ISSN: 1803-3814

CONTEXT OUT CLASSIFIER

Radek Hrebik[1*], Jaromir Kukal[1]

[1]Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Department of Software Engineering
Brehova 7, 115 19 Prague 1
Czech Republic

Abstract: *Novel context out learning approach is discussed as possibility of using simple classifiers which is background of hidden class system. There are two ways how to perform final classification. Having a lot of hidden classes we can build their unions using binary optimization task. Resulting system has the best possible sensitivity over all output classes. Another way is to perform second level linear classification as referential approach. The presented techniques are demonstrated on traditional iris flower task.*

Keywords: *classification, binary programming, cluster union, imperfect learning*

1 Introduction

Multidimensional statistical methods represent traditional tools for decision support and pattern classification, e.g., multi-step decision making processes and hierarchical systems. The paper deals with alternative approach to two-step decision processes that is based on the solution of optimization tasks. The first step of a deterministic decision process can be imperfect, but the second step is designed to enhance the perfectness of the whole system. The aim of this research is to design a new method of context out classification with embedded optimal unioning of hidden classes.

2 Context Out Learning

2.1 Motivation

In the case of classification we are well motivated to define the final (output) classes because of their close connection to the solution of the problem. The cardinal question is why to define and use another classes called hidden classes. There is a good analogy with ore mining. The formulation of mining problem is clear. It is necessary to separate the material into two classes: the ore and the residual material. But for a large stone the task is too complex. First, it is necessary to break it into small pieces as symbols of hidden classes and then carefully sort them into two output classes: the ore and the rest by using an effective procedure. Although technical aspects of separation are also useful, they are not discussed here. We will focus on the decision whether given piece of stone is or is not ore. Using majority rule, the pieces with more than fifty per cent of ore belong to the first class (ore). It is a traditional approach but with very low efficiency in general. For example the concentration of gold is very low and therefore no piece of stone would be selected for future gold extraction. This paradox can be easily solved by using cluster sensitivity and critical sensitivity which help to construct more sophisticated strategies not only for ore mining but mainly for general classification of patterns [6]. A general classification task distributes m patterns into N classes, and our method is based on preprocessing which places them into H hidden classes. Cluster analysis is a good example of formation of hidden classes. There are many other approaches to performing hidden classification using various kinds of local classifiers, and they are generally imperfect [5]. Our approach is based on basic characteristics of classification quality which are frequently used in many applications: accuracy [8], class sensitivity [2] and critical sensitivity [7].

2.2 Imperfect Linear Classifiers

Any hierarchical classifier consists of one hidden layer at least. There are many possibilities of designing such hidden layer. We prefer the parallel system of imperfect classifiers. The imperfect classifier is any system which classifies the patterns to given output classes but with low efficiency. Having a lot of such classifiers we are able to use them for alternative pattern description and forming of hidden classes. As in real life we can focus only on several cases and particular features. The imperfect classifier can be learned in any traditional

way of perfect classification but only using several patterns and several properties. This approach is called *context out learning* here.

Having m patterns from N classes where each pattern is a vector from $\mathbb{R}^n$, we can define context out classifier as follows: having original pattern set $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{y}^* \in \{1,...,N\}^m$, we select K patterns and L properties randomly to obtain context out sample $\mathbf{X}_{\mathrm{CO}} \in \mathbb{R}^{K\times L}$, $\mathbf{y}^*_{\mathrm{CO}} \in \{1,...,N\}^K$. Therefore, we plan to design context out classifier which makes only imperfect response related to original pattern set. Any traditional classification principle can be used to obtain its response $\mathbf{y} \in \{1,...,N\}^m$ for the complete set.

This imperfect classification procedure can be repeated G-times using Monte Carlo approach. Resulting responses $\mathbf{y}_1,...,\mathbf{y}_G$ form matrix $\mathbf{Y} \in \{1,...,N\}^{m\times G}$ where i-th row represents i-th pattern as row vector $\mathbf{r}_i = y_{i,1},...,y_{i,G}$ of imperfect memberships. Only unique columns are saved of course. There are N^G possible values of $\mathbf{r}_i$ but we will focus on unique rows. Their number is constrained as $1 \le H \le \min(m, N^G)$. Obviously, the unique rows represent the hidden classes.

2.2.1 Max Margin Context Out Classifier

The abilities of context out classification can be demonstrated as direct application of max margin classifier [1]. Traditionally, this classifier is applied only for $N = 2$ using bipolar notation as follows: $\mathbf{X}_{\mathrm{CO}} \in \mathbb{R}^{K\times L}$, $\mathbf{y}^*_{\mathrm{CO}} \in \{\pm 1\}^K$ where $+1$ represents $\mathscr{C}_1$ and -1 represents $\mathscr{C}_2$. The bipolar response is driven by formula

$$y = \mathrm{sign}(w_0 + \sum_{j=1}^{L} w_j x_{CO,j}) \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{n+1}$ is unknown vector satisfying

$$(w_0 + \sum_{j=1}^{L} w_j x_{CO,i,j}) y^*_{CO,i} \ge 1 \tag{2}$$

for all $i = 1,...,K$ as separability conditions. Using maximum margin condition

$$\sum_{j=1}^{L} w_j^2 = \min, \tag{3}$$

the optimum weights are the solution of quadratic optimization task.

The max margin classifier is applicable only to linear separable pattern set. When the context out pattern set does not satisfy this condition we select another one randomly.

2.3 Optimal Class Unions

Novel formulation of cluster unioning is based on following notation. The pattern set $\mathscr{S} = \{d_1,...,d_m\}$ is partitioned into a disjoint system of non-empty classes $\mathscr{C}_i \subset \mathscr{S}$ for $i = 1,...,N$. The disjoint system of hidden non-empty groups $\mathscr{H}_j \subset \mathscr{S}$ for $j = 1,...,H$ is also known as the result of imperfect classification. The relation between the classes and the hidden groups is presented via the contingency table $\mathbb{F} \in \mathbb{N}_0^{N\times H}$, where $f_{i,j} = \mathrm{card}\{k : d_k \in \mathscr{C}_i \bigcap \mathscr{H}_j\}$ is the result of pattern counting. Here, $f_{i,j}$ is the number of patterns belonging to both classes $\mathscr{C}_i$ and groups $\mathscr{H}_j$ as joint frequency, which can be relativized as

$$q_{i,j} = \frac{f_{i,j}}{\sum_{k=1}^{H} f_{i,k}}, \tag{4}$$

where $i = 1,...,N$, $j = 1,...,H$.

The hidden classes should be useful in final classification as discussed in previous section. But it is necessary to design the algorithms which will transform the hidden classes into the output ones. The deterministic approach is based on assumption that every hidden class belongs to just one output class. This fact is a kind of mapping which is easy to represent by binary matrix where every column consists of just one unit.

The novel deterministic approach is based on the relationship between hidden groups and the output classes. Our aim is to optimize this relationship as the best mapping from hidden to output classes. Here, strict classifier is defined as mapping

$$\mathrm{c} : \mathscr{L}_H \to \mathscr{L}_N$$

from the set $\mathscr{L}_H$ of hidden class indices to the set $\mathscr{L}_N$ of final class indices, where $\mathscr{L}_n = \{1,...,n\}$. This mapping can be expressed via the matrix $\mathbb{X} \in \{0,1\}^{N\times H}$, where $x_{i,j} = 1$ iff $d_k \in \mathscr{H}_j \Rightarrow d_k \in \mathscr{C}_i$. Therefore, $x_{i,j} = 1$ just when for any pattern belonging to $\mathscr{H}_j$ it also belongs to $\mathscr{C}_i$. The uniqueness conditions $\sum_{i=1}^{N} x_{i,j} = 1$ have to be satisfied for $j = 1,...,H$. There are many quantitative measures for classification efficiency.

First, the accuracy of classification can be expressed as

$$acc = \frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{H} f_{i,j}x_{i,j} \tag{5}$$

and this will be the subject of maximization.

Using the concept of class sensitivity as a relative frequency of true classification, we can calculate it by the equation

$$se_i = \sum_{j=1}^{H} q_{i,j}x_{i,j} \tag{6}$$

for $i = 1,...,N$.

Average sensitivity can be defined as

$$ase = \frac{1}{N}\sum_{i=1}^{N} se_i. \tag{7}$$

The lower estimate of class sensitivity is defined as critical sensitivity

$$se^* = \min\{se_i : i = 1,...,N\}. \tag{8}$$

We can now formulate several linear programming tasks related to optimum classifier design using the planning matrix $\mathbb{X} \in \{0,1\}^{N\times H}$, where $x_{i,j} = 1$ indicates $\mathscr{H}_j$ as a part of $\mathscr{C}_i$.

In the case of class equity, we can use a minimax approach and maximize critical sensitivity (8). An adequate non-linear optimization task is

$$se^* = \min\{se_i : i = 1,...,N\} = \max \tag{9}$$

subject to

$$\sum_{i=1}^{N} x_{i,j} = 1 \text{ for } j = 1,...,N, \tag{10}$$

$$x_{i,j} \in \{0,1\} \text{ for } i = 1,...,N,\ j = 1,...,H. \tag{11}$$

This task can be converted into a linear optimization task

$$se^+ = \max \tag{12}$$

subject to

$$\sum_{i=1}^{N} x_{i,j} = 1 \text{ for } j = 1,...,H, \tag{13}$$

$$\sum_{j=1}^{H} q_{i,j}x_{i,j} - se^+ \geq 0 \text{ for } i = 1,...,N, \tag{14}$$

$$x_{i,j} \in \{0,1\} \text{ for } i = 1,...,N,\ j = 1,...,H, \tag{15}$$

$$se^+ \in [0,1]. \tag{16}$$

Here se^+ is not only value of objective function (9) but also artificial variable (16). The inequalities (14) guarantee that se^+ is a lower bound of critical sensitivity se^* during optimization process and $se^+ = se^*$ in the optimum point. This approach is preferred in the experimental part of our study.

2.4 Optimal Linear Classifier

Using form of imperfect response matrix $\mathbf{Y}$ but respecting class inseparability we can design referential classifier using only linear regression between the hidden layer and classifier output. To avoid over-fitting effect when $G > m$ we have to apply ridge regression [4, 9] with parameter $\lambda > 0$ and calculate final weights as

$$\mathbf{v} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y} + \lambda\mathbf{I})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{y}^*, \tag{17}$$

we can apply max margin classifier again and obtain the optimal weights between the hidden layer and classifier output. This approach is not preferred but will be used as referential one.

3 Numerical Experiments

The imperfect learning technique is of stochastic nature. Its theoretical properties can not be investigated theoretically in general case due to large variety of classification tasks. Therefore, we have to demonstrate the efficiency of novel method on two examples. Basic principle of imperfect classification and its properties is demonstrated in the first case. The second example is focused on traditional problem of iris flower classification as simple real task.

3.1 2D Classifier

The context out classification approach with embedded max-margin method followed by union of hidden classes has been tested on artificial example. It is focused on the case of separable dataset but with non-linear decision rule. Therefore, the separation rule exists but it is non-linear. This testing example will demonstrate the efficiency of imperfect linear classifier for the solution of non-linear classification task.

In our testing example the points close to the the origin form the first class, i.e. $y = +1$ for $\|\mathbf{x}\|_2 \leq \rho$, while the far points form the second class and therefore $y = -1$ for $\|\mathbf{x}\|_2 \geq R$, where $0 < \rho < R$. The patterns with $\|\mathbf{x}\|_2 \in (\rho, R)$ are prohibited in order to avoid class mixing. The coordinates of individual patterns are generated randomly as $\mathbf{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ with eliminated prohibited patterns.

We generate 100 two-dimensional patterns using $\rho = 1$ and $R = 1.1$. We design ten random context out classifiers based on five samples and two properties as a model of complete pattern description in the first case. The results of imperfect learning are depicted in Figure 1a as switching lines of imperfect classifier. Any plane segment consisting of at least one pattern represents unique hidden class. There are 25 hidden classes in Figure 1a. Using maximization of critical sensitivity, we obtain $se_1 = 0.9500$, $se_2 = 0.9000$ and $se^* = 0.9000$. The referential ridge regression leads to $se_1 = 0.9300$, $se_2 = 0.9100$ and $se^* = 0.9100$ as better solution.

Using 20 context out classifiers to previous case, we obtain the better classification systems with $H = 39$, $se_1 = 0.9750$, $se_2 = 0.9667$ and $se^* = 0.9667$, which is depicted in Figure 1b. But the ridge regression offers $se_1 = 0.9750$, $se_2 = 0.9333$ and $se^* = 0.9333$.

The experiments get more interesting in the case of $L = 1$ when the pattern dimensionality is also reduced. When $m = 100$, $K = 5$, $L = 1$, $G = 10$, we obtain $H = 17$, $se_1 = 0.7750$, $se_2 = 0.7667$ and the imperfect classifiers are depicted in Figure 1c. The ridge regression comes to the same result.

When the number of generation is increased to $G = 20$, we obtain $H = 34$, $se_1 = 0.9250$, $se_2 = 0.9167$ and $se^* = 0.9167$, and the imperfect classifiers are depicted in Figure 1d. But $se_1 = 0.9000$, $se_2 = 0.9167$ and $se^* = 0.9000$ for ridge regression.

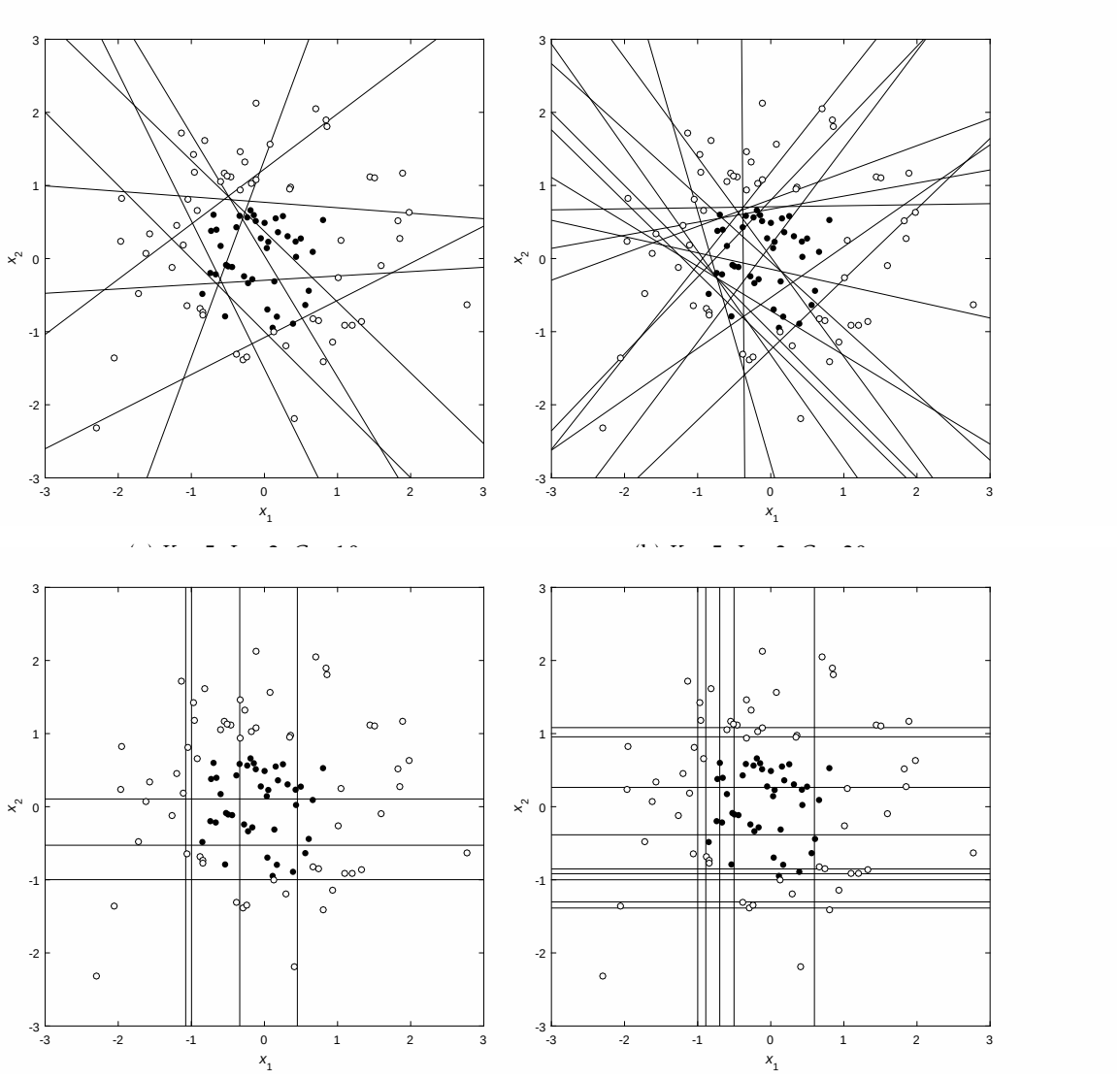
Using higher values of G, we obtain systems with $se^* = 1.0000$ in previous two cases. But the number of classifiers and hidden classes is too large and resulting system has a low practical importance. The combination of context out learning with imperfect classifiers and hidden class unions is also suitable for solving the linear inseparable problems and comparable with ridge regression as referential method.

3.2 Iris Flower Classifier

The novel classifier with imperfect hidden layer followed by class unioning or ridge regression has been applied to traditional iris flower task. As generally known, the first class of iris flowers (Iris setosa) is lineary separable against the rest classes. But we can demonstrate the real abilities of our system in the case of Iris versicolor and Iris virginica separations. The Iris flower task [3] consists of 150 four dimensional patterns. We used $K = 10$, $L = 4$ and $G = 20$. The results of one-to-all classifications are summarised in table 1. As in previous artificial example the novel method of imperfect learning seems to be better than the referential one. related to critical sensitivity se^* the ridge regression after imperfect learning (IRR) has less efficiency than optimal unioning (IOU) in case of Iris flower classifications. But direct application of ridge regression to original four dimensional data (DRR) is not recommended.

Table 1: Imperfect learning and optimal union (IOU), ridge regression (IRR) and ridge regression on original data (DRR)

	Iris versicolor			Iris virginica		
	IOU	IRR	DRR	IOU	IRR	DRR
se_1	0.96	0.92	0.86	1.00	0.98	0.93
se_2	0.97	0.93	0.48	0.99	0.93	0.92
se^*	0.96	0.92	0.48	0.99	0.93	0.92



(c) $K = 5$, $L = 1$, $G = 10$ (d) $K = 5$, $L = 1$, $G = 20$

Figure 1: Context Out Learning for Different Parameters

4 Conclusions

Novel method of imperfect classification has been developed. It is based on stochastic method of imperfect hidden layer forming. Both optimal unioning of hidden classes and ridge regression are able to offer final decision with high critical sensitivity. But the efficiency of this approach depends on the number of Monte Carlo generations. Therefore, the number of hidden classes can be unacceptable large in more complicated case.

Acknowledgement: This paper is supported by SGS17/196/OHK4/3T/14 grant of Czech Technical University in Prague.

References

[1] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152. COLT '92, New York, NY, USA (1992)

[2] Chang, L., Slikker, W.: Neurotoxicology: Approaches and Methods. Elsevier Science (1995)

[3] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)

[4] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

[5] Hrebik, R., Kukal, J.: Application of kohonen som learning in crisis prediction. In: Mathematical Methods in Economics. pp. 254–258. University of Hradec Kralove, Hradec Kralove (2017)

[6] Hrebik, R., Kukal, J., Jablonsky, J.: Optimal unions of hidden classes. Central European Journal of Operations Research pp. 1–17 (2017)

[7] Novakova, K.: Application of Transforms in Object Recognition (in Czech). Ph.D. thesis, FNSPE, CTU in Prague (2008)

[8] Taylor, J.: An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. A series of books in physics, University Science Books (1997)

[9] Vittinghoff, E., Glidden, D.V., Shiboski, S.C., McCulloch, C.E.: Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Springer Science & Business Media (2011)