A method of nonorthogonal discretization in multidimensional feature space in the process of decision rule discovery

Anna Samborska

Institute of Computer Graphics and Multimedia Systems
Department of Computer Science and Information Systems
Szczecin University of Technology
ul. Zolnierska 49, Szczecin, Poland
asamborska@wi.ps.pl

Abstract. This work proposes a new approach to discretization in multidimensional space of conditional attributes in the classification process. The introduced method divides feature space into irregular sectors using nonorthogonal cuttings and produces decision rules of higher predictive accuracy than the orthogonal discretization. The comparison with the traditional approach revealed that the introduced method can be especially useful when decision classes are not monotonically distributed in domains of particular conditional attributes.

Keywords: nonorthogonal discretization, classification, rule discovery, roulette wheel selection

1. Introduction

Many of classification tasks cannot be solved by analytical methods. In such cases there are often performed methods of decision rules discovery, mostly in the following form [4]:

IF < *statement* > THEN < *conclusion* >,

where *statement* is a set of terms:

(*conditional attribute* < arithmetical operator > *value*)

united by logical operators.

The decision attribute is usually expressed in nominal or ordinal scale (divided into so-called decision classes). If conditional attributes are real values, it is often required to perform discretization – transfer all the objects (defined by a feature vector X **(1)** and a decision class they belong to) from a continuous to an ordinal feature space [1].

$$X = [x_1\ x_2 \ldots x_n] \quad \textbf{(1)}$$

Usually, discretization consists in dividing domain of each conditional attribute into intervals (each defines one class of the current attribute) using a set of 'cuttings':

$$C_{ij}=A_{ij}\,, \tag{2}$$

where i – an index of the attribute x_i , j – an index of a cutting for the attribute x_i and, assuming that the domain has been normalized – $A_{ij}\in$ [0,1]. To obtain n classes for an attribute x_i we need to perform n-1 cuttings.

The resulted system of independent cuttings, which discretizes the feature space orthogonally, can be an appropriate and efficient solution when both: the conditional attributes are strongly correlated with decision and decision classes are monotonically distributed within attributes' domains. In some cases, it is difficult to discretize each attribute's domain separately. Of course, it can be done arbitrarily [1] (for example 5 intervals, each 0.2 long) but the cuttings based on the data distribution are used more often although may cause some problems (fig. 1). Sometimes it is suggested to increase the number of the cuttings, but unfortunately, it can result in overfitting. Therefore, if one attempts to discover high-confidence and strong classification rules in multidimensional feature space, it is crucial to analyze spatial distribution of the data while performing discretization [5].

In adaptive classification systems (when dataset is successively supplied by new samples, and rules are modified accordingly), system of orthogonal cuttings can have insufficient number of degrees of freedom to adapt to the mutable dataset.

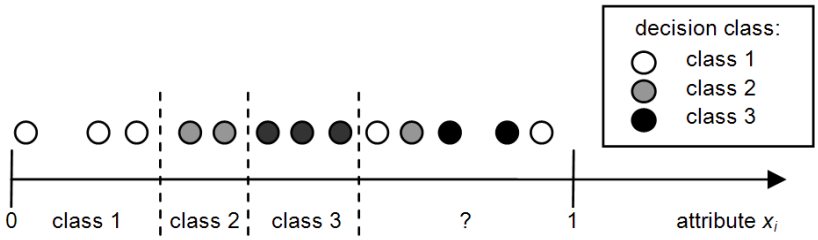


Fig. 1. An example of discretization problem in continuous domain of conditional attribute x_i

2. Nonorthogonal vs. orthogonal discretization

Author proposes a new method of discretization that takes into consideration all dimensions of feature space and produces irregularly shaped cells, fitting to the spatial distribution of the data (fig. 2). For further consideration we assume that feature space is normalized.

The idea is to split the feature space by nonorthogonal cuttings C_{ij}, each defined by a set of values A_{ijk}:

$$C_{ij} = \{ A_{ij1}, A_{ij2}, \dots , A_{ijk} , \dots , A_{ijn} \}, \tag{3}$$

where i – an index of attribute x_i, j – an index of a cutting for the attribute x_i, $n\geq 2$ – dimensionality of the feature space. Taking all A_{ijk} values (k=1…n), we can calculate coefficients of a plane (or a hyperplane) equation for the cutting C_{ij} by solving:

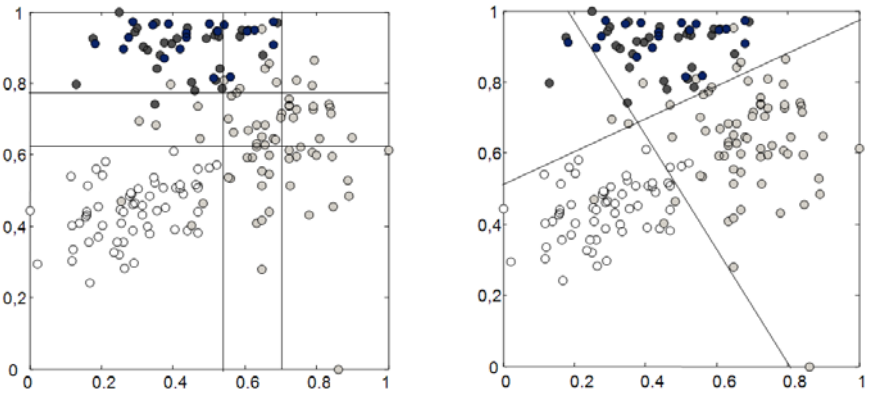
$$|M_{ij}| = 0 \tag{4}$$

where:

$$M_{ij} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_i & \dots & x_n & 1 \\ 0 & 0 & 0 & \dots & A_{ij1} & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & A_{ij2} & \dots & 0 & 1 \\ 0 & 1 & 0 & \dots & A_{ij3} & \dots & 0 & 1 \\ 0 & 0 & 1 & \dots & A_{ij4} & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A_{ij(n-1)} & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & A_{ijn} & \dots & 1 & 1 \end{bmatrix} \tag{5}$$

and $x_1 \dots x_n$ – variables associated with conditional attributes.

It seems to be much simpler to use plane (hyperplane) equation coefficients for cuttings definition instead of A_{ijk} values, but in practice the coefficients do not provide direct information about plane's orientation and position. Additionally, manipulating A_{ijk} values prevents planes' 'escape' from feature space, because of their accessible domain: [0,1]. Thus, it is easier to operate the cuttings – one does not have to check many border conditions for each of the coefficients.



$C_{11} = 0.53$ $C_{12} = 0.70$ $C_{21} = 0.62$ $C_{22} = 0.78$

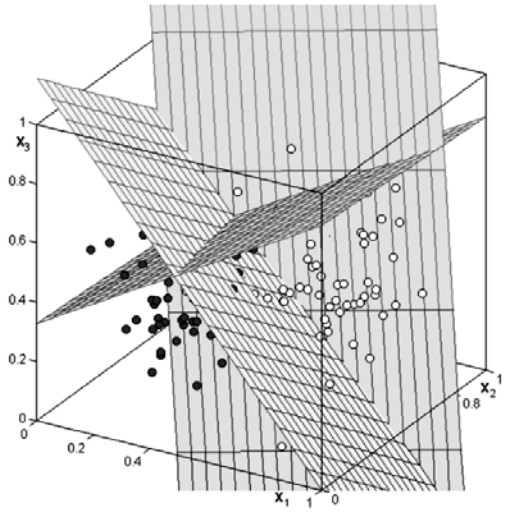
$C_{11} = \{ 0.81, 0.19 \}$ $C_{12} = \{ 0.52, 0.97 \}$

Fig. 2. An example of the cuttings in 2-dimensional feature space, a) orthogonal discretization b) nonorthogonal discretization

Nonorthogonal feature space partitioning can be more efficient way of discretization – especially when the dataset is successively expanding and rule set has to adapt to varying conditions by slight changes in cuttings system:

– more degrees of freedom gives wider range of cuttings manipulation,
– irregularly shaped cells are easier to adjust to spatial distribution of the data (fig.3),

– in many cases we can reduce the number of nonorthogonal cuttings, comparing to orthogonal discretization in the same classification task (fig. 2b).



C_{11} = { 0.99, 0.56, 0.13 }
C_{21} = { 0.80, 0.83, 0.70 }
C_{31} = { 0.33, 0.89, 0.29 }

Fig. 3. An example of nonorthogonal discretization in 3-dimensional feature space

3. Classification rules discovery

Each irregularly shaped cell (further called sector) after performing the nonorthogonal cuttings can produce one decision rule. So, it is essential to calculate spatial position of the sample *s* in reference to all the cutting planes (hyperplanes). Therefore, in nonorthogonally discretized feature space, each decision rule can be defined as follows:

$$\text{IF } (s \square C_{11}) \text{ and } (s \square C_{12}) \text{ and } \ldots \qquad \text{THEN } (s \in p) \tag{6}$$

where *s* – classified sample, □ – arithmetical operator (≤ or >), *p* – predicted decision class.

The location of the sample *s* in reference to the cutting C_{ij} can be calculated using determinant of the matrix M_{ij} (**5**) for feature vector $[x_1\ x_2 \ldots x_n]$ (**1**):

$$\begin{aligned} (|M_{ij}| \le 0) &\Rightarrow (x \le C_{ij}) \\ (|M_{ij}| > 0) &\Rightarrow (x > C_{ij}) \end{aligned} \tag{7}$$

In order to simplify the formula (**6**) and to make the sectors manipulation more effective, each cell is labeled by a series of binary values. They define cell position in reference to all the cuttings – if the sector is located below the plane (hyperplane) C_{ij}, the proper bit is set to '0', otherwise '1'. Transferring the binary values to a decimal

numbers we obtain shorter, unique label for each of the sectors. Now we can simplify notation of the formula (**6**):

$$\text{IF } (s \in S_L) \text{ THEN } (s \in p) \tag{8}$$

where S_L – decimally labeled sector. Predicted decision class p for sector S_L is determined as the dominant decision class within the sector.

Rule acceptance takes place after performing 2 tests:

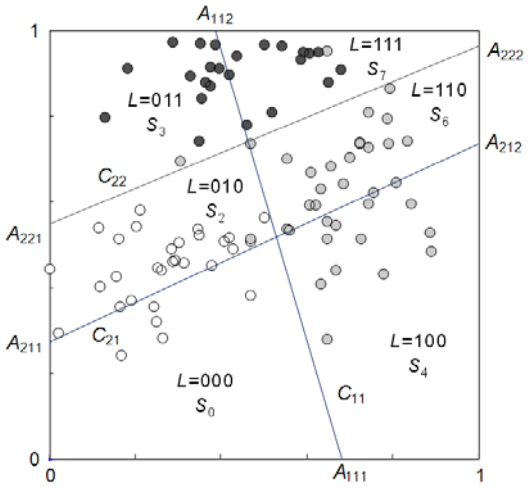
1. *confidence test* – the dominant decision class representatives - to - all samples in the cell ratio must exceed the *confidence threshold*,
2. *strength test* – the dominant decision class in the cell must be of a higher percentage of all the class representatives than the *strength threshold*.

An additional mechanism that supports rule acceptance consists in uniting adjacent cells if they produce the same conclusion (dominant decision class). Although it seems to be difficult to find all the neighbors of an irregularly shaped sector S_L in n-dimensional space, but using the introduced labeling system we can reduce it to search of all the sectors with label that differ with L in 1 bit. Uniting cells before *strength test* can help to accept single rules, that are not strong enough.

Because more than one sector can produce the same conclusion, finally, the decision rule can be defined as follows:

$$\text{IF } (s \in S_{L1}) \text{ OR } (s \in S_{L2}) \text{ OR } \ldots \quad \text{THEN } (s \in p) \tag{9}$$

where *L1, L2* – decimal labels of the sectors.



***Discovered rules*:**

1. IF ($s \in S_0$) OR ($s \in S_2$) THEN ($s \in$ *class 1*)
2. IF ($s \in S_4$) OR ($s \in S_6$) THEN ($s \in$ *class 2*)
3. IF ($s \in S_3$) OR ($s \in S_7$) THEN ($s \in$ *class 3*)

Fig. 4. An example of rule discovery in 2-dimentional feature space discretized nonorthogonally.

4. Experiment

The research aims to compare prediction accuracy of the rule sets discovered through orthogonal and nonorthogonal discretization of the feature space. In order to estimate and rate the rule set quality there were introduced:

1. *efficiency* $E \in [0,1]$

$$E = \prod_{p} \left(\frac{TP(p)}{TP(p) + FN(p)} \cdot \frac{TN(p)}{TN(p) + FP(p)} \right) \tag{10}$$

2. *predictive accuracy* $A \in [0,1]$

$$A = \sum_{p} TP(p) \tag{11}$$

where:
- p – the index of decision class (predicted class)
- TP(p) – *true positives* – the number of cases that belong to p and are covered by the rule predicting class p;
- FP(p) – *false positives* – the number of cases covered by the rule predicting class p but belonging to different class;
- FN(p) – *false negatives* – the number of cases that belong to p but are not covered by the rule predicting p;
- TN(p) – *true negatives* – the number of cases that are not covered by the rule predicting p and that do not belong to class p [4].

In order to test the new approach to discretization there has been performed classification process for 3 datasets [3] that differ in both number of instances, attributes and decision classes (main characteristics are summarized in table 1). For further research, all attributes in datasets were normalized.

dataset name	***#cases***	***#continuous attributes***	***#decision classes***	***multiple correlation with decision***
Wines	178	13	3	0.9487
Wisconsin breast cancer	683	9	2	0.9183
Iris	150	4	3	0.9646

Table 1. Datasets used in the experiment (from repository [3])

Searching for the cuttings system that produces the most efficient and accurate rules for current feature space (in case of both orthogonal and nonorthogonal discretization) was performed randomly, supported by roulette wheel selection. Accessible domain of each A_{ijk} was divided into 20 equal bins (0.05 long). After random selection of the cuttings, the discovered rule set was rated. Next, the probability of drawing the same bin for A_{ijk} was increased in proportion to E.

These steps have been repeated 10000 times for the training set. Then using the winning system of cuttings C_{ij} (with the highest efficiency), predictive accuracy of the rules for testing set was calculated. The procedure has been repeated 15 times. The population was randomly split anew into training (50%) and testing (50%) set after each iteration.

The simulation has been performed for 3 datasets, with the same simulation parameters, such as confidence threshold = 0.75, strength threshold = 0.2 and with varying number of cuttings – 1 or 2 cuttings for each attribute x_i, and the third option – 1 cutting for each of selected features (with the highest correlation with decision).

5. Results and conclusions

The results of the experiment show that the method of the heuristic search – the roulette wheel selection appears to be effective in such classification tasks. The proposed amount of drawings (10000) in most cases was more than enough to find an efficient system of cuttings for rule discovery (table 2).

dataset name	***#cuttings***		***mean E (training set)***		***mean A [%] (training set)***		***mean A [%] (testing set)***	
			orth.	***nonorth.***	***orth.***	***nonorth.***	***orth.***	***nonorth.***
Wines	26	2 for each x_i	0.1714	0.7908	56.20	93.98	34.19	87.31
	13	1 for each x_i	0.8478	0.7973	95.00	94.44	78.60	87.22
	6	$x_1\ x_6\ x_7\ x_8\ x_{11}\ x_{12}$	0.9079	0.8660	97.32	96.30	83.33	88.54
Wisc. breast cancer	18	2 for each x_i	0.5227	0.9150	77.53	96.78	39.93	91.30
	9	1 for each x_i	0.9448	0.9516	98.37	98.46	93.21	95.06
	3	$x_2\ x_3\ x_6$	0.8832	0.9469	96.49	98.32	95.28	96.51
Iris	8	2 for each x_i	0.9730	0.9894	99.37	99.76	91.85	92.59
	4	1 for each x_i	0.9060	0.9802	97.67	99.52	90.78	94.67
	2	$x_3\ x_4$	0.8514	0.9030	96.51	97.78	92.41	95.37

Table 2. Experiment results – efficiency and predictive accuracy for 3 datasets

In the table 2 we can observe that for nonorthogonal discretization, the obtained mean E value is more stable in each dataset (for different cutting sets) and in most cases higher than for orthogonal approach. Moreover, the nonorthogonal discretization gives better predictive accuracy in general, especially for testing set. Also it is noticeable that predictive accuracy is the highest, while taking least numerous cuttings sets (fig. 5). We can explain it with:

- better generalization – the less cuttings, the less risk of overfitting,
- stronger rules – the less cuttings, the less sectors (in general),
- more space covered by rules – the less cuttings, the more probably the sectors would be more capacious.

Additionally, the roulette wheel selection, used in the experiment, has to optimize less variables A_{ij} when there is less cuttings considered. And also, we can assume that the orthogonal cutting sets were found closer to the optimal through 10000 drawings than the nonorthogonal cuttings (*n-1* variables A_{ijk} more for each single cutting).

Therefore, there is a high probability that using more convergent methods we could prove the greater advantage of the nonorthogonal over the orthogonal discretization.

It is also noticeable that the decrease of the predictive accuracy (training set to testing set) seems to be proportional to the number of cuttings. We can observe similar, but even more significant effect in results of the orthogonal discretization. At the same time they are visibly worse (fig. 5) – particularly, the obtained predictive accuracy was especially poor for the orthogonal discretization, when there were taken 2 cuttings for each feature. The reason seems to be a significant number of rejected rules that were not strong enough. Hence, one should consider the fact, that when each conditional attribute is normally discretized (separately), there is used usually more than one cutting (if there are more than 2 decision classes). The research results seem to undermine the advisability of this approach for some datasets (Wines, Wisconsin).

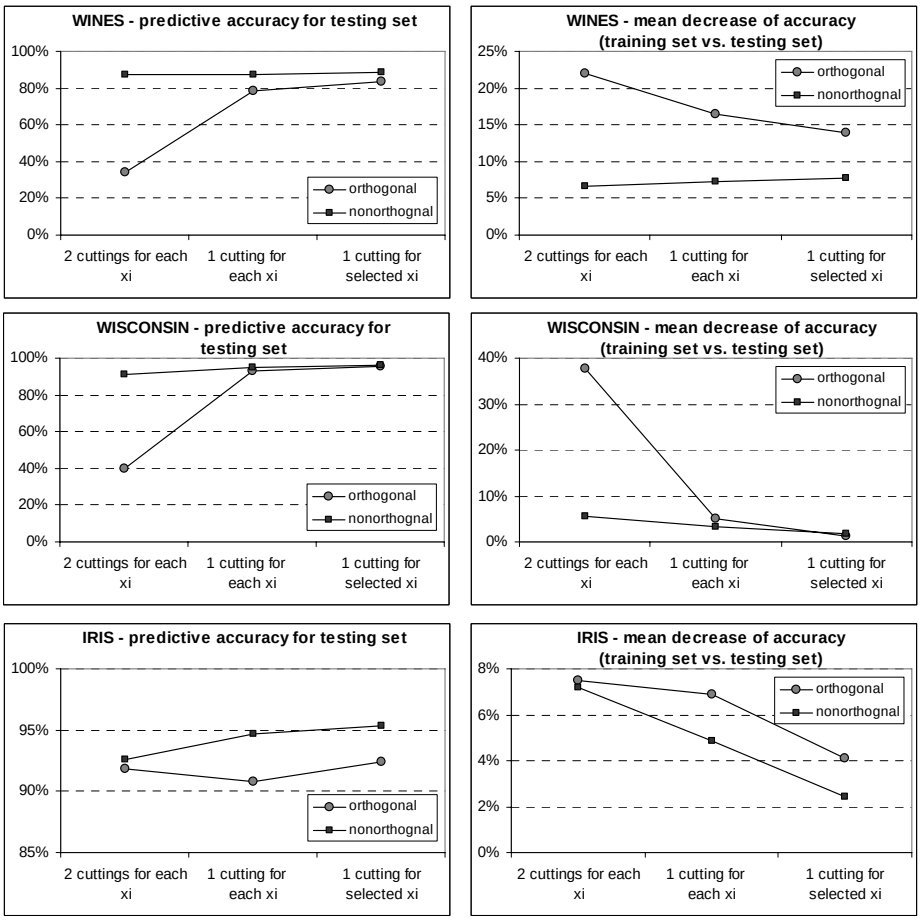


Fig. 5. Experiment results – predictive accuracy for testing set and mean decrease of accuracy – training set to testing set, for 3 datasets

It is worth to consider the fact that if one takes less number of cuttings than features for orthogonal discretization (there are no cuttings for a part of attributes), it automatically means reduction of space dimensionality. Nonorthogonal cuttings are still able to use information from 'unconsidered' features, because they still operate on full attribute space.

6. Summary

On the basis of the very promising experiment results, the introduced method of nonorthogonal discretization in multidimensional feature space provides higher prediction accuracy and generalization of discovered rules than traditional orthogonal cuttings. In the greater part of analyzed cases the mentioned quality measures were significantly higher for the nonorthogonal cuttings. Moreover, increased number of degrees of freedom for each cutting (n instead of 1) gives wider range of possible cuttings manipulation and better ability to fit spatial distribution of the data (due to irregular cells). Therefore, it can be especially efficient in systems of adaptive classification, when rule set has to adapt in response to changes in dataset.

The research revealed that nonorthogonal discretization is especially useful when decision classes are not monotonically distributed in domains of particular conditional attributes (dataset Wines). Additionally, there can be used less number of cuttings than features, what does not mean dimensionality reduction (unlike in case of orthogonal discretization).

Of course, for some classification tasks, data distribution can indicate that traditional cutting system would be more efficient. However, orthogonal discretization can be considered as a special case of nonorthogonal discretization – in such situation we can observe that:

$$A_{ij1} \cong A_{ij2} \cong \ldots \cong A_{ijn} \qquad (\mathbf{12})$$

Summing up, the proposed approach appears to be an effective solution for feature space discretization and rule discovery.

7. References

[1] Dougherty J., Kohavi R., Sahami M., K., *Supervised and unsupervised discretization of continuous features*, Machine Learning: Proceedings of the Twelfth International Conference, Morgan Kaufmann Publishers, San Francisco, 1995

[2] Grąbczewski K., *A separation criterion applied to classification rules induction based on databases (polish)*, Phd Thesis, IBS PAN, Warszawa, 2003

[3] Newman D.J., Hettich S., Blade C.L., Merz, C.J., *UCI Repository of machine learning databases*, Department of Information and Computer Science, University of California, Irvine 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html

[4] Parpinelli R.S., Lopes H.S., Freitas A. A., *An Ant Colony Algorithm for Classification Rule Discovery*, CEFET-PR, Curitiba, 2002.

[5] Wiliński A., Samborska A., *About convergention of symptom space discretization systems in medical diagnostics support (polish)*, Metody informatyki stosowanej w technice i technologii, Wydział Informatyki, Politechnika Szczecińska, nr 8, Szczecin 2005