SHORT COMMUNICATIONS

Rough set-based feature selection method

ZHAN Yanmei, ZENG Xiangyang * and SUN Jincai

(College of Marine Engieering, Northwestern Polytechnical University, Xi' an 710072, China)

Received June 7, 2004; revised September 3, 2004

Abstract A new feature selection method is proposed based on the discern matrix in rough set in this paper. The main idea of this method is that the most effective feature, if used for classification, can distinguish the most number of samples belonging to different classes. Experiments are performed using this method to select relevant features for artificial datasets and real-world datasets. Results show that the selection method proposed can correctly select all the relevant features of artificial datasets and drastically reduce the number of features at the same time. In addition, when this method is used for the selection of classification features of real-world underwater targets, the number of classification features after selection drops to 20% of the original feature set, and the classification accuracy increases about 6% using dataset after feature selection.

Keywords: underwater target; rough set; discern matrix; feature selection.

Classification feature is a key factor for high classification accuracy of underwater target. In order to obtain effective classification features, researchers have done much work in feature extraction. Many features of underwater target have been successfully extracted using signal processing techniques, such as high order spectrum analysis, chaos and fractal, etc. However, high classification accuracy cannot be obtained unless different types of features are combined together. The combination of different kinds of features leads to a larger dimension of feature set and an increase in learning time of a learning algorithm. In some cases, a large number of features can even result in a decrease of classification accuracy. Therefore, selecting only a part of effective classification features from a large feature set call for more research work in feature selection.

Feature selection is a combination problem. Ideally, feature selection methods search 2^n (where n is the number of features in original feature set) candidate feature subsets for the best one according to some evaluation function. However, this selection procedure is exhaustive, and it may be too costly for practical use even for a medium-size feature set. Therefore, many feature selection methods have been investigated to pick an optimum or suboptimum feature subset according to certain evaluation function to avoid exhaustive search.

In Ref. [1], Dash et al. sum up the feature selection as a three-step process: search strategy, evaluation function and stopping criterion. In this paper, we break through this traditional feature selection frame, and propose a simple feature selection method based on rough set. Simulation results indicate that this new method is an effective selection method and it can be used in feature selection for underwater target.

1 Basic ideas of rough set

1.1 Rough set model

In rough set, an information system is defined as $\mathbb{S}=(\mathbb{U}, \mathbb{Q}, \mathbb{V})$. Here, $\mathbb{U}=\{x_1, x_2, \cdots, x_n\}$ is a non-empty finite set of instances, called universe; $\mathbb{Q}=\mathbb{A}\cup d$, where $\mathbb{A}$ is a finite set of features, called the condition feature set, and d is called the decision feature; $\mathbb{V}$ is the union collection of ranges of all the condition features $\mathbb{V}=\bigcup_{a\in A}\mathbb{V}_a$, where $\mathbb{V}_a$ is the range of feature a.

1.2 Indiscernibility relation and equivalence class

Given any subset $\mathbb{B}\subset\mathbb{A}$, let IND denote the indiscernibility relation of a feature set, then the in-

* To whom correspondence should be addressed. E-mail: zenggxy@nwpu.edu.cn

Table 1. Example of the information system

$\mathbb{U}$	$\mathbb{A}1$	$\mathbb{A}2$	$\mathbb{A}3$	$\mathbb{A}4$	d
x_1	0	0	1	0	0
x_2	1	0	2	1	1
x_3	1	1	1	0	0
x_4	0	2	1	1	1
x_5	1	2	1	0	1
x_6	1	0	1	0	0
x_7	1	2	2	1	1
x_8	0	0	2	1	1

discernibility relation of the feature set $\mathbb{B}$, denoted by IND($\mathbb{B}$), is defined as:

$$\mathrm{IND}(\mathbb{B})=\{(x_1, x_2)\in \mathbb{U}\times \mathbb{U}:\ \forall a\in \mathbb{B},\ a(x_1)=a(x_2)\}. \quad (1)$$

Here, IND($\mathbb{B}$) is called B-indiscernibility relation: if objects x_1 and x_2 belong to IND($\mathbb{B}$), then x_1 and x_2 are discernible from each other by all features from $\mathbb{B}$; the universe $\mathbb{U}$ can be divided into several equivalence classes according to the B-indiscernibility relation, denoted by $\mathbb{U}/\mathrm{IND}(\mathbb{B})$; the equivalence classes of x produced by IND($\mathbb{B}$) are denoted by $[x]_{\mathrm{IND}(\mathbb{B})}$. All the equivalence classes of IND($\mathbb{B}$) are called the elementary sets of $\mathbb{B}$. The equivalence classes formed by dividing the universe $\mathbb{U}$ with decision feature d are called decision classes.

1.3 Approximations of set

For any object subset $\mathbb{X}\subset\mathbb{U}$, the lower approximation of $\mathbb{X}$ using B-indiscernibility relation is defined as the union collection of all the elementary sets of $\mathbb{B}$ that are contained in $\mathbb{X}$

$$\underline{\mathbb{B}}\ \mathbb{X}=\bigcup\{x_i\in\mathbb{U}\mid[x_i]_{\mathrm{IND}(\mathbb{B})}\subseteq\mathbb{X}\} \quad (2a)$$

and the upper approximation is

$$\overline{\mathbb{B}}\ \mathbb{X}=\bigcup\{x_i\in\mathbb{U}\mid[x_i]_{\mathrm{IND}(\mathbb{B})}\mid\mathbb{X}\neq\Phi\}. \quad (2b)$$

1.4 Positive region, negative region and boundary region

The lower approximation and the upper approximation of $\mathbb{X}\subset\mathbb{U}$ divide the universe $\mathbb{U}$ into three regions: the positive region POS($\mathbb{X}$), the negative region NEG($\mathbb{X}$) and the boundary region BND($\mathbb{X}$):

$$\mathrm{POS}(\mathbb{X})=\underline{\mathbb{B}}\ \mathbb{X}, \quad (3a)$$

$$\mathrm{NEG}(\mathbb{X})=\mathbb{U}-\overline{\mathbb{B}}\ \mathbb{X}, \quad (3b)$$

$$\mathrm{BND}(\mathbb{X})=\overline{\mathbb{B}}\ \mathbb{X}-\underline{\mathbb{B}}\ \mathbb{X}. \quad (3c)$$

1.5 Significance of features

In feature selection, the significance of a feature $a\in\mathbb{A}$ can be determined by the degree of dependence of two feature sets $\mathbb{R}$, $\mathbb{P}\subseteq\mathbb{A}$. The dependence of $\mathbb{P}$ on $\mathbb{R}$ is defined as:

$$\gamma_{\mathbb{R}}(\mathbb{P})=\frac{|\mathrm{POS}_{\mathbb{R}}(\mathbb{P})|}{|\mathbb{U}|}. \quad (4)$$

where $\mathrm{POS}_{\mathbb{R}}(\mathbb{P})$ is the positive region of all the equivalence classes of feature set $\mathbb{R}$ in classification $\mathbb{U}/\mathrm{IND}(\mathbb{P})$.

The dependence of condition feature on decision feature is different for different features. Suppose the decision feature d is the only feature in feature set $\mathbb{P}$, remove feature a from condition feature subset $\mathbb{R}\subseteq\mathbb{A}$, the significance of feature a for classification $\mathbb{U}/\mathrm{IND}(\mathbb{P})$ is defined as:

$$\mathrm{SGF}(a,\mathbb{R},\mathbb{P})=\gamma_{\mathbb{R}}(\mathbb{P})-\gamma_{\mathbb{R}-\{a\}}(\mathbb{P}). \quad (5)$$

Here, SGF(a, $\mathbb{R}$, $\mathbb{P}$) represents the change of the dependence of $\mathbb{P}$ on $\mathbb{R}$ after feature a is removed from $\mathbb{R}$. The bigger the change of dependence is, the higher the significant feature a is. In this way, the significance of feature a is also reflected. It can be seen from Eq. (4) that the significance of feature a can also be scaled by the positive region. Given two feature sets $\mathbb{P}$ and $\mathbb{R}$, and feature $a\in\mathbb{R}$, if $\mathrm{POS}_{\mathbb{R}}(\mathbb{P})=\mathrm{POS}_{\mathbb{R}-\{a\}}(\mathbb{P})$, then feature a is called a redundant feature in feature set $\mathbb{R}$. Otherwise, feature a is indispensable in $\mathbb{R}$ to the feature set $\mathbb{P}$.

1.6 Reduction of feature set

Given the feature set $\mathbb{B}\subset\mathbb{A}$ and the decision feature d, if $\mathrm{POS}_{\mathbb{B}}(d)=\mathrm{POS}_{\mathbb{A}}(d)$, and every feature in $\mathbb{B}$ is indispensable to the decision feature d, then we call feature set $\mathbb{B}$ a reduction of information system $\mathbb{S}$. There may be many reductions for a single information system.

1.7 Core of feature set

For the decision feature d, the intersection of all the reductions of information system is called the core of feature set $\mathbb{A}$. The core of feature set can be obtained from the discern matrix.

1.8 Discern matrix

In the information system $\mathbb{S}$, the discern matrix of condition feature set $\mathbb{A}$, denoted by $M(\mathbb{A})=(m_{ij})_{n\times n}$, is defined as

$$M(\mathbb{A})=\begin{cases}\Phi, & x_i, x_j \text{ belong to the same equivalence class of } d,\\ \{a\in A:\ a(x_i)\neq a(x_j)\}, & x_i, x_j \text{ belong to different equivalence classes of } d.\end{cases} \tag{6}$$

Discern matrix $M(\mathbb{A})$ is a symmetric matrix, so only $m_{ij}(1\leqslant j\leqslant i\leqslant n)$ needs to be calculated. The discern matrix of information system in Table 1 is given in Table 2.

Table 2. The discern matrix

	x_1	x_3	x_6	x_2	x_4	x_5	x_7	x_8
x_1								
x_3								
x_6								
x_2	$\mathbb{A}1, \mathbb{A}3, \mathbb{A}4$	$\mathbb{A}2, \mathbb{A}3, \mathbb{A}4$	$\mathbb{A}3, \mathbb{A}4$					
x_4	$\mathbb{A}2, \mathbb{A}4$	$\mathbb{A}1, \mathbb{A}2, \mathbb{A}4$	$\mathbb{A}1, \mathbb{A}2, \mathbb{A}4$					
x_5	$\mathbb{A}1, \mathbb{A}2$	$\mathbb{A}2$	$\mathbb{A}2$					
x_7	$\mathbb{A}1, \mathbb{A}2, \mathbb{A}3, \mathbb{A}4$	$\mathbb{A}2, \mathbb{A}3, \mathbb{A}4$	$\mathbb{A}2, \mathbb{A}3$					
x_8	$\mathbb{A}3, \mathbb{A}4$	$\mathbb{A}1, \mathbb{A}2, \mathbb{A}3, \mathbb{A}4$	$\mathbb{A}1, \mathbb{A}3, \mathbb{A}4$					

The core of condition feature set $\mathbb{A}$ is composed of the units that have only one feature in the discern matrix. For example, the core of the information system in Table 1 is $\{\mathbb{A}2\}$.

2 Feature selection method

Based on the discern matrix of feature set $\mathbb{A}$, a new information system $\mathbb{S}'$ can be constructed. The composition of information system $\mathbb{S}'$ is as follows:

universe: $\mathbb{U}'=\{(x_i, x_j)\in\mathbb{U}\times\mathbb{U}:\ d(x_i)\neq d(x_j)\}$, decision feature: d': $\mathbb{U}'\rightarrow\{1\}$.

The features in condition feature set $\mathbb{A}'$ correspond to those in condition feature set $\mathbb{A}$, only differing in the feature values of each feature. In the new information system, given a condition feature $a'\in\mathbb{A}'$, the feature values of a' are

$$a'=\begin{cases}1, & a'(x_i)\neq a'(x_j),\ \forall(x_i, x_j)\in\mathbb{U}',\\ 0, & \text{else.}\end{cases} \tag{7}$$

According to the procedure above, the new information system constructed from the information system in Table 1 is presented in Table 3.

The purpose of feature selection is to select those features that can distinguish as many instances belonging to different classes as possible. It can be seen from Table 3 that the more the feature values are set to 1, the more powerful the feature is for classification. Based on this idea, a simple feature selection method is proposed in this paper. This method chooses the feature that has the most feature values set to 1 as the first feature selected. Suppose the first feature selected is feature a, remove all the instance pairs distinguished by feature a from universe $\mathbb{U}'$, and also remove feature a from feature set $\mathbb{A}'$. Use the set of features left as the condition feature set of a new information system and the instance pairs left as the universe of the new information system, and repeat the above selection process for the new information system. The selection process continues till the universe $\mathbb{U}'$ is empty.

Table 3. The information system $\mathbb{S}'$

$\mathbb{U}'$	$\mathbb{A}1'$	$\mathbb{A}2$	$\mathbb{A}3$	$\mathbb{A}4$	d'
(x_1, x_2)	1	0	1	1	1
(x_1, x_4)	0	1	0	1	1
(x_1, x_5)	1	1	0	0	1
(x_1, x_7)	1	1	1	1	1
(x_1, x_8)	0	0	1	1	1
(x_3, x_2)	0	1	1	1	1
(x_3, x_4)	1	1	0	1	1
(x_3, x_5)	0	1	0	0	1
(x_3, x_7)	0	1	1	1	1
(x_3, x_8)	1	1	1	1	1
(x_6, x_2)	0	0	1	1	1
(x_6, x_4)	1	1	0	1	1
(x_6, x_5)	0	1	0	0	1
(x_6, x_7)	0	1	1	1	1
(x_6, x_8)	1	0	1	1	1

From the definition of core of feature set, we know that all features in the core are indispensable to the decision feature d. That is, these features are indispensable to classification. Therefore, the selection method can start with core of the condition feature set. Firstly, core of the condition feature set is calculated, then the feature selection process continues using the method described above. Core of the condition feature set can be calculated easily based on the new information system $\mathbb{S}'$. According to the definition of core, core of the condition feature set is composed

of features corresponding to feature value 1 in those rows which have only one feature value set to 1.

The feature selection process is as follows:

(1) $Fea = \Phi$, where Fea is the selected feature set;

(2) Construct new information system $\mathbb{S}' = (\mathbb{U}', \mathbb{A}' \cup d', \mathbb{V}')$ from the information system $\mathbb{S}$;

(3) Find rows that have only one feature value that is set to 1 in the information system $\mathbb{S}'$ (feature value of decision feature d is not included). Among these rows, features in the column that 1 occurs consist of core($\mathbb{A}'$) (the core of feature set $\mathbb{A}'$);

(4) For each feature in core($\mathbb{A}'$), remove sample pairs in those rows that 1 occurs from universe $\mathbb{U}'$;

(5) Remove core($\mathbb{A}'$) from feature set $\mathbb{A}'$, that is, $\mathbb{A}' = \mathbb{A}' - \text{core}(\mathbb{A}')$;

(6) If $\mathbb{U}'$ is not empty, then, firstly, find a column that 1 occurs most frequently in $\mathbb{S}'$, suppose the feature of this column is a; secondly, set $Fea = Fea \cup a$; finally, remove sample pairs in rows that the feature values of a are set to 1 from $\mathbb{U}'$, and remove a from $\mathbb{A}'$, $\mathbb{A}' = \mathbb{A}' - a$;

(7) Output the feature set selected core($\mathbb{A}'$) $\cup$ Fea.

3 Experiment

Firstly, artificial datasets Corral, mofn-3-7-10 and parity5+5 are used to test the effectiveness of the feature selection method proposed in this paper; then the feature selection method is employed to select useful features for real-world datasets "wine" and classification features of underwater targets. The artificial datasets and dataset "wine" used in this paper come from http://www.ics.uci.edu/~mlearn/MLRepository.html. All the datasets are introduced briefly as follows and the composition of all these datasets is presented in Table 4.

Corral: Corral is a Boolean dataset which has six features: $\mathbb{A}0$, $\mathbb{A}1$, $\mathbb{B}0$, $\mathbb{B}1$, "irrelevant" and "correlated", each feature has two feature values {0, 1}. The target concept of this dataset is $(\mathbb{A}0 \wedge \mathbb{A}1) \vee (\mathbb{B}0 \wedge \mathbb{B}1)$. The feature named "irrelevant" is uniformly random, and the feature named "correlated" matches the class label 75% of the time.

mofn-3-7-10: There are ten features numbered 1~10 in this Boolean dataset. The target concept of this dataset is: The labels of the samples that at least three features with feature number 3 to 9 are set to one are 1, otherwise, are 0. Features 1, 2, 10 are irrelevant features. It is usually difficult to select all the relevant features of this dataset. Since there are interactions among the seven relevant features, most algorithms are unable to identify the relevant features correctly.

parity5+5: This Boolean dataset has ten features in all. Features numbered 2, 3, 4, 6 and 8 are five relevant features, the rest features numbered 1, 5, 7, 9 and 10 are irrelevant.

wine: This dataset has 13 continuous features. The aim is to classify two different kinds of wine.

Classification features for underwater targets: The features in this dataset include 24 wave structure features[3], 16 envelope energy features[4], and two wavelet fractal features of power spectrum[5]. There are altogether 456 samples in this dataset, choose two thirds of them as the training samples and the rest as the testing samples.

Table 4. Datasets used in the experiment

Dataset	Feature No. a)	Class No. b)	Tr. No. c)	Ts. No. d)
Corral	6	2	32	128
mofn-3-7-10	10	2	300	1024
Parity5+5	10	2	100	1024
Wine	13	3	118	178
Underwater target	42	3	304	456

a) Number of features; b) number of classes; c) number of instances in training set; d) number of instances in testing set.

The features of the two real-world datasets used in this paper are all continuous features. Since the selection method proposed here can be used for only discrete features, the continuous features should be discretized before the feature selection method is used for the two real-world datasets. Here we use the Recursive Minimal Entropy discretization method[6] to discretize the continuous features. After discretization, the feature selection method is used to select the relevant features of the discretized real-world datasets.

For artificial datasets, the effectiveness of a feature selection method can be tested by comparing the selected features with the known relevant features. If

the features selected are the same as the known relevant features of artificial datasets, then the feature selection method is effective. However, the classification information about the real-world datasets is unknown, so the effectiveness of the feature selection method can only be judged by comparing the classification accuracies before and after feature selection of real-world dataset. In this paper, we utilize C4.5[7] to calculate the classification accuracies of the feature set before and after feature selection. Since the training set and the testing set of artificial datasets are fixed, the feature selection process and classification can be calculated only once. However, for real-world datasets the training set and the testing set can be selected randomly, and feature selection and classification must run multiple times to get a stable solution. Therefore, we calculate 50 times for each real-world dataset in this paper. Results in Table 5 are the average of 50 calculations. C4.5 in Table 5 denotes the classification accuracy of the dataset before feature selection, and (C4.5+feature selection) denotes the classification accuracy after feature selection.

From the results in Table 5, we can see that the feature selection method proposed in this paper can select the relevant features correctly for all the three artificial datasets, and the classification accuracy with C4.5 can also be improved dramatically after feature selection. The increase in classification accuracy is especially obvious for dataset parity5+5. The classification accuracy of the original dataset is only 50% for parity5+5, and it is improved by 31.2% after feature selection. The average results for real-world dataset show that the number of features reduces drastically after the feature selection method is used. The number of features for dataset "wine" drops from original 13 to about 4, and from 42 to 8 for underwater target classification features (cutting off 80% or so of the original feature set). The variance of 50 calculations shows that the difference of the number of features selected each time is very small for dataset "wine", while the number of features selected for under water targets has bigger fluctuation. The classification accuracy for underwater targets with C4.5 improves about six percent after feature selection. Although the classification accuracy for wine drops after feature selection, it only decreases slightly and hardly has any bad effect on classification.

Table 5. Feature selection results

Dataset	Feature selected	C4.5	C4.5+feature selection
Corral	A0, A1, B0, B1	0.812	1.000
Mofn-3-7-10	3, 4, 5, 6, 7, 8, 9	0.879	0.961
Parity5+5	2, 3, 4, 6, 8	0.500	0.812
Wine	4.02±0.714	0.921±0.037	0.918±0.033
Underwater target	7.8±2.967	0.875±0.049	0.933±0.048

4 Conclusion

The experimental results have verified that the feature selection method proposed in this paper is an effective feature selection method. This method can correctly select all the relevant features of artificial datasets, and the classification accuracies of the artificial datasets using C4.5 are also improved after feature selection. When the selection method is used for real-world dataset, it can not only drastically reduce the number of features, but also improve the classification accuracy with C4.5 for certain dataset (for example, the classification features of underwater target).

References

1 Dash M. and Liu H. Feature selection for classification. Intelligent Data Analysis, 1997, 1(3): 131—156.

2 Liu T. M. Data Mining Technology and Its Application (in Chinese). 1st ed. Beijing: National Defence Industry Publishing House, 2001, 55—101.

3 Cai Y. B., Zhang M. Z., Shi X. Z. et al. Extraction and classification of wave structure features in ship noise. Acta Electronica Sinica (in Chinese), 1999, 27(6): 129—130.

4 Hu C. H., Zhang J. B., Xia J. et al. System Analysis and Design Based on MATLAB—Wavelet Analysis (in Chinese). 1st ed. Xi'an: Xi'an Electron Scientific and Technical University Publishing House, 1999, 264—271.

5 Chen J., Chen K. A. and Sun J. C. A new method for feature extraction in ship noise. Journal of Northwestern Polytechnical University, 2000, 18(2): 241—244.

6 Dougherty J., Kohavi R. and Sahami M. Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th International Conference on Machine Learning, 1995, 194—202.

7 http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html