# Feature Selection via Modified RSBRA for SVM Classifiers

Ye Li, Zhonghui Hu, Yunze Cai, and Xiaoming Xu

*Abstract*—**Discretization can remove redundant and irrelative attributes during converting continuous attributes into discretized ones and therefore can be used for feature selection. Rough sets and Boolean reasoning based discretization approach (RSBRA), put forward by Nguyen in 1995 [8], is very noticeable for its efficiency of reduction. However, the RSBRA is not a suitable feature selection method for machine learning algorithm such as neural network or SVM because too much useful information loses due to the discretization. In this paper, we present a modified RSBRA for feature selection and evaluate it with SVM classifiers. In the presented algorithm, the level of consistency, coined from the rough sets theory, is introduced to substitute the stop criterion of circulation of the RSBRA, which maintains the fidelity of the training set after discretization. Experiment results show the modified algorithm has better predictive accuracies and less training time than the original RSBRA.**

## I. Introduction

FEATURE selection is an important problem in the research of pattern recognition. In some cases, too many redundant or irrelative features may overpower main features for classification. Feature selection can remedy this problem and therefore improve the prediction accuracy and reduce the computational overhead of classification algorithms. Discretization can convert continuous attributes into discretized ones and synchronously remove redundant and irrelative attributes. Therefore, discretization methods can be used for feature selection. For example, the Chi2 algorithm as a discretization method is also an effective approach of feature selection [6]. The merit of discretization is it can deal with hybrid attributes and multi-class problems readily.

Manuscript received September 15, 2004. This work was supported by National Basic Research Program of China under Grant 2002cb312200-01-1 and National Nature Science Foundation of China under Grant 60174038.

Ye Li is with the Department of Automation, Shanghai Jiaotong University, Shanghai 200030 China (phone: 86-21-62826946; e-mail: liyemail@sjtu.edu.cn).

Zhonghui Hu is with the Department of Automation, Shanghai Jiaotong University, Shanghai 200030 China (e-mail: huhzh@sjtu.edu.cn).

Yunze Cai is with the Department of Automation, Shanghai Jiaotong University, Shanghai 200030 China (e-mail: yzcai@sjtu.edu.cn).

Xiaoming Xu is with University of Shanghai for Science and Technology, Shanghai 200030 China (e-mail: xmxu@mail.sjtu.edu.cn).

Discretization methods can be supervised or unsupervised. If no information of instance labels is utilized during discretization, we call the method unsupervised; or, we call it supervised. The simplest equal-width-intervals and equal-frequency-intervals methods [1] are unsupervised. Supervised discretization methods include class-driven statistical discretization [2], 1-rules [3], entropy-based discretization [4], ChiMerge [5], Chi2 [6] and modified Chi2 [7], and so on. Supervised discretization methods can keep key information for classification and therefore show more favorable performance in feature selection.

Rough sets and Boolean reasoning based discretization approach (RSBRA), which is put forward by Nguyen in 1995 [8] as a global and supervised discretization method, is very noticeable for its great efficiency of reduction. It mines the intrinsic knowledge from the original dataset and utilizes the capability of cut discerning samples from different class as the criterion to choose proper cuts from an initial cut set. Though it is a good discretization method for many machine learning algorithms, however, it is not suitable for feature selection for neural network or SVM because the information loss due to discretization is too large. In this paper, we present a modified RSBRA for feature selection and evaluate it with SVM classifiers. Though discretization is usually a needless preprocessing step for neural network or SVM, which can deal with continuous or hybrid attributes directly, it is still attractive because it can improve the generation performance, reduce the training time and decrease the storage space requirement of a machine learning algorithm.

This paper is organized as follows. Section I is the introduction. Section II gives some basic concepts of the rough sets theory and section III gives the general description of discretization based feature selection problems. Section IV analyzes the RSBRA and presents the modified algorithm. Section V are experiments and section VI draws conclusions.

## II. Rough Sets Preliminaries

An information system is a 4-tuple $S = (U, A, V, f)$, where $U$ is a non-empty finite set of objects, $A$ is a non-empty finite set of attributes, $V$ is the union of

attributes domains, i.e., $V = \bigcup Va$ for $\forall a \in A$, where $V_a$ denotes the domain of the attribute $a$, $f : U \times A \to V$ is an information function which for $\forall a \in A$ and $x \in U$, $f(x,a) \in Va$. A 5-tuple $T = (U, C \cup D, V, f)$ is called a decision table, if $S = (U, A, V, f)$ is an information system and $A = C \cup D$, where $C$ is the set of condition attributes and $D$ the set of decision attributes.

Each subset of attributes $B \in A$ determines a binary indiscernibility relation:

$$IND(B) = \{(x,y) \in U \times U \mid \forall a \in B, f(x,a) = f(y,a)\}$$

The indiscernibility relation $IND(B)$ partitions $U$ into some equivalence classes. An equivalence class of $IND(B)$ is denoted by $B(x)$, while the family of all equivalence classes is denoted by $U/IND(B)$, or simply by $U/B$. If $(x,y)$ belongs to $IND(B)$ we will say that $x$ and $y$ are $B-$ indiscernible. Equivalence classes of the relation $IND(B)$ are referred to as $B-$ granules.

The $B-$ lower and $B-$ upper approximation of $X \subset U$ are respectively defined as follows:

$$B_*(x) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\}$$

$$B^*(x) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \varnothing\}$$

The set $BN_{B(X)} = B^*(x) - B_*(x)$ is called $B-$ boundary region of $X$.

The lower approximation of a set $X$ with respect to $B$ is the set of all objects which can be for certain classified as $X$ using $B$. The upper approximation of a set $X$ with respect to $B$ is the set of all objects which can be possibly classified as $X$ using $B$. The boundary region of a set $X$ with respect to $B$ is the set of all objects which can be classified neither as $X$ nor as not $\overline{X}$ using $B$.

$B$-positive region of $D$ is defined as:

$$POS_B(D) = \bigcup_{X \in U/D} B_*(X)$$

The level of consistency, denoted as $L_c$, is defined as:

$$L_c = | POS_B(D) | / | U | \tag{1}$$

$L_c$ represents the percentage of instances in $U$ which can be classified into the equivalence classes determined by indiscernibility relation $IND(D)$. For a consistent data table there is always $L_c = 1$.

## III. Discretization Based Feature Selection

The discretization problem can be described as follows:

For an information system $T = (U, C \cup D, V, f)$, assume $\forall a \in C, Va = [la, ra]$ and $Pa$ a partition of $Va$ i.e.

$$Pa = \{[c_a^0, c_a^1), [c_a^1, c_a^2), ..., [c_a^k, c_a^{k+1})\}$$

where $k$ is an integer and

$$la = c_a^0 < c_a^1 < ... < c_a^{k+1} = ra$$

$$Va = [c_a^0, c_a^1) \cup [c_a^1, c_a^2) \cup ... \cup [c_a^k, c_a^{k+1})$$

$c$ is called a cut. Any family $P = \{Pa : a \in A\}$ of partition defines a new decision table $T^P = (U, C \cup D, V^P, f^P)$, where $\forall x \in U$, $f^P(x_a) = i \Leftrightarrow f(x_a) \in [c_a^i, c_a^{i+1})$, $i \in \{0, 1, ..., k\}$. Therefore, discretization is in nature a process of partitioning data space into several finite subspaces in which instances belong to the same class. For some attributes, discretization can not find any cut, which means these attributes are useless for keeping discernible relations of the information system and can be deleted. This is the key of discretization based feature selection.

Usually, discretization based feature selection includes three steps:

1) Compute candidate cut set;
2) Select practical cuts from the initial candidate cut set according to some criterion;
3) Discretize the data by the practical cuts and remove attributes without any practical cut selected.

## IV. The Modified RSBRA

The RSBRA is based on the rough sets theory and Boolean reasoning. The core of the rough sets theory is the indiscernible relation between samples; hence from the view of rough sets theory, discretization needs to reserve the indiscernible relation information of the decision system; or it may lead to information loss or introduce error information and therefore decrease the accuracy of results.

For discretization, the selection of practical cuts (i.e. step 2) is of most importance. In the RSBRA, the candidate cuts are chosen as follows: for each attribute, sort the attribute values and choose the average attribute value of any two neighbor instances, whose labels are different, as a candidate cut. The $k$ th candidate cut of attribute $a$ is denoted as $c_a^k$. And then a new information table $A^*$ which expresses the same indiscernible relations as the original information table $A$ is constructed. In the new information table, the columns are candidate cuts and the rows are pairs of samples from different classes $i$ and $j$ i.e. $(x_i, x_j)$. The value $A^*((x_i, x_j), c_a^k)$ is computed as follows: if $(x_i, x_j)$ can be discerned by the cut $c_a^k$ i.e. $a(x_i) < c_a^k < a(x_j)$ or $a(x_j) < c_a^k < a(x_i)$, then set $A^*((x_i, x_j), c_a^k) = 1$; or, set $A^*((x_i, x_j), c_a^k) = 0$. Therefore the number of '1' in each column represents the influence of the corresponding candidate cut on decision-making. The candidate cut with

the largest number is chosen as a practical cut and then the corresponding column and all pairs of samples which can be discerned by the chosen cut are removed from the information table $A^*$. After that, it is needed to recalculate the discernible capability of residual candidate cuts. The process is repeated until all pairs of samples from different class can be discerned by the result set of cuts and then the original data is discretized using the result cuts.

The RSBRA is a very efficient method of reduction. Nevertheless, the RSBRA is not a suitable feature selection method for SVM and leads to low prediction accuracy of SVM because from the view of SVM too much useful information loses due to discretization. Therefore, it is necessary to use some measure of information loss to control the discretization process. Usually there are three kinds of measures for feature selection i.e. accuracy measure, consistency measure and classic measures. The accuracy measure evaluates feature subset by accuracy of machine learning algorithms; the consistency measure by the percentage of inconsistent samples; the classic measures include information based measures, distance based measures, relativity based measures, and so on. In this paper, we use the consistency measure and have the following proposition.

**Proposition** Suppose that $T^P = (U, C \cup D, V^P, f^P)$ and $T^{P'} = (U, C \cup D, V^{P'}, f^{P'})$ are discretized decision tables of $T = (U, C \cup D, V, f)$ defined by partitions $P = \{P_a : a \in C\}$ and $P' = \{P'_a : a \in C\}$ respectively and $S_c$ and $S_c' = S_c \cup \{c_a\}$ are the set of cuts corresponding to the partition $P$ and $P'$ respectively, where $c_a$ is a cut of attribute $a \in C$. Let $L_c = |POS_C(D)| / |U|$ be the level of consistency of $T^P$ and $L_c'$ the level of consistency of $T^{P'}$. Then there is $L_c' \geq L_c$.

**Proof** The cut $c_a$ of attribute $a$ separates some equivalent class $B(x)$ of $T^P$ into two classes $B_1(x)$ and $B_2(x)$ where $B(x) = B_1(x) \cup B_2(x)$. Obviously, $B_1(x)$ and $B_2(x)$ are equivalent classes of $T^{P'}$.

If $B(x)$ is consistent, then $B_1(x)$ and $B_2(x)$ are sure to be consistent; hence, the level of consistency does not change after $c_a$ is added to the cut set i.e. $L_c' = L_c$.

In the case that $B(x)$ is inconsistent, if the cut $c_a$ is between the continuous attribute values of any two inconsistent samples, then after discretization, the number of inconsistent samples may decrease and there is $L_c' > L_c$; or, there is $L_c' = L_c$.

Therefore, there is always $L_c' \geq L_c$. □

The above proposition shows that the level of consistency of discretized data rises monotonously with the increase of the number of cuts. There is a limit for such increase, i.e., the initial level of consistency of the dataset. For a consistent dataset, it is 1, while for an inconsistent dataset, it is smaller than 1.

Because the consistency measure is monotone, in the modified the RSBRA, we can use the consistency measure as the stop criterion in order to maintain the fidelity of the dataset after discretization. After choosing a practical cut, we discretize the dataset by the chosen cuts and judge whether the level of consistency of the discretized data table reaches that of the original data table. If not, the next cut with greatest discernible capability is added to the practical cuts set. By this a final practical cuts set is obtained.

If there are several candidate cuts with the same discernible capability, choose the candidate cut which can increase the level of consistency most. This may help for choosing less practical cuts. Another modification to the RSBRA is that we don't recalculate the discernible capability of residual candidate cuts after choosing a practical cut. We mean to delete irrelative and redundant features from the dataset but reserve enough information for SVM classifiers. It may lead to a little decrease of efficiency of feature selection; however, the performances of posterior machine learning algorithms are better. In addition, by discretization many samples may become entirely same and all but one of them can be deleted from the data table. Thus, the size of the data decreases horizontally, which is helpful for reducing training time.

The RSBRA is too expensive in both time and space complexity to be practical. A more economical version of RSBRA is presented in [9] and gives an easier method to compute discernible capability of a cut. For a classification problem of $r$ class, $c_a^j$ ( $j = 1, ..., r$ ) the $j$ th candidate cut of attribute $a$, let

$$l_j(c_a^m) = |\{x \in U : [a(x) < c_a^m] \wedge [d(x) = j]\}|$$

$$r_j(c_a^m) = |\{x \in U : [a(x) > c_a^m] \wedge [d(x) = j]\}|$$

Then the discernible capability of $c_a^j$ is computed as

$$W_a^m = \sum_{j=1}^r l_j(c_a^m) \cdot \sum_{j=1}^r r_j(c_a^m) - \sum_{i=1}^r l_i(c_a^m) \cdot r_i(c_a^m) \quad (2)$$

Now, we describe the modified algorithm as follows:

**Algorithm**: Feature Selection for SVM Classifiers

*Step 1*. Set $P_{practical} = \phi$ and compute the level of consistency of original data $L_{c-original}$ according to (1).

*Step 2*. Generate the candidate cut set. For any sorted attribute, the average attribute value of any two neighbor

instances with different labels, is a candidate cut of it.

*Step 3*. Compute the discernible capability of every candidate cut according to (2).

*Step 4*. Set the candidate cut with largest discernible capability $c_{max}$ as a practical cut. Add $c_{max}$ to the practical cut set and remove it from the candidate cut set; namely, set $P_{practical} = P_{practical} \cup \{c_{max}\}$ , $P_{possible} = P_{possible} / \{c_{max}\}$ . If there are several candidate cuts with the same discernible capability, choose the candidate cut which can increase the level of consistency most.

*Step 5*. Discretize the data using $P_{practical}$ and compute the level of consistency of the discretized data $L_{c-discretized}$.

*Step 6*. If $L_{c-discretized} < L_{c-original}$ , then go to step 4 else go to step 7.

*Step 7*. Delete from the original data attributes with no practical cut chosen and then discretize the data using $P_{practical}$.

## V. Experiments

We select two datasets, Breast-cancer-wisconsin and Glass, from the UCI repository [10] and a dataset Heart from the Statlog repository [11] for our experiments. Table I gives a summary of these datasets. 16 samples with missing values have been removed from Breast-cancer-wisconsin dataset. Every dataset is separated randomly into a training set with 70% samples and a testing set with 30% samples. For comparison, the same training and testing sets are used to do three kinds of experiments:

1) SVM. The data is fed to a SVM classifier directly without preprocessing.
2) RSBRA + SVM. The RSBRA based feature selection is applied before the data is fed to a SVM classifier.
3) Modified RSBRA + SVM. The presented algorithm is applied before the data is fed to a SVM classifier.

We repeat each kind of experiment ten times for every dataset. In the experiments, LIBSVM (Version 2.6) [12] is chosen to be the benchmark for evaluating the performance of the three kinds of experiments. C-SVC [13] and RBF kernel are chosen as the model type of SVM classifiers and kernel function respectively. In every experiment, the parameters c and g of the model are searched by 5-folds cross-validation. The search ranges are [-10, 15] and [10, -15] respectively and the search steps are 1 and -1 respectively. Table II and table III show respectively the average predictive accuracies of ten experiments and the change of attribute numbers.

Experiment results show that RSBRA is efficient in reduction, however, it produces low predictive accuracies when it is used as a feature selection method for SVM classifiers. Obviously, the modified algorithm presented in this paper is more suitable for feature selection for SVM. It not only removes irrelevant or redundant attributes from the data, but also improves the performance of SVM classifiers. In addition, all datasets gain smaller size after discretization. Especially, the size of Breast Cancer Wisconsin dataset decreases to an average 41.80%, which greatly reduces the training time.

TABLE I
DATA SETS INFORMATION

| Names | Examples | Continuous attributes | Discrete attributes | Classes |
|---|---|---|---|---|
| Breast cancer wisconsin | 683 | 9 | 0 | 2 |
| Glass | 214 | 9 | 0 | 6 |
| Heart | 270 | 6 | 7 | 2 |

TABLE II
PREDICTIVE ACCURACY (PERCENT)

| Names | SVM | RSBRA + SVM | Modified RSBRA + SVM |
|---|---|---|---|
| Breast cancer wisconsin | 96.42 | 77.01 | 96.67 |
| Glass | 68.28 | 59.69 | 70.94 |
| Heart | 82.96 | 73.70 | 84.69 |

TABLE III
A COMPARISON OF ATTRIBUTE NUMBERS BEFORE/AFTER DISCRETIZATION

| Names | Before Discretization | RSBRA | Modified RSBRA |
|---|---|---|---|
| Breast cancer wisconsin | 9 | 5.3 | 7.50 |
| Glass | 9 | 6.3 | 7.30 |
| Heart | 13 | 6.4 | 11 |

## VI. Conclusion

In this paper, the modified RSBRA is proposed as a feature selection method. Several modifications to RSBRA are made:

1) The level of consistency, coined from the rough sets theory, is introduced to maintain the fidelity of the training set after discretization.
2) After choosing a practical cut we don't recalculate the discernible capability of residual candidate cuts because we don't mean to select an optimal feature subset but to remove the irrelevant and most redundant features and therefore reserve key classification information needed for SVM classifiers.
3) If there are several candidate cuts with the same discernible capability, choose the candidate cut which can increase the level of consistency most. This helps for a smaller set of practical cuts.

Though discretization may cause information loss and therefore in many cases it would be better to deal with data

directly other than take discretization as a preprocessing measure [14], however, we show that proper discretization can both improve performance of classifiers and reducing training time.

Compared with the original RSBRA, the modified algorithm generates a bigger attribute number. This is not favorable and the problem will be studied in future researches.

## REFERENCES


[1] Wong, A. K. C. and Chiu, D. K. Y., "Synthesizing statistical knowledge from incomplete mixed-mode data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9 no. 6, pp.796-805, November 1987.

[2] Richeldi M and Rossotto M., "Class-driven statistical discretization of continuous attributes (extended abstract)" [A]. In: *N Lavrac & S. Wrobel, dcs, Machine Learning: ECML-95, Lecture Notes in Artificial Intelligence 914*, Springer Verlag [C]. Berlin, Heigelberg, New York, 1995, 335-338

[3] Holte R C., "Very simple classification rules perform well on most commonly used datasets" [J]. *Machine Learning*, 1993, 11:63-90

[4] Chmielewski M R and Grzymala-Busse J W, "Global discretization of attributes as pre-precessing for machine learning" [A]. *Proc of the III International Workshop on RSSC94 [C]*. 1994,294-301

[5] R. Kerber, "ChiMerge: Discretization of Numeric Attributes," *Proc. AAAI-92,Ninth Int'l Conf. Artificial Intelligence*, Pp.123-128. AAAI Press/The MIT Press, 1992.

[6] Huan Liu and Rudy Setiono. "Feature selection via discretization," *IEEE Transaction on Knowledge and Data Engineering*. Vol 9, No. 4, July/August 1997

[7] Francis E.H. Tay and Lixiang Shen. "A modified Chi2 algorithm for discretization". *IEEE Transaction on Knowledge and Data Engineering*. Vol 14, No. 3, May/June 2002

[8] H. S. Nguyen and A. Skowron, "Quantization Of Real Value Attributes Rough Set and Boolean Reasoning Approach", *Proc. Of the Second Joint Annual Conference on Information Sciences*, 1995, 37-37. Wrightsville Beach, NC

[9] H. S. Nguyen, "Some efficient algorithms for rough set methods", in *Proceedings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'96*, July 1996, Granada, Spain, pp. 1451-1456

[10] UCI KDD Archive, Available: http://kdd.ics.uci.edu/

[11] Statlog repository, Available: http://www.liacc.up.pt/ML/statlog/datasets.html

[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM, Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Boser, B., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, PA: ACM Press, 1992. 144~152

[14] Dan Ventura and Tony R. Martinez. "An Empirical Comparison of Discretization Methods". *Proceedings of the Tenth International Symposium on Computer and Information Sciences*, pp. 443-450, 1995