discriminating class labels. Our experiments on the UCI Repository data sets have successfully shown that the proposed approach namely, FSFSF is effective and efficient.

The time cost of FSFSF mainly lies in the construction of *CoverSet*, which partitions a given feature space into continuous subspaces with clear class labels that possess statistical significance according to the spatial distribution of objects. The time complexity of the *CoverSet* partitioning process is the same as that of a fuzzy clustering approach. Although maybe there are intersections among subspaces, every subspace is associated to a single label. Therefore there is no object falling into the intersections. The subspaces with class labels can be separated by other subspaces with unknown class labels, by means of which the class labels are discriminated.

To sum up, our proposed approach has following characteristics:

(1) Features may have equal discriminant abilities if two subspaces with different labels are projected to the features such that the gap of two projections exists no matter how long in distance between the two subspaces.

(2) FSFSF just considers the ability that can partition objects into subspaces with different class labels. There is no need for an exhaustive search for all possible subspaces of the given feature space. The algorithm emphasizes the discriminant abilities of features corresponding to those subspaces and therefore improves the performance of feature selection.

(3) FSFSF ranks features mainly according to discriminant ability of features by using a discriminant ability matrix to make the features ranking being more righteous and objective, and different from those by the sum of the measurement of discriminant abilities of every feature for all Covers.

(4) On the basis of feature ranking, FSFSF selects the features automatically on a best-so-far basis, which have powerful discriminant abilities to incrementally complete the feature selection.

FSFSF integrates the features with strong discriminant abilities to obtain a much better feature subset in terms of "ranking the features by their discriminant abilities" and "the gain of discriminant ability of a feature set" as the heuristics. To this end, FSFSF is a local optimal algorithm for finding the best subset features. In our experiments, the combination of the features by choosing strong-with-strong discriminant abilities is effective. Our future research work will continue on those with strong-with-weak or weak-with-weak discriminant abilities of subset features, or the solution of a variety of feature subsets.

Acknowledgment

We are greatly indebted to colleagues at Data and Knowledge Engineering Center, School of Information Technology and Electrical Engineering, the University of Queensland. We thank Prof. Xiaofang Zhou for his special suggestions and many interesting discussions.

This work is partly supported by the Nature Science Foundation of China under Grant No. 60473125; National Key Project Foundation of Science under Grant No. G5800-08-ZS-WX.

References

[1] A. Gorban, B. Kegl and D. Wunsch, “Principal Manifolds for Data Visualization and Dimension Reduction”, LNCSE 58, Springer, Berlin - Heidelberg - New York, **(2007)**.
[2] M. R. Garey and D. S. Johnson, “Computers and Intractability a Guide to the Theory of NP-Completeness”, W. H. Freeman and Company, New York, **(1979)**.
[3] H. Liu and H. Motoda, “Feature Selection for Knowledge Discovery and Data Mining”, New York: Kluwer Academic Publishers, **(1998)**.
[4] R. Kohavi and G. H. John, “Wrappers for Feature Subset Selection”, Artificial Intelligence, no. 1–2, **(1997)**, pp. 273-324.
[5] U. Stanczyk, “RELIEF-based selection of decision rules”, Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference. Gdynia, Poland, **(2014)**, pp. 299-308.

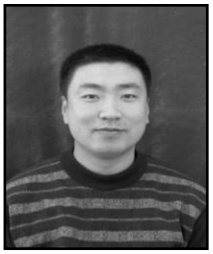
[6] J. Jia, N. Yang and C. Zhang, "Object-oriented feature selection of high spatial resolution images using an improved Relief algorithm", Mathematical and computer modeling, vol. 58, no. 3-4, **(2013)**, pp. 619-626.
[7] P. Wang, C. Sanin and E. Szczerbicki, "Prediction based on integration of decisional DNA and a feature selection algorithm Relief-F", Cybernetics and systems, vol. 44, no. 2-3, **(2013)**, pp. 173-183.
[8] G. Li, "Feature Subset Selection of Information System Based on Similar Extension Matrix", Computer Engineering, vol. 32, no. 17, **(2006)**, pp. 52-54.
[9] UCI Repository Data Sets Download Web Site: http://archive.ics.uci.edu/ml/datasets.html
[10] H. Liu, L. Yu, M. Dash and H. Motoda, "Active Feature Selection Using Classes", In Proceedings of PAKDD, **(2003)**, pp. 474-485.
[11] M. Richeldi and P. L. Lanzi, "ADHOC: a Tool for Performing Feature Selection", Proceedings of Eighth IEEE International Conference on Tools with Artificial Intelligence, **(1996)**, pp. 102-105.
[12] H. Liu and R. Setiono, "Feature Selection via Discretization", IEEE Transaction on Knowledge and Data Engineering, vol. 9, no. 4, **(1997)**, pp. 642-645.

Authors

Hongwei Wang, he was born in Jilin, China in 1980. He is studying at China University of Petroleum (Beijing) for a doctorate. Graduated from Daqing Petroleum Institute in 2005 and received a master's degree in engineering.

He is working at Bohai University, lecturer.

Main research include data mining, feature selection



Guohe Li, he was born in Zhangzhou, Fujian, China in 1965. He graduated in Applied Geophysics in 1988 and Computer Science in 1991 from China University of Petroleum, Beijing, China, obtaining the BSc degree and MSc degree respectively. He obtained the PhD degree in Computer Science from Beihang University, Beijing, China, in 2005.

He is a Professor of Computer Science in the School of Information Engineering at China University of Petroleum in Beijing, China. He worked as an Assistant Professor (1991-1998), an Associate Professor(1999-2005),a Professor(2006-) at China University of Petroleum in Beijing, China. He was a Visiting Scholar in the School of Information Technology and Electrical Engineering at University of Queensland (UQ) in Brisbane, Queensland, Australia in 2009.

Dr Guohe Li's major research interests and expertise include: Artificial Intelligence, Knowledge Discovery and Data Mining, Information Management Systems. He is a member of China Rough Set and Soft Computing (CRSSC) and reviewer of several journals of computer and information technology.

Xue Li, he was born in Chongqing, China in 1955. He graduated in Computer Software from Chongqing University, Chongqing, China 1982. He obtained the MSc degree in computer Science from University of Queensland 1990. He obtained a Ph.D degree in information systems from Queensland University of Technology in 1997.

He is an Associate Professor in the School of Information Technology and Electrical Engineering at University of Queensland

(UQ) in Brisbane, Queensland, Australia. He worked as a Lecturer at National University of Defense Technology (NUDT), Changsha, China (1982-1986). He worked as a lecturer in the Department of Computing at the Queensland University of Technology (QUT), Australia (1990-1998). From Jan. 1998 to Dec 2000, he worked as a senior lecturer in School of ISTM, University of New South Wales (UNSW), Sydney, Australia.

Dr Xue Li's major areas of research interests and expertise include: Data Mining, Multimedia Data Security, Database Systems, and Intelligent Web Information Systems. He is a member of ACM, IEEE, and SIGKDD.

