A New Evaluation Function for Entropy-Based Feature Selection from Incomplete Data

Wenhao Shu[1], Hong Shen[2,3], Yingpeng Sang[1], Yidong Li[1], and Jun Wu[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2] School of Information Science and Technology, Sun Yat-sen University, China
[3] School of Computer Science, University of Adelaide, Australia
11112084@bjtu.edu.cn, hongsh01@gmail.com

Abstract. In data mining and knowledge discovery, evaluation functions for evaluating the quality of features have great influence on the outputs of feature selection algorithms. However, in the existing entropy-based feature selection algorithms from incomplete data, evaluation functions are often inadequately computed as a result of two drawbacks. One is that the existing evaluation functions have not taken into consideration the differences of discernibility abilities of features. The other is that in the feature selection algorithms of forward greedy search, if the feature with the same entropy value is not only one, the arbitrary selection may affect the classification performance. This paper introduces a new evaluation function to overcome the drawbacks. A main advantage of the proposed evaluation function is that the granularity of classification is considered in the evaluation computations for candidate features. Based on the new evaluation function, an entropy-based feature selection algorithm from incomplete data is developed. Experimental results show that the proposed evaluation function is more effective than the existing evaluation functions in terms of classification accuracy.

Keywords: Evaluation function, Conditional entropy, Feature selection, Rough sets, Incomplete data.

1 Introduction

Feature reduction has been shown effective in dealing with high-dimensional data for efficient data mining, which refers to the study of methods for reducing the number of dimensions describing data [4, 10]. Its general purpose is to select relevant features to represent data and reduce computational cost, without deteriorating discriminative capability. It can bring many potential benefits: alleviating the curse of dimensionality, speeding up the learning process, and improving the generalization capability of a learning model. Many feature reduction algorithms have been developed at present. In general, they can be broadly classified into two categories: feature extraction and feature selection [5]. Feature extraction constructs new features with a linear or nonlinear transformation by projecting the original feature space to a lower dimensional one. Unlike feature

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 98–109, 2014.

extraction methods, feature selection methods preserve the original meaning of the features after reduction, which can be broadly categorized into wrapper [1] and filter [7, 9] methods. The wrapper method uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features. One drawback of the wrapper method, however, is that it is very expensive to run for data with numbers of features. The filter method separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with a feature selection algorithm. It relies on many feature measures such as distance [3], consistency [11], correlation [2] and so on. Much attention has been paid to filter feature selection.

Generally speaking, filter feature selection methods work under the framework consisting of four components [4]: subset generation, evaluation, stopping criterion and result validation. The main difference among various feature selection algorithms lies in how to evaluate the candidate features. Obviously, evaluation functions have great influence on outputs of feature selection algorithms. Rough set theory offers a formal methodology for filter feature selection. The main advantage of rough set theory is that no additional information about the data is required for data analysis such as thresholds or expert knowledge on a particular domain. It provides a mathematical tool to handle uncertainty in many data analysis tasks [6, 13]. The feature subset obtained by rough set-based feature selection is called a reduct. The features in the reduct are not only strongly relevant to the classification task, but also no redundant with each other, which keep consistency with the objective of feature selection.

It is clear that the feature selection work in classical rough set theory is based on complete data. However, in many real-world applications, it may happen that some feature values are missing because of many factors such as noise in data, prediction capability [12, 13, 15]. Here we briefly review the state of the art about feature selection algorithms from incomplete data. Sun et al. [12] introduced rough entropy to evaluate the roughness of knowledge in incomplete data, and developed a rough entropy-based feature selection algorithm. Slezak [14] proposed an algorithm based on information entropy to compute a reduct. As the uncertainty measure, conditional entropy, is one key issue in rough set theory, Dai et al. [15] proposed conditional entropy for incomplete data, and studied the application of feature selection based on conditional entropy. Evaluation functions, used to evaluate the quality of features, have great influence on the outputs of feature selection algorithms. However, there are some drawbacks in the existing evaluation functions. On the one hand, the existing evaluation functions only consider the differences of entropy values' variation, but there exists the differences of discernibility abilities for candidate features. As much as we know, the existing research work has not considered this aspect. Even if there are multiple features leading to the same entropy values, we can still compare the discernibility power of the features according to the granularity measure. On the other hand, for the forward greedy search, if the feature with the same entropy values is not only one, we often arbitrarily choose one of them, but the arbitrariness may affect the classification performance. Therefore, the main contribution of this paper is to present a new evaluation function to overcome the above stated drawbacks.

This paper is organized as follows. In Section 2, we review some basic concepts from the theory of rough sets. In Section 3, a simple example is firstly given to illustrate the drawbacks of existing evaluation functions, and then a new evaluation function together with an entropy-based feature selection algorithm are presented. In Section 4, comparison experiments are made to show the validity of the proposed evaluation function. Finally, the conclusions are presented in Section 5.

2 Preliminaries

Data sets are usually given as the form of tables, we call a data table as an information system, formulated as $IS =< U, A, V, f >$, where U is a set of nonempty and finite objects, called the universe; A is the set of features characterizing the objects; V is the union of feature domains, i.e., $V = \cup_{a\in A} V_a$, where V_a is the value set of feature a, called the domain of a; and $f : U \times A \to V$ is an information function, which assigns feature values to objects such as $\forall a \in A$, $x \in U$, and $f(x, a) \in V_a$, where $f(x, a)$ denotes the value of feature a for object x. If the feature set is divided into condition feature set C and decision feature set D, the information system is called a decision system. If there exist $x \in U$ and $a \in A$ such that $f(x, a)$ is equal to a missing value (a null or unknown value, denoted as "*"), i.e., $* \in V_a$, then the information system is an incomplete information system (IIS). If $* \notin V_D$ but $* \in V_C$, then the decision system is an incomplete decision system (IDS).

Given a complete information system $CIS =< U, A, V, f >$, for $\forall B \subseteq A$, the equivalence relation generated by B is defined by $IND(B) = \{(x, y)|\forall a \in B, f(x, a) = f(y, a)\}$. The family of all equivalence classes of $IND(B)$ is denoted as $U/IND(B)$. An equivalence class of $IND(B)$ containing x is denoted by $[x]_B$. Since there are missing values for some objects, the equivalence relation $IND(B)$ is not suitable for incomplete information systems.

Given an incomplete information system $IIS =< U, A, V, f >$, for $\forall B \subseteq A$, a tolerance relation between objects that are possibly indiscernible in terms of B is defined by $TR(B) = \{(x, y)|\forall a \in B, f(x, a) = f(y, a) \vee f(x, a) = * \vee f(y, a) = *\}$. It can be easily shown that $TR(B) = \cap_{a\in B} TR(\{a\})$. The tolerance class of object x with reference to a feature set B is denoted as $T_B(x) = \{y|(x, y) \in TR(B)\}$. Let $U/TR(B)$ denote the family set $\{T_B(x)|x \in U\}$, which is the classification induced by B. For $X \subseteq U$, the lower and upper approximation of X with respect to B can be defined as $\underline{B}(X) = \{x \in U|T_B(x) \subseteq X\}$ and $\overline{B}(X) = \{x \in U|T_B(x) \cap X \neq \text{ÃŸ}\}$. The lower approximation is called the positive region, that is $POS_B(X) = \underline{B}(X)$. X is called $B-$definable iff $\overline{B}(X) = \underline{B}(X)$. Otherwise, $\overline{B}(X) \neq \underline{B}(X)$ and X is rough.

Given an incomplete decision system $IDS =< U, C \cup D, V, f >$, for $\forall B \subseteq C$, the objects are partitioned into n mutually exclusion crisp subsets $U/IND(D) = \{D_1, D_2, \cdots, D_n\}$ by the decision features D. The lower and upper approximations with respect to B of D are defined as $\underline{B}(D) = \{\underline{B}(D_1), \underline{B}(D_2), \cdots, \underline{B}(D_n)\}$

and $\overline{B}(D) = \{\overline{B}(D_1), \overline{B}(D_2), \cdots, \overline{B}(D_n)\}$. Denoted by $POS_B(D) = \bigcup_{i=1}^{n} \underline{B}(D_i)$, which is called the positive region of D with respect to B in the IDS. The lower approximation is a description of the domain objects which are known with absolute certainty to belong to the decision classes.

3 An Evaluation Function for Entropy-Based Feature Selection

In this section, a simple example is firstly given to illustrate the drawbacks of existing evaluation functions, and then a new evaluation function together with a entropy-based feature selection algorithm are presented.

The conditional entropy of Definition 1 can be used as a reasonable information measure in incomplete decision tables[15], and it is quite representative among other entropies. Correspondingly, the evaluation function in terms of conditional entropy is also defined.

Definition 1. Let $IDS =< U, C \cup D, V, f >$be an incomplete decision table, $U = \{x_1, x_2, \ldots, x_n\}$, for $B \subseteq C$, the classification induced by B is $U/TR(B) = \{T_B(x_1),\ T_B(x_2), \ldots, T_B(x_n)\}$, and $U/IND(D) = \{D_1, D_2, \cdots, D_m\}$ is a partition on decision attribute set D. The conditional entropy of D with respect to B is defined by

$$H(D|B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|T_B(x_i) \cap D_j|}{|U|} log \frac{|T_B(x_i) \cap D_j|}{|T_B(x_i)|}.$$

Definition 2. Given an incomplete decision table $IDS =< U, C \cup D, V, f >$, suppose $B \subseteq C$ is the selected feature subset, and $a \in C - B$ is a candidate feature. Then the evaluation function of candidate feature a is defined as $e(a) = H(D|B) - H(D|B \cup \{a\})$.

From Definition 2, the existing evaluation function can be used to evaluate the importance of features. The smaller the evaluation value is, the more important the feature will be. However, the drawbacks of above evaluation function can be explained with reference to the following example.

Example. Suppose there is an incomplete decision table $IDS =< U, C \cup D, V, f >$, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ and $C = \{c_1, c_2, c_3, c_4\}$. In the feature selection process, Definition 2 is applied to compute the evaluation values of features. By computing, the descending sequence of four candidate features is listed as follows: $e(c_1) > e(c_2) > e(c_3) = e(c_4)$. Obviously, the features with the minimum evaluation value are c_3 and c_4. By direct computation the classifications induced by two features, $U/TR(c_3) = \{\{x_1, x_2, x_5, x_6\}, \{x_3, x_4, x_7, x_8\}\}$ and $U/TR(c_4) = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\},\ \{x_7, x_8\}\}$, obviously, the discernibility abilities of them are different, feature c_3 can describe the stronger discernibility power than c_4. However, Definition 2 does not take into consideration this difference. Thus the evaluation function given by Definition 2 is inadequately computed as a result of this aspect.

On the other hand, in the feature selection algorithm of forward greedy search, due to $e(c_3) = e(c_4)$, we can select one feature arbitrarily. Consequently, feature c_3 or c_4 are chosen to the selected feature subset. The arbitrariness can surely not guarantee a selected feature subset is a reduct. Suppose that the selected feature subset containing feature c_3 and c_2 exhibit the best performance, but we obtain the final feature subset is $\{c_4, c_2\}$ due to the arbitrary selection. Obviously, this result may affect the classification performance. Therefore, we give a new evaluation function from a reasonable perspective to improve the above mentioned problems.

Definition 3. Given an incomplete decision table $IDS =< U, C \cup D, V, f >$, suppose $B \subseteq C$ is the selected feature subset, and $a \in C - B$ is a candidate feature, the classification induced by a consists of tolerance class $A_i (1 \leq i \leq k)$. Then a new evaluation function of candidate feature a is defined as $f(a) = e(a) + g(a)$, where $g(a) = \frac{1}{|U|^2} \sum_{i=1}^{k} |A_i|^2$, which is the granularity measure of feature a.

Theorem 1. *Given an incomplete decision table $IDS =< U, C \cup D, V, f >$, suppose $B \subseteq C$ is the selected feature subset, for $\forall a, b \in C - B$, there is $f(a \cup b) < f(a)$ or $f(a \cup b) < f(b)$.*

Proof. Suppose the classification induced by a consists of tolerance classes $A_i (1 \leq i \leq k)$, and the classification induced by $a \cup b$ consists of tolerance classes $B_j (1 \leq j \leq l)$, by Definition 2 and the definition of conditional entropy, it is obvious that $e(a \cup b) < e(a)$. Since $a \subseteq a \cup b$, according to the definition of tolerance class, there is $|B_j| < |A_i|$, obviously, it holds that $\sum_{j=1}^{l} |B_j|^2 < \sum_{i=1}^{k} |A_i|^2$, thus $g(a \cup b) < g(a)$. Therefore, $f(a \cup b) < f(a)$. In the same way, it can proof that $f(a \cup b) < f(b)$. □

Theorem 1 shows the rationality of the new evaluation function, which states the uncertainty decreases when the available knowledge increases. Obviously, the granularity measure can represent discernibility ability of candidate feature a, the smaller $g(a)$ is, the stronger its discernibility ability. Through comparison, the selection of survival features can be achieved. From above example, there is $g(c_3) > g(c_4)$, thus it also holds that $f(c_3) > f(c_4)$, the discernibility ability of candidate feature c_4 is stronger than that of feature c_3. Therefore, the survival feature is c_4. It is obvious that the new evaluation function is more reasonable. Combine the new evaluation function into feature selection, a selected feature subset (called reduct) can be characterized by the following statement.

Definition 4. Given an incomplete decision table $IDS =< U, C \cup D, V, f >$, a selected feature subset $B \subseteq C$ is called a reduct of the IDS if and only if $H(D|B) = H(D|C)$, and for$\forall B' \subset B$, $H(D|B') \neq H(D|C)$.

In this definition, the first one indicates that the selected feature subset preserves the same information measure as the whole set of features; the second one guarantees that all of the features are indispensable, i.e., there is not any redundant feature in the reduct.

In the following, we combine the proposed evaluation function with forward greedy search to construct the feature selection algorithm.

Algorithm 1. Entropy-based Feature Selection Algorithm from Incomplete Data

```
Input: An incomplete decision table $IDS =< U, C \cup D, V, f >$;
Output: A feature subset $Red$.
Begin
Initialize $Red = \emptyset$;
For each $c \in C$ do
    compute $H(D|C - \{c\}) - H(D|C)$;
    if $H(D|C - \{c\}) - H(D|C) > 0$, then $Red = Red \cup \{c\}$;
End for
While $H(D|Red) \neq H(D|C)$ do
    compute $f(c)$ for all $c \in C - Red$;
    choose the feature $c_k$ that minimizes $f(c)$, and let $Red = Red \cup \{c_k\}$,
    $C = C - \{c_k\}$;
End while
For each $c \in Red$ do
    compute $H(D|Red) - H(D|Red - \{c\})$;
    if $H(D|Red) - H(D|Red - \{c\}) = 0$, then $Red = Red - \{c\}$;
End for
Return $Red$.
End
```

The algorithm begins with an empty subset *Red*, and adds some indispensable features to *Red* gradually. Then select the features with the minimal value by the new evaluation function into *Red* each loop until satisfying the stopping condition. Finally, a redundancy-removing step is carried out to avoid the redundancy in the selection result. The feature subset selected by this algorithm obtains the same information as the original feature set from incomplete data.

4 Experimental Analysis

In order to test the validity of the new proposed evaluation function, we conduct some experiments on a PC with Windows 7, Intel (R) Core(TM) Duo CPU 2.93 GHz and 4GB memory. Algorithms are coded in C++ and the software being used is Microsoft Visual 2008. The objective of the following experiments is to show the effectiveness of feature selection algorithm based on the new evaluation function. We perform the experiments on six real UCI data sets, which are downloaded from UCI Repository of machine learning databases in [16]. The characteristics of six data sets are described in Table 1. For the complete data sets, we randomly change 5% of the known features values from each original data set into missing values to create incomplete data sets. For the numerical features, we use the data tool Rosetta (http://www.lcb.uu.se/tools/rosetta/index.php) to discretize them.

Table 1. A description of six data sets

Data sets	Objects	Features	Classes
Hepatitis	155	19	2
Soybean-large	307	35	19
Synthetic	600	60	6
Cardiotocography	2126	21	3
Ticdate 2000	5822	85	2
Mushroom	8124	22	2

In what follows, we first make a comparative study on the feature selection algorithms in terms of feature subset size. The results are shown in Table 2 in which PFS represents the proposed feature selection algorithm, EFS represents the feature selection algorithm constructed in [15] and LFS denotes the lower approximation-based feature selection algorithm in [13]. Note that PFS selects candidate features by Definition 4, while EFS finds candidate features by Definition 2. The main difference between PFS and EFS is the evaluation function.

Table 2. Comparison of feature subset size by Algorithms PFS, EFS and LFS

Data sets	Original feature set size	Feature subset size		
		PFS	EFS	LFS
Hepatitis	19	12	14	14
Soybean-large	35	9	11	10
Synthetic	60	13	13	16
Cardiotocography	21	12	13	12
Ticdate 2000	85	24	24	24
Mushroom	22	4	5	5

As shown in Table 2, we can observe that Algorithm PFS selects fewer features comparing with EFS and LFS in most data sets. For example, as data set Hepatitis, PFS selects 12 features, while both of EFS and LFS select 14 features. The reason can be attributed to that the total number of objects in the data sets keep invariant, the more objects can be discerned with the selected features by proposed evaluation function in PFS than that of EFS at certain iterations, such that fewer features needed to discern all the objects in the data sets by PFS.

And it does shows that there is a decrease in feature subset size between PFS and LFS, demonstrating that there is other information contained in the entropy other than that in the lower approximation. This phenomenon indicates that the proposed feature selection algorithm can reduce data dimensions effectively, thus it verifies the validity of new evaluation function.

We employ two classifiers NaiveBayes and J48 to evaluate the classification performance of the selected feature subset. Each data set is divided into two parts: one for training and the other for test. On the basis of the training data, we employ feature selection algorithms to reduce the data sets. By NaiveBayes and J48, the rules are extracted from the training set. Using the rules the test set is classified and the classification results are obtained. The average classification accuracies and standard deviation are acquired based on tenfold cross-validation shown in Tables 3 and 4, where Raw depicts the classification performance on data sets with the original features, and the average classification accuracies are expressed in percentage. The "Average(ACC)" row records the average classification accuracy of the three algorithms on six data sets.

Table 3. Comparison of classification accuracy for NaiveBayes Classifier

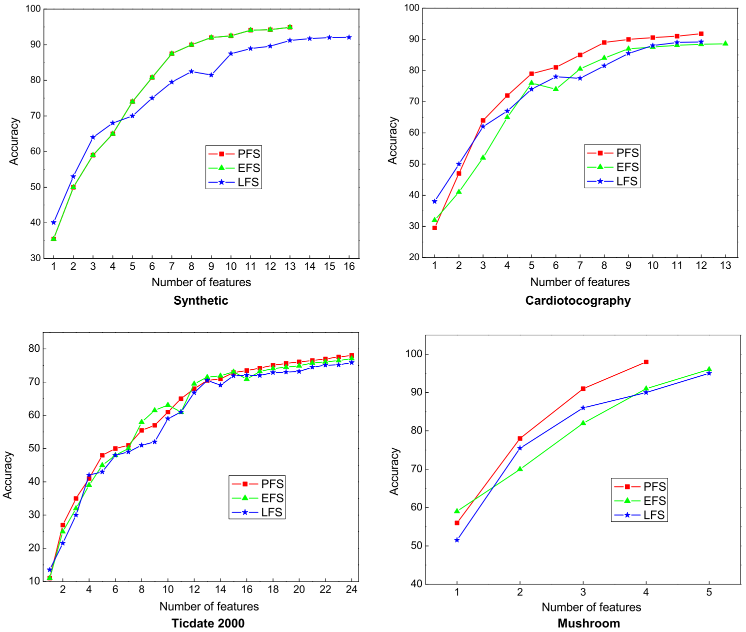
Data sets	NaiveBayes Classifier			
	Raw	PFS	EFS	LFS
Hepatitis	84.07±0.99	86.12±0.75	85.30±0.61	85.28±0.73
Soybean-large	91.43±1.07	92.50±1.12	90.89±1.20	90.11±1.54
Synthetic	95.58±2.20	94.97±1.93	94.97±1.93	92.06±1.87
Cardiotocography	89.79±0.61	91.85±0.40	88.56±0.76	89.23±0.31
Ticdate 2000	76.04±1.14	78.07±0.86	77.58±1.39	76.90±2.15
Mushroom	95.52±0.76	98.19±0.58	96.72±0.60	98.95±0.71
Average(ACC)	88.73	90.28	89.00	88.76

Table 4. Comparison of classification accuracy for J48 Classifier

Data sets	J48 Classifier			
	Raw	PFS	EFS	LFS
Hepatitis	79.35±1.16	84.60±1.09	82.32±1.25	80.81±1.74
Soybean-large	88.01±0.63	87.92±0.54	87.09±0.41	87.75±0.90
Synthetic	84.51±1.02	89.40±0.66	89.40±0.66	86.03±0.82
Cardiotocography	95.07±0.84	97.26±0.59	94.01±1.20	95.92±1.03
Ticdate 2000	79.55±0.91	81.70±1.37	79.34±1.44	82.15±1.58
Mushroom	100.00±0.0	100.00±0.0	100.00±0.0	100.00±0.0
Average (ACC)	87.74	90.15	88.69	88.78

The results shown in Tables 3 and 4 indicate that PFS produces the better classification performances after feature selection based on the new evaluation function than those of EFS and LFS as to NaiveBayes and J48. Regarding NaiveBayes, PFS is better than EFS on all the data sets other than data set Synthetic, and PFS also shows increases in classification accuracies comparing with LFS. As to J48, PFS outperforms EFS on four of six data sets; PFS outperforms LFS on most of the data sets. Considering the results between PFS and EFS, it can demonstrate the effectiveness of new evaluation function in feature selection. In addition, the three approaches improve the classification capability by selecting a small portion of the original features. From the experimental results, we can confirm that the proposed evaluation function leads to promising improvement on classification performance.

To further explain the reason why the classification performances are improved using the new evaluation function, we conduct the experiments on four large data sets using NaiveBayes classifier with Algorithms PFS, EFS and LFS. Fig.1 displays more detailed change trend of the three algorithms in classification accuracy with the number of selected features.



Synthetic

Cardiotocography

Ticdate 2000

Mushroom

Fig.1. Trends of accuracies by NaiveBayes with number of features

From Fig.1, the curves between PFS and EFS in the data set Synthetic are overlapping. The reason is that PFS and EFS select the same features, thus the classification accuracies are the same for selecting the same number of features. However, most points in the curves of PFS are higher than those of EFS and LFS in the data sets. Take data set Cardiotocography as an illustration, the classification accuracies of PFS are higher than those of EFS and LFS since the beginning of selecting three features. The underlying reason perhaps is that though the number of features is the same by PFS and EFS, the selected features are different, PFS employed the new evaluation function always find the candidate features that can discern more objects for classification learning, such that the classification performance is better than that of EFS. The similar situations can be found in two other data sets. Observing the curves, we can find that PFS can keep a steady increase in accuracy value, whereas EFS and LFS incur a fluctuant increase, even a decrease. This phenomenon may result from one possible reason that PFS has a redundancy-removing step, while EFS and LFS does not consider the redundant information between the selected features. It shows some dispensable features in the selected feature subset are superfluous, which deteriorate the classification performance.

Furthermore, we conduct the experiments on the four larger data sets using J48 classifier with the three algorithms. Fig.2 displays more detailed change trend of the three algorithms in classification accuracy with the number of selected features.

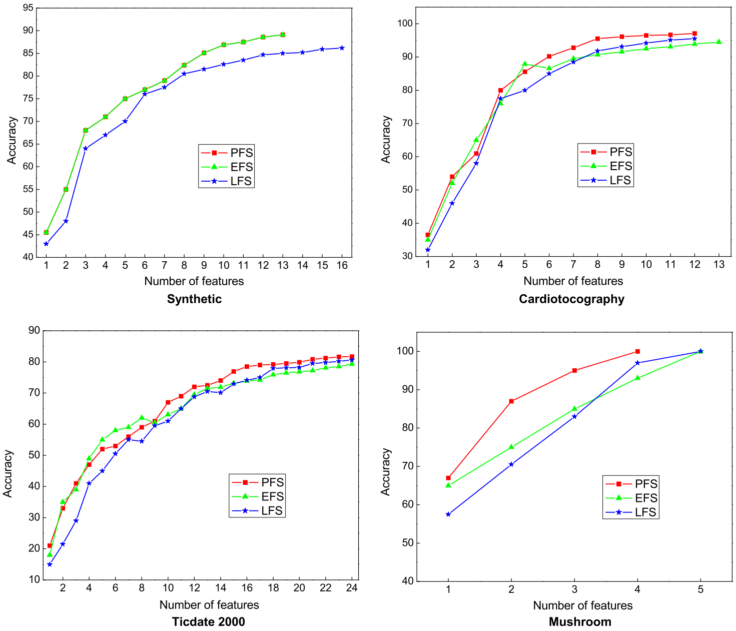


Fig.2. Trends of accuracies by J48 with number of features

As shown in Fig.2, the curves between PFS and EFS in the data set Synthetic are overlapping. However, one may observe that there are many points in the curves where the classification performance of PFS clearly surpasses those of EFS and LFS. We can see that, as data set Mushroom, when the selected feature number is two, the classification accuracy of PFS is higher than those of EFS and LFS. Though the same number of selected features, PFS can select the feature that discerns more objects for classification learning, correspondingly, the selected features are different, and the classification accuracy is higher than that of EFS. And comparing with LFS, PFS can find some other useful information contained in the entropy other than lower approximation, which would result in better classification performance. For the other three data sets, one may observe that the similar situations.

Based on the aforementioned experimental results, we can conclude that the new evaluation function gives an effective way to select satisfactory feature subset in the process of feature selection from incomplete data.

5 Conclusions

In this paper, we introduce a new evaluation function to overcome the drawbacks of existing evaluation functions. Based on the new evaluation function, we construct a conditional entropy-based feature selection algorithm with forward greedy search from incomplete data. The numerical experiments show the validity of the new evaluation function. Two main conclusions are drawn as follows. On the one hand, compared with the existing evaluation function, the new evaluation function reflects not only the conditional entropy values' variation, but also the discernibility ability of a candidate feature. Thus the new evaluation function is more reasonable than the existing evaluation function to describe the discernibility ability. On the other hand, in feature selection, even if there are more features with same importance in the conditional entropy, our feature selection algorithm can select one with the greatest classification ability, while the arbitrary selection in the existing feature selection algorithm may affect the classification performance. Therefore, the new evaluation function is more effective in the process of feature selection from incomplete data.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China (61170232), Fundamental Research Funds for the Central Universities (2012JBZ0 17), Independent research project of State Key Laboratory of Rail Traffic Control and Safety (RCS2012ZT011), Innovation Funds of Excellence Doctor of Beijing Jiaotong University (2014YJS040) and Research Initiative Grant of Sun Yat-sen University. The corresponding author is Hong Shen.

References

1. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
2. Qu, G., Hariri, S., Yousif, M.: A new dependency and correlation analysis for features. IEEE Transactions on Knowledge and Data Engineering 17(9), 1199–1207 (2005)

3. Liang, J., Yang, S., Winstanley, A.: Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition 41(5), 1429–1439 (2008)
4. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis 1(3), 131–156 (1997)
5. Steppe, J.M., Bauer, K.W., Rogers, S.K.: Integrated feature and architecture selection. IEEE Transactions on Neural Networks 7(4), 1007–1014 (1996)
6. Pawlak, Z., Skowron, A.: Rough sets and Boolean reasoning. Information Sciences 177(1), 41–73 (2007)
7. Xue, B., Cervante, L., et al.: A multi-objective particle swarm optimisation for filter-based feature selection in classification problems. Connection Science 24(2-3), 91–116 (2012)
8. Cervante, L., Xue, B., Shang, L., Zhang, M.J.: Binary particle swarm optimisation and rough set theory for dimension reduction in classification. In: IEEE Congress on Evolutionary Computation (CEC), pp. 2428–2435 (2013)
9. Sebban, M., Nock, R.: A hybrid filter / wrapper approach of feature selection using information theory. Pattern Recognition 35(4), 835–846 (2002)
10. Farahat, A.K., Ghodsi, A., Kamel, M.S.: An efficient greedy method for unsupervised feature selection. In: The 11th IEEE International Conference on Data Mining (ICDM), pp. 161–170 (2011)
11. Hu, Q.-H., Zhao, H., Xie, Z.-X., Yu, D.-R.: Consistency based attribute reduction. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 96–107. Springer, Heidelberg (2007)
12. Sun, L., Xu, J.C., Tian, Y.: Feature selection using rough entropy-based uncertainty measures in incomplete decision systems. Knowledge-Based Systems 36, 206–216 (2012)
13. Qian, Y.H., Liang, J.Y., Pedrycz, W., Dang, C.Y.: An efficient accelerator for attribute reduction from incomplete data in rough set framework. Pattern Recognition 44, 1658–1670 (2011)
14. Slezak, D.: Approximate entropy reducts. Fundamenta Informaticae 53, 365–390 (2002)
15. Dai, J.H., Wang, W.T., Xu, Q.: An uncertainty measure for incomplete decision tables and its applications. IEEE Transactions on Cybernetics 43(4), 1277–1289 (2013)
16. UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html