Adapt the mRMR Criterion for Unsupervised Feature Selection

Junling Xu

School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
jlxu@seu.edu.cn

Abstract. Feature selection is an important task in data analysis. mRMR is an equivalent form of the maximal statistical dependency criterion based on mutual information for first-order incremental supervised feature selection. This paper presents a novel feature selection criterion which can be considered as the unsupervised version of mRMR. The concepts of relevance and redundancy are both concerned in the feature selection criterion. The effectiveness of the new unsupervised feature selection criterion is confirmed by the theoretical proof. Experimental validation is also conducted on several popular data sets, and the results show that the new criterion can select features highly correlated with the latent class variable.

Keywords: Feature selection, Unsupervised feature selection, Mutual information.

1 Introduction

Data analysis often deals with large data sets containing not only a huge amount of instances but also a significant number of features. Some of the features are redundant, while some are irrelevant and noisy. Therefore, it is an important step to preprocess the data to remove the noisy and redundant features when analyzing high-dimensional data sets. This process is commonly termed as feature selection. Generally, feature selection methods can be categorized as supervised and unsupervised, based on the class information of the data. When class labels of the data are available, supervised feature selection can be used, otherwise, unsupervised feature selection is the right option. Feature selection methods can also be categorized as filter and wrapper, based on their dependence on the learning algorithm that will finally use the selected subset [1]. Filter methods are independent of the learning algorithm, whereas wrapper methods use the learning algorithm as the evaluation function.

The objective of unsupervised feature selection is to select important features which are representative and can characterize the main property of all the original features in the absence of class labels. There have been some works on it [2,3,4,5,6,7,8]. The algorithm described in [2] evaluates the clustering tendency of each feature by an entropy index, which is based on the observation that data

L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part II, LNCS 6441, pp. 111–121, 2010.

with clusters has very different point-to-point distance histogram from data without clusters. In [3], a maximum information compression index is used to measure feature similarity so that feature redundancy is detected. In [4], a forward orthogonal search algorithm by maximizing the overall dependency is proposed to detect significant features and select a subset of all the original features. However, all of these methods are designed to deal with numerical features. In [5], weights are assigned to different feature spaces for k-means clustering based on within-cluster and between-cluster matrices. Feature saliency is integrated in EM algorithm in [6] so that feature selection is performed simultaneously with clustering process. A wrapper criterion for clustering [7], which evaluates the quality of clusters using normalized cluster separability (for k-means) or normalized likelihood (for EM clustering), was proposed by Dy and Brodley. Recently, Li *et al.* adapted their scatter separability criterion to localized feature selection [8]. All of the above-mentioned four methods are all wrapper methods which are computationally very expensive and do not scale well to large datasets.

In order to design a unsupervised-filter feature selection method which can simultaneously deal with numerical and non-numerical features, this paper propose a novel feature selection criterion based on mutual information (UmRMR). The motivation for considering MI is derived from its capability to measure a general dependence between two features. Though mutual information has been used in supervised feature selection [9,10], to the best of our knowledge, it is the first time which has been used as an evaluation measure in unsupervised feature selection. The process of our feature selection is a sequential forward search which ranks features according to UmRMR. Experiments on several popular data sets show the effectiveness of the proposed method.

The rest of the paper is organized as follows. In the next section, we present the details of the proposed feature selection criterion, and reveal the relationship with mRMR. The experimental study, as well as the results, is described in Section 3. Section 4 concludes the paper and outlines directions for future work.

2 Unsupervised Feature Selection Criterion

This section briefly introduces the mRMR feature selection criterion; and then presents the unsupervised feature evaluation criterion UmRMR which can be considered as an unsupervised version of mRMR; finally, the relationships of UmRMR and mRMR are discussed.

2.1 The mRMR Criterion

The process of feature selection used here is a sequential forward search which ranks the features according to an evaluation measure. Assume that at step $m-1$ the set U of unselected features, and a feature subset S_{m-1}, consisting of $m-1$ features has been determined. How should the mth significant feature $\mathbf{y}_m$ be chosen? According to the "minimal-redundancy-maximal-relevance" (mRMR) criterion proposed in [10], the mth feature is selected based on the following formula:

$$\ell_m = arg \max_{1 \leq i \leq n, \mathbf{x}_i \in U} \{I(\mathbf{x}_i; \mathbf{c}) - \frac{1}{m-1} \sum_{\mathbf{x}_j \in S_{m-1}} I(\mathbf{x}_i; \mathbf{x}_j)\}. \tag{1}$$

The first item in (1), the maximal relevance (Max-Relevance) condition, tends to select the feature which has the largest dependency on the target class $\mathbf{c}$. It is likely that features selected according to Max-Relevance could have rich redundancy, i.e., the dependency among these features could be large. When two features highly depend on each other, the respective class-discriminative power would not change much if one of them was removed. Therefore, the minimal redundancy (Min-Redundancy) condition is added to select mutually exclusive features.

2.2 Unsupervised mRMR Criterion (UmRMR)

Since the class information is unavailable in unsupervised feature selection process, we need to give new definitions for relevance and redundancy. We now first define the relevance of a feature.

Definition 1 (Relevance). *The relevance of a feature $\boldsymbol{x}_i$ is its average mutual information to the whole feature set:*

$$Rel(\boldsymbol{x}_i) = \frac{1}{n} \sum_{j=1}^{n} I(\boldsymbol{x}_i; \boldsymbol{x}_j) = \frac{1}{n}(H(\boldsymbol{x}_i) + \sum_{1 \leq j \leq n, j \neq i} I(\boldsymbol{x}_i; \boldsymbol{x}_j)). \tag{2}$$

In the definition of the relevance of a feature, $H(\mathbf{x}_i)$ indicates the information content contained in feature $\mathbf{x}_i$: the larger $H(\mathbf{x}_i)$ is, the more information it can supply the learning algorithm; and $\sum_{1 \leq j \leq n, j \neq i} I(\mathbf{x}_i; \mathbf{x}_j)$ is the amount of information decreased from the information content contained in all the other features due to the knowledge of $\mathbf{x}_i$: the larger $\sum_{1 \leq j \leq n, j \neq i} I(\mathbf{x}_i; \mathbf{x}_j)$ is, the less new information other features can supply the learning algorithm; If we select feature $\mathbf{x}_i$ which has the maximal $Rel(\mathbf{x}_i)$, then it can lead to the loss of information to the least extent.

In order to define the redundancy of a feature, we assume that the $Rel(\mathbf{x}_i)$ of a feature $\mathbf{x}_i$ is proportional to its entropy $H(\mathbf{x}_i)$ (i.e. the relevance value provided by per information unit is a constant $\mathscr{C}_{\mathbf{x}_i}$ for a certain feature $\mathbf{x}_i$, while it may be various for different features). If a feature $\mathbf{x}_i$ in U is considered to be selected, then for any feature $\mathbf{y}_j$ in S_{m-1}, its conditional information content on $\mathbf{x}_i$ is $H(\mathbf{y}_j \mid \mathbf{x}_i)$. Obviously, the information supplied by $\mathbf{y}_j$ decreases due to the existence of $\mathbf{x}_i$, so does the relevance of $\mathbf{y}_j$. Based on our assumption, we have the definition of conditional relevance:

Definition 2 (Conditional Relevance). *The conditional relevance of feature $\boldsymbol{y}_j$ on feature $\boldsymbol{x}_i$ is*

$$Rel(\boldsymbol{y}_j \mid \boldsymbol{x}_i) = \frac{H(\boldsymbol{y}_j \mid \boldsymbol{x}_i)}{H(\boldsymbol{y}_j)} Rel(\boldsymbol{y}_j) = \mathscr{C}_{\boldsymbol{y}_j} H(\boldsymbol{y}_j \mid \boldsymbol{x}_i). \tag{3}$$

As can be seen from (3), the relevance of feature $\mathbf{y}_j$ decreases due to the selection of feature $\mathbf{x}_i$, as $H(\mathbf{y}_j \mid \mathbf{x}_i) \leq H(\mathbf{y}_j)$. So we can define the redundancy of $\mathbf{x}_i$ to $\mathbf{y}_j$ as follows:

Definition 3 (Redundancy). *The redundancy of feature $\boldsymbol{x}_i$ relative to feature $\boldsymbol{y}_j$ is given by*

$$Red(\boldsymbol{x}_i; \boldsymbol{y}_j) = Rel(\boldsymbol{y}_j) - Rel(\boldsymbol{y}_j \mid \boldsymbol{x}_i). \tag{4}$$

To select the mth significant feature $\mathbf{y}_m$, both the relevance of the feature to all the original feature set and the redundancy of the feature to the already-selected features should be considered, we define the feature selection criterion UmRMR as follows:

$$\ell_m = arg \max_{1 \leq i \leq n, \mathbf{x}_i \in U} \{Rel(\mathbf{x}_i) - \max_{\mathbf{y}_j \in S_{m-1}} Red(\mathbf{x}_i; \mathbf{y}_j)\} \tag{5}$$

or

$$\ell_m = arg \max_{1 \leq i \leq n, \mathbf{x}_i \in U} \{Rel(\mathbf{x}_i) - \frac{1}{m-1} \sum_{\mathbf{y}_j \in S_{m-1}} Red(\mathbf{x}_i; \mathbf{y}_j)\}. \tag{6}$$

The mth significant feature can be selected as $\mathbf{y}_m = \mathbf{x}_{\ell_m}$, which decreases the uncertainty about other features with a higher percentage, compared with other single feature in the feature set U, and brings little redundant information.

2.3 Relationships of UmRMR and mRMR

The experiments in [10] showed that the mRMR incremental selection scheme provides a better way to maximize the dependency of the selected features and the target class. But for unsupervised feature selection, the class distribution underlying the data sets is unknown, can we still maximize the dependency of the selected features and the underlying target class without the information about the class? After studying the relationship of UmRMR and mRMR, we found that UmRMR can solve this problem to some extent. First we will show the relationship between the two definitions of relevance.

Proposition 1. *The relevance of $\boldsymbol{x}_i$ in UmRMR is a lower bound of the relevance condition of mRMR under the naive bayes assumption, i.e. $Rel(\boldsymbol{x}_i) \leq I(\boldsymbol{x}_i; \boldsymbol{c})$.*

Proof: Let $\mathbf{c}$ be the target class of instances, then it can be viewed as a function of features $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Because $\mathbf{x}_1, \cdots, \mathbf{x}_n$ can be seen as random variables, so is $\mathbf{c}$. All of these variables here are assumed to be of discrete type, otherwise, they will be discretized first. We have

$$\begin{aligned} H(\mathbf{x}_i|\mathbf{c}) &= \sum_c \sum_{x_i} P(x_i, c) log \frac{1}{P(x_i|c)} \\ &= \sum_c \sum_{x_i} \sum_{x_j} P(x_i, x_j) \frac{P(x_i, x_j, c)}{P(x_i, x_j)} log \frac{1}{P(x_i|c)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x_i}\sum_{x_j} P(x_i, x_j) \sum_{c} \frac{P(x_i, x_j, c)}{P(x_i, x_j)} log \frac{1}{P(x_i|c)} \\
&\le \sum_{x_i}\sum_{x_j} P(x_i, x_j) log[\sum_{c} \frac{P(x_i, x_j, c)}{P(x_i, x_j)} \frac{1}{P(x_i|c)}] && (7) \\
&= \sum_{x_i}\sum_{x_j} P(x_i, x_j) log[\frac{1}{P(x_i, x_j)} \sum_{c} \frac{P(x_i, x_j, c)}{P(x_i|c)}] \\
&= \sum_{x_i}\sum_{x_j} P(x_i, x_j) log[\frac{1}{P(x_i|x_j)} \frac{1}{P(x_j)} \sum_{c} \frac{P(x_i, x_j, c)}{P(x_i|c)}] \\
&= \sum_{x_i}\sum_{x_j} P(x_i, x_j) [log \frac{1}{P(x_i|x_j)} + log \sum_{c} \frac{P(x_i, x_j, c)}{P(x_i|c) P(x_j)}] \\
&\le H(\mathbf{x}_i|\mathbf{x}_j) + log \sum_{x_i}\sum_{x_j}\sum_{c} \frac{P(x_i, x_j) P(x_i, x_j, c)}{p(x_j) P(x_i|c)}. && (8)
\end{aligned}
$$

(7) and (8) are attained due to the using of the Jensen's Inequality. The naive bayes assumption supposes that the features are not independent but conditionally independent given the value of **c**. Under the naive bayes assumption, $P(x_i, x_j, c) = P(c)P(x_i|c)P(x_j|c)$, substituting for $P(x_i, x_j, c)$ in (8) we get

$$
\begin{aligned}
&H(\mathbf{x}_i|\mathbf{c}) \\
&\le H(\mathbf{x}_i|\mathbf{x}_j) + log \sum_{x_i}\sum_{x_j}\sum_{c} \frac{P(x_i, x_j) P(c) P(x_i|c) P(x_j|c)}{p(x_j) P(x_i|c)} \\
&= H(\mathbf{x}_i|\mathbf{x}_j) + log \sum_{x_i}\sum_{x_j} P(x_i, x_j) \sum_{c} \frac{P(x_j, c)}{p(x_j)} \\
&= H(\mathbf{x}_i|\mathbf{x}_j) + log \sum_{x_i}\sum_{x_j} P(x_i, x_j) \\
&= H(\mathbf{x}_i|\mathbf{x}_j). && (9)
\end{aligned}
$$

According to (2) and (9),

$$
\begin{aligned}
Rel(\mathbf{x}_i) &= \frac{1}{n} \sum_{j=1}^{n} I(\mathbf{x}_i; \mathbf{x}_j) \\
&= \frac{1}{n} \sum_{j=1}^{n} (H(\mathbf{x}_i) - H(\mathbf{x}_i|\mathbf{x}_j)) \\
&\le \frac{1}{n} \sum_{j=1}^{n} (H(\mathbf{x}_i) - H(\mathbf{x}_i|\mathbf{c})) \\
&= I(\mathbf{x}_i; \mathbf{c}). && (10)
\end{aligned}
$$

Proposition 1 means that maximizing $Rel(\mathbf{x}_i)$ can increase the value of $I(\mathbf{x}_i; \mathbf{c})$ to some extent.

The redundancy of mRMR is defined as the dependance of the the feature to be selected and the already selected features. From (4), It seems that our definition of redundance only considered the relevance and conditional relevance of the selected features, which is different from the definition in the mRMR criterion. Now we will show the relationship between the two definitions of redundancy.

Proposition 2. *The redundancy of $\boldsymbol{x}_i$ relative to $\boldsymbol{y}_j$ in UmRMR equals the redundancy in mRMR times a constant.*

Proof

$$\begin{aligned} Red(\mathbf{x}_i;\mathbf{y}_j) &= Rel(\mathbf{y}_j) - Rel(\mathbf{y}_j \mid \mathbf{x}_i) \\ &= Rel(\mathbf{y}_j) - \frac{H(\mathbf{y}_j \mid \mathbf{x}_i)}{H(\mathbf{y}_j)} Rel(\mathbf{y}_j) \\ &= \frac{I(\mathbf{x}_i, \mathbf{y}_j)}{H(\mathbf{y}_j)} Rel(\mathbf{y}_j) \\ &= \mathscr{C}_{\mathbf{y}_j} I(\mathbf{x}_i, \mathbf{y}_j). \end{aligned} \tag{11}$$

3 Experimental Results

In this section, we first test whether UmRMR can select features highly correlated with the latent class. Then we test the effectiveness of UmRMR on improving the performance of the learning algorithm. Finally we compare UmRMR with other two unsupervised feature selection criteria. The process of feature selection used here is a sequential forward search which ranks the features according to UmRMR, so we call this algorithm as SFS-UmRMR. Data sets used here are all taken from the UCI machine learning repository [11], they originally are either with discrete features or continuous features. For continuous features they were discretized by supervised method provided in [12] before the feature selection process.

3.1 Can SFS-UmRMR Select Features Highly Correlated with the Latent Classes?

Here we compare the orderly list of features produced by SFS-UmRMR and IG (information gain), a supervised feature selection method which is widely employed in machine learning. For each of the seven data sets shown in Table 1 (ecoli, iris, lymph, dermatology, breast-w, spambase and haberman), two orderly list of features were generated as shown in Table 2. The **bold** numbers are features with the same order in the two lists or features occurred in both lists within the top given number of features. It can be seen from Table 2 that important features certificated by IG can often be ranked in front by SFS-UmRMR, though the results attained by SFS-UmRMR did not using the class information as IG (except in the process of discretization).

Table 1. List of data sets

Data Set	Number of Features	Number of Instances	Number of Classes
ecoli	7	336	8
iris	4	150	2
lymph	18	148	4
dermatology	34	366	6
breast-w	9	699	2
spambase	57	4601	2
haberman	3	306	2

Table 2. Feature ranking by IG and SFS-UmRMR

Data Set	Ranking by IG	Ranking by SFS-UmRMR
ecoli	{**6,7,1,2,5,3,4**}	{**6,7,1,2,5,3,4**}
iris	{**3,4,1,2**}	{**3,4,1,2**}
lymph	{**13**,*18*,**15,14**,*2*,**10**...}	{**14,13,10**,*12*,**15**,*5*...}
dermatology	{**21,20,22**,*33*,*29*,**27**...}	{**20,27,21**,*16*,**22**,*9*...}
breast-w	{**2,3,6,7,5**...}	{**2,7,3,5,6**...}
spambase	{**52,53,56**,7,**21**...}	{57,**56,53,21,52**...}
haberman	{**3,2,1**}	{**3,2,1**}

3.2 How Effective Is SFS-UmRMR?

To demonstrate the efficacy of our algorithm from another view, we test SFS-UmRMR on more data sets to inspect its effectiveness. Since data sets, of which the features are all non-numerical are relatively few, the data sets used here originally are either with discrete features or continuous features. For continuous features they are discretized by supervised method provided in [12] before the feature selection process. Eight data sets are considered here (the information about the data sets is shown in Table 3). The desired number of features is not given, so the output of SFS-UmRMR is an orderly list of features. A wrapper method with k-nearest-neighbor (k-NN) algorithm is then used to seek the minimal feature subset, which can attain the classification accuracy provided by the complete data, and the optimal feature subset, which can attain the best classification accuracy. The classification accuracy is calculated by performing the 10-fold cross-validation procedure, and then the average classification accuracy of 10 runs of the k-NN algorithm is calculated. Features are selected one by one to form a subset according to their order in the output of SFS-UmRMR. The value of k, in the k-NN rule, is chosen by performing many experiments for different values of k, where $1 \leq k \leq \sqrt{N_{tr}}$ and N_{tr} is the number of the samples in the training set, and k is chosen as the one that gives the best classification performance.

A minimal feature subset and an optimal feature subset for each of the eight data sets including ionosphere, zoo, sponge, sonar, glass, arrhythmia, lung-cancer, and vote are selected. The numbers of features in the minimal subsets and in the optimal subsets for the eight data sets are {10, 12, 17, 22, 5, 2, 3, 2} and {10, 12, 32, 54, 6, 48, 5, 7} respectively. A comparison between the classification accuracy based on the complete data and the reduced data for the eight data sets is reported in Table 4. It can be seen that the classification accuracy based on the selected subsets outperformes those based on the complete data. This means that the selected feature subsets are representative and informative and thus can be used to replace the complete data for pattern classification.

Table 3. List of data sets

Data Set	Number of Features	Number of Instances	Number of Classes
ionosphere	33	351	2
zoo	17	101	7
sponge	45	76	3
sonar	60	208	2
glass	9	214	7
arrhythmia	279	452	16
lung-cancer	56	32	2
vote	16	435	2

Table 4. Classification accuracy over the complete data and reduced data

	No. of Features			Accuracy(%)		
Dataset	C	M	O	C	M	O
ionosphere	33	10	10	89.77±0.71{2}	90.57±1.14{2}	90.57±1.14{2}
zoo	17	12	12	96.14±0.50{1}	98.02±0.00{1}	98.02±0.00{1}
sponge	45	17	32	92.50±0.66{3}	92.63±0.66{2}	93.68±1.32{2}
sonar	60	22	54	86.44±1.44{1}	86.88±1.92{1}	88.08±2.64{2}
glass	9	5	6	70.00±2.10{3}	75.75±2.34{1}	77.57±0.94{1}
arrhythmia	279	2	48	59.27±0.55{6}	59.60±1.33{4}	67.61±0.55{3}
lung-cancer	56	3	5	79.69±1.56{3}	85.00±1.56{1}	88.44±1.56{2}
vote	16	2	7	93.15±0.46{4}	95.17±0.00{1}	95.59±0.92{1}

C/M/O: Complete/Minimal/Optimal Data. {}: the value of k used in k-NN rule

3.3 Comparison with Other Feature Selection Methods

We compare the performance of SFS-UmRMR with two unsupervised feature selection methods ENTROPY [2] and FOS-MOD [4]. Now, we first give a brief introduction of the two feature selection methods.

ENTROPY (entropy-based ranking) is proposed by Dash and Liu in [2]. The entropy is defined as the equation: $E(\mathbf{x}) = -\sum_{i=1}^{N}\sum_{j=1}^{N}(S_{i,j} \times logS_{i,j}) + (1 - S_{i,j}) \times log(1 - S_{i,j})$, where $S_{i,j}$ is the similarity value between the ith and the jth instances. $S_{i,j}$ is defined as the equation: $S_{i,j} = e^{\alpha \times dist_{i,j}}, \alpha = -ln(0.5)/\overline{dist}$, where $dist_{i,j}$ is the distance between the ith and the jth instances after the feature $\mathbf{x}$ is removed, $\overline{dist}$ is the average distance among the instances after the feature $\mathbf{x}$ is removed. Features are ranked in an ascending order according to the value of E.

FOS-MOD is an unsupervised forward orthogonal search algorithm. In this method, features are selected in a stepwise way, one at a time, by estimating the capability of each specified candidate feature subset to represent the overall features in the measurement space. The dependency between features is measured by a squared correlation function. The squared-correlation coefficient between feature $\mathbf{x}$ and $\mathbf{y}$ is $sc(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^2/[(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})]$. The concept of relevance in FOS-MOD is very similar to (2), where mutual information is replaced with squared-correlation coefficient. The concept of redundancy in FOS-MOD is implicity considered by an orthogonalization process, that is, the relevance of the mth significant feature is computed after it is orthogonalized with the $m - 1$ previous selected features.

Table 5. List of data sets

Data Set	Number of Features	Number of Instances	Number of Classes
liver-disorders	6	345	2
glass	9	214	7
segment	19	2310	7
sonar	60	208	2
vehicle	18	846	4
ionosphere	33	351	2

The six data sets used here are all with continuous features (the information about the data sets is shown in Table 5). For SFS-UmRMR, data sets are discretized with simple equal-width binning where the number of bins is chosen automatically by maximizing the likelihood via leave-one-out cross-validation before the feature selection process. First, all of the three feature selection methods are applied on each data set without giving the desired number of features, and thus we can attain three orderly lists of features for each data set; then features are selected one by one to form a subset according to their order in the output of each algorithm. k-NN algorithm is used to attain the classification accuracy with different sizes of the reduced feature subset for each data set, and the best classification performance for a certain k is chosen. A comparison in terms of the k-NN classification accuracy for different sizes of the reduced feature subset is shown in Fig. 1.

As can be seen from Fig. 1, SFS-UmRMR outperforms the other two methods almost on all of the six data sets. SFS-UmRMR can attain as good performance as the complete data when the feature number is relatively small, while the other two methods seem to prefer more features. Noticeably, the three classification accuracy curves have distinct tendencies. For SFS-UmRMR, as the feature number increases, the accuracy constantly increases and then converges at some point or declines. On the contrary, the accuracies for ENTROPY and FOS-MOD almost constantly increase as more and more features are used, indicating that redundant features or even irrelevant features are thought to be important by the two methods.

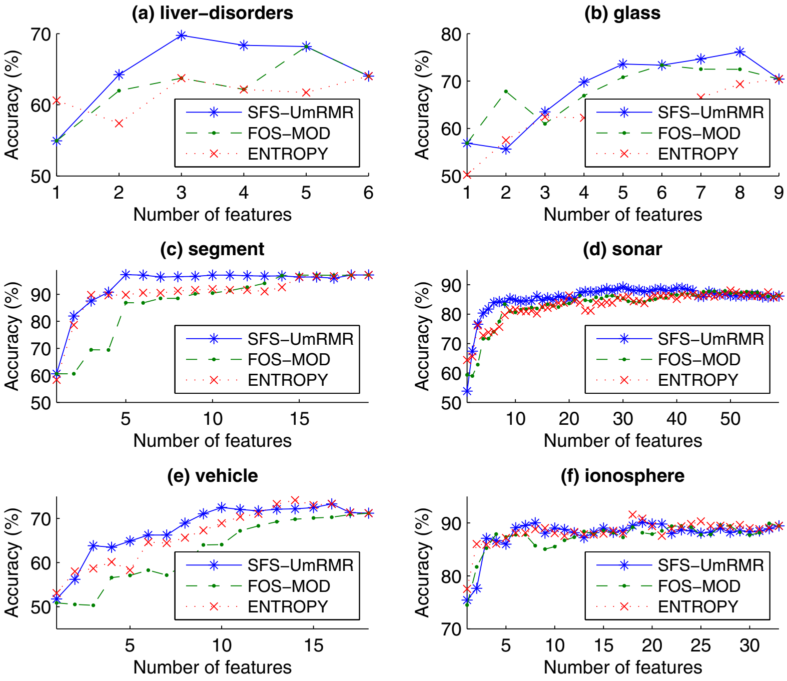

Fig. 1. Performance comparison on several data sets

4 Conclusion

In cases of the class information of the data set is unavailable or lost, unsupervised feature selection is a hard work. In this paper, we have proposed a novel unsupervised feature selection criterion UmRMR based on mutual information where the concepts relevance and redundancy of features are both defined from a standpoint of information theory. We also discussed the relationship between UmRMR with the famous mRMR criterion, which can be seen as a theoretical proof for the effectiveness of UmRMR. Experimental result also conformed that UmRMR can select features highly correlated with the latent class.

Unlike the method proposed in [4] which assumes a linear relationship exists between sample features, UmRMR can deal with general relationship between features. In many cases, where features are not linked by linear relationship, UmRMR can be competent for the feature selection task. In the future, we will compare the performance of UmRMR with more unsupervised feature selection methods on more data sets.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 60503033, 60703086).

References

1. Langley, P.: Selection of Relevant Features in Machine Learning. In: AAAI Fall Symposium on Relevance, pp. 1–5. AAAI Press, New Orleans (1994)
2. Dash, M., Choi, K., Scheuermann, P., Liu, H.: Feature Selection for Clustering - a Filter Solution. In: 2nd IEEE International Conference on Data Mining, pp. 115–122. IEEE Press, Washington (2002)
3. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Trans. Pattern Analysis and Machine Intelligence 24, 301–312 (2002)
4. Wei, H.L., Billings, S.A.: Feature Subset Selection and Ranking for Data Dimensionality Reduction. IEEE Trans. Pattern Analysis and Machine Intelligence 29, 162–166 (2007)
5. Modha, D.S., Spangler, W.S.: Feature Weighting in k-Means Clustering. Machine Learning 52, 217–237 (2003)
6. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous Feature Selection and Clustering Using Mixture Models. IEEE Trans. Pattern Analysis and Machine Intelligence 26, 1154–1166 (2004)
7. Dy, J.G., Brodley, C.E.: Feature Selection for Unsupervised Learning. Journal of Machine Learning Research 5, 845–889 (2004)
8. Li, Y.H., Dong, M., Hua, J.: Localized Feature Selection for Clustering. Pattern Recognition Letters 29, 10–18 (2008)
9. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Trans. Neutral Networks 5, 537–550 (1994)
10. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. Pattern Analysis and Machine Intelligence 27, 1226–1238 (2005)
11. Blake, C., Merz, C.: UCI Repository of Machine Learning Database, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
12. Fayyad, U.M., Irani, K.B.: Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In: 13th International Joint Conference on Articial Intelligence, pp. 1022–1027. Morgan Kaufmann, Chambery (1993)