FUZZY TEMPLATES FOR PAIR-WISE MULTI-CLASS CLASSIFICATION

Rimantas Kybartas
Vilnius University, Naugarduko st. 24, Vilnius, Lithuania

Keywords: Fuzzy Templates, Single layer perceptron, Support vectors, Pair-wise classifiers, Multi-class classification.

Abstract: The paper analyzes pair-wise classifier fusion by Fuzzy Templates. Single layer perceptron and support vectors are used as pair-wise classifiers. Weakness and strength of this fusion method are analyzed. Experimental results and theoretical considerations show that in some cases such an approach could be competitive or even outperform other pair-wise classifier fusion methods.

1 INTRODUCTION

Multi-class classification is applied in many fields, but there is no universal method which performs best in all cases. The shortcoming of complex solutions such as Neural networks, e.g. (Bishop, 1995), is the fact that proper parameter selection is time consuming. As an alternative there are classification methods based on fixed architecture. One of them is the two stage pair-wise classifier fusion method. It was shown in (Raudys et al., 2010) that such classification strategy is much more promising than the use of a single complex multi-class classifier. Moreover, such approach assures the possibility of taking pair-wise misclassification costs into account.

The superiority of the fusion of pair-wise classifiers to multiple-class classifiers is in that the former ones are more lightweight and less prone to adapt to training data while exploiting the useful statistical properties of pairs of classes.

Some kind of voting rules (Platt et al., 2000), probability based methods (Haste and Tibshirani, 1998) or any other fusion methods e.g. (Krzysko and Wolynski, 2009) are used in the second stage of the pair-wise classifiers. In (Kuncheva et al., 1998) the Fuzzy Templates method, where decision templates are used for fusion outputs of multi-class classifiers, was proposed. The interest in decision templates still remains. E.g. the decision templates extended by additional neural network layer are analyzed in (Haghigi et al., 2011). Thus there is a need to explore the possibilities to use decision templates in fusion of pair-wise classifiers.

The paper is organized as follows. The Fuzzy Templates approach is presented in Section 2. Section 3 discusses weaknesses and strengths of such pair-wise fusion method. The results on real world data are presented in Section 4. Section 5 contains conclusions, discussion and suggestions for future research.

2 FUZZY TEMPLATES

The aim of the original Fuzzy Templates method (Kuncheva et al., 1998) is to fuse continuous outputs of several K-category classifiers. In Pair-wise Fuzzy Templates (PWFT) L=K(K-1)/2 pair-wise classifiers are used. The fuzzy template vector for class Π_i is $F_i = \{ f_i(l) \}$ with K-1 attributes, where

$$f_i(l) = \frac{\sum_{z=1}^{N_k} C_{m,n}(\boldsymbol{x}_z)}{N_k}\text{, for all } l=1..K\text{-}1, \tag{1}$$

$\{x_z\}$ is crisply labelled training data, N_k is the number of training vectors in class Π_k, and $C_{m,n}(x_z)$ is the output of the pair-wise classifier with either $m=i$ or $n=i$.

When a new vector x_z is submitted for classification, its decision profiles (vectors) for each class $DP_i(x_z) = \{C_{m,n}(x_z)\}$ are calculated.

Final decision making is made according to $max(S(F_i,DP_i(x_z)))$, where

Kybartas R. (2012).
FUZZY TEMPLATES FOR PAIR-WISE MULTI-CLASS CLASSIFICATION.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 564-567
DOI: 10.5220/0003766505640567

$$S(F_i, DP_i(\mathbf{x}_z)) = 1 - \frac{1}{K-1}\sum_{l=1}^{K-1}(f_i(l) - DP_{i,l}(\mathbf{x}_z))^2 \tag{2}$$

Single layer perceptron (SLP) and Support vector classifier (SVC) were used as pair-wise classifiers in this research.

SLP may be expressed as $f(w^T x+b)$, where w and b are, correspondingly, the weight vector and bias obtained during perceptron training, x is a p-dimensional data vector, and f is the output activation function. In this study the sigmoid activation function was used:

$$f(x) = 1/(1+e^{-x}) \tag{3}$$

While training nonlinear single layer perceptron by gradient descent approach (Bishop, 1995), one may obtain seven well known statistical classifiers (Raudys, 1998) which are optimal for some particular data sets: Euclidean distance classifier, linear regularized discriminant analysis, standard linear Fisher classifier or the Fisher classifier with a pseudo-inverse of covariance matrix, robust discriminant analysis and minimum empirical error or maximal margin (i.e. support vector) classifiers.

SVC (Boser et al., 1992) was chosen as another pair-wise classifier. Since SLP is linear method in the context of classification tasks, linear SVC was selected. This paper proposes to use modified SVC outputs by applying sigmoid function (3) before providing them to the Pair-wise Fuzzy Templates.

3 WEAKNESS AND STRENGTH OF PAIR-WISE FUZZY TEMPLATES METHOD

PWFT should work effectively when outputs of pair-wise classifiers are highly diverse for different pairs of classes. For some particular vector x from class Π_i the pair-wise classifiers of different classes may produce the same results thus confusing the fusion rule. PWFT method should be avoided when using data sets which allow such situations. With SLP as the pair-wise classifier, this situation occurs when all the classes are arranged in a similar manner which makes SLP training stop in the same learning phases, thus getting approximately the same weights.

Let's examine two examples of artificially generated 2-dimensional data sets. Five data classes having the same covariance matrixes and arranged in a symmetric manner were generated (see Figure 1.a) for the first one. In this situation the Fisher discriminant function should optimally separate all pairs of classes. In the second data set five data classes having different covariance matrixes and arranged in random manner (see Figure 1.b) were generated. In this situation different statistical classifiers should be optimal for different pairs.

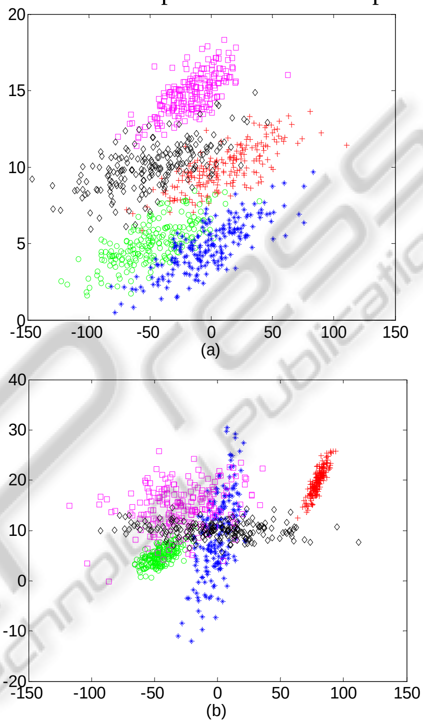

Figure 1: Five two dimensional Gaussian classes a) plotted in symmetric manner with the same covariance matrices; b) plotted in random manner with different covariance matrixes.

Table 1: Results of the two generated data sets with various fusion methods (best of them marked as bold).

Method/Pairwise classifier	Data (a)		Data (b)	
	SLP	SVC	SLP	SVC
PWFT	0.294	0.119	**0.193**	**0.175**
H-T	**0.115**	**0.117**	0.210	0.194
Voting	0.121	0.118	0.257	0.237
DAG	0.121	0.118	0.266	0.230

The results of experiments using different pair-wise classifier fusion rules (see Section 4) are shown in Table 1. With dataset (a), PWFT method performs very badly compared to other fusion rules when SLP is used as a pair-wise classifier. This is because all pair-wise classifiers produce almost the same output values. The situation is much better when SVC with outputs modified by eq. (3) is used as a pair-wise classifier instead of SLP. The attribute values of fuzzy template vectors were scattered in a rather narrow interval, but it did not affect the decision of SVC. Such an effect was gained due to the application of proposed modification using sigmoid function (3) for widely scattered SVC output values.

The data set (b) provides a much more favourable situation for PWFT method with SLPs as pair-wise classifiers (PWFT+SLP) because different pair-wise classifiers were obtained during pair-wise SLP training processes. PWFT with modified output SVC as the pair-wise classifier (PWFT+SVC) also showed considerable improvement.

Now let's introduce a new parameter α which is used as parameter multiplier in equation (3), i.e. $f(\alpha x)$. This way the range of input values is stretched, but the range of output values is narrowed. When employed in forming the values of PWFT+SLP fuzzy template vectors, the results are better due to the wider range of input values. Even in unfavourable cases, with proper α, PWFT+SLP method performance is not worse than the performance of standard voting methods. E.g. if the parameter α=50 is used, the PWFT+SLP method with data set (a) (Figure 2.a) results in error rate of 0.121 – the same as *voting* and *DAG* methods.

4 RESULTS WITH REAL WORLD DATA

Iris, *Letter*, *Satimage* and *Wine* data sets were taken from UCI machine learning repository (Frank and Asuncion, 2010). The *Wheat*, *Yeast* and *Chromosomes* data sets were donated by other researchers. All data sets were normalized and rotated according to their eigenvalues (Raudys, 2001) before training.

In order to optimally stop SLP and get the regularization parameter *C* for SVC, validation data obtained from training by noise injection technique (Skurichina et al., 2000) was used. The regularization parameter *C* for SVC was selected using grid search. The weights in SVC class weighting were set inversely proportional to the sample count of each class. The *LIBSVM* library (Chang and Lin, 2001) was used for SVC classifiers realization.

Three pair-wise classifier fusion methods were used for benchmark comparison.

Hastie-Tibshirani (H-T) method **(**Hastie and Tibshirani, 1998) is probability estimation based fusion method which uses Kulback-Leibler distance for conditional probability $r_{ij}=Prob(i|i \text{ or } j)$ estimation.

The voting rule performs the allocation of *K*(*K*-1)/2-dimensional vector formed by the first stage classifiers according to the majority of class labels in this vector. This method is also known as "Max Wins" method.

The directed acyclic graph (DAG) method is also known as DAGSVM (Platt et al., 2000). It organizes pair-wise SVCs in rooted binary direct acyclic graph to make final decision. The pair-wise SLPs were also used for experiments, instead of SVCs used in original paper.

Table 2: Comparison of four rules designed to fuse outputs of pair-wise classifiers based on SLP and SVC. The values indicate the average of incorrectly classified data used for testing. The last column shows pessimistic inaccuracy of the best method error rate.

Base	PWFT	H-T	Voting	DAG	Inacc.
Chromosomes					
SVC	*.204*/.262	.204	.210	.221	.004
SLP	*.193*/.194	**.192**	.193	.196	
Iris					
SVC	*.038*/.060	.037	.038	.038	.013
SLP	.036/*.032*	**.027**	.038	.039	
Letter					
SVC	**.153**/.282	.157	.153	.167	.008
SLP	.172/*.170*	.175	.173	.180	
Satimage					
SVC	*.148*/.167	.147	.151	.152	.005
SLP	**.141**/.142	.147	.143	.143	
Wheat					
SVC	*.073*/.090	.072	.074	.074	.012
SLP	.226/.070	**.063**	.070	.072	
Wine					
SVC	*.032*/.035	.032	.032	.031	.011
SLP	.095/*.031*	**.024**	.032	.033	
Yeast					
SVC	*.146*/.164	.147	.148	.149	.011
SLP	.149/*.142*	**.130**	.141	.145	

In order to get proper results, 500 experiments were performed for each data set by using half of randomly permuted data for training and the other half for testing in each experiment. Due to limited computational resources and large number of classes with rather large amount of data vectors, only 50 such experiments were performed for *Chromosomes* and *Letter* data sets. The same optimal newly introduced parameter α of equation (3) for each pair-wise classifier was selected from some predefined set by using validation data.

The obtained average results of such experiments grouped by data sets are presented in Table 2. The values in the cells mean the rate of incorrect classification of testing data. The cells in the second column of the table contain two values. The first value in the SVC row represents results of the case when modification with sigmoid function (3) was used, and the second value – when pure outputs of SVC were used. In the SLP row the first value means results where no parameter α was used in (3) and the second value – the result with selected optimal α. The best result in the PWFT column is written in italic, while the best one within all the methods for the data is marked in bold. The last column in Table 2 shows the inaccuracy of the best method.

Despite the differences between "the best" and other methods in Table 2 being statistically insignificant, it may be concluded that in some cases the newly presented method may successfully compete with others.

5 CONCLUSIONS AND DISCUSSIONS

It was shown that due to the inclination of SLP to obtain a set of classical statistical classifiers during its training evolution, the PWFT+SLP classifier is an effective method when the data classes are scattered in such a way that the classifiers of diverse complexities are needed for different pairs.

When using SVC as a pair-wise classifier, the modification with sigmoid function (3) for outputs has to be done. Experimental results showed that both, the SVC and SLP based pair-wise classifiers may be used as proper pair-wise classifiers in PWFT fusion method. Therefore there still remains room for investigation of PWFT method suitability according to the number of classes, its dimensionalities and other statistical characteristics.

A new parameter α was proposed for PWFT+SLP error rate improvement. Preliminary results showed that in situations unfavourable for the PWFT+SLP method, the use of proper scaling parameter α could allow it to perform at least with the same error rate as voting methods. Employment of some particular procedure of selecting optimal α has to be investigated more deeply, because improper α selection may make results worse, and of course, selection of a proper value may considerably improve overall classification results.

ACKNOWLEDGEMENTS

This paper is part of research financed by Research Council of Lithuania (contract number MIP-043/2011).

REFERENCES

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Oxford Univ. Press.

Boser, B., Guyon, I., Vapnik, V., 1992. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fith Annual Workshop on Computational Learning Theory*, pages 144-152. ACM Press.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines; *http://www.csie.ntu.edu.tw/~cjlin/libsvm*.

Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science *http://www.ics.uci.edu/~mlearn/MLRepository.html*.

Haghigi, M. S., Vahedian, A., Yazdi, H. S., 2011. Extended Decision Template Presentation for Combining Classifiers. *Expert Systems with Applications 38, p*ages. 8414-8418.

Hastie, T. and Tibshirani, R., 1998. Classification by pairwise coupling. *The Annals of Statistics*, 26(1): 451-471.

Krzysko, M., Wolynski, W., 2009. New variants of pairwise classification. *European Journal of Operational Research* (EOR) 199(2):512-519.

Kuncheva, l., Bezdek, J. C., Sutton, M. A., 1998. On Combining Multiple Classifiers by Fuzzy Templates, *Proceedings of the 1998 Annual Meeting of the North American Fuzzy Information Processing Society*, 193 – 197.

Platt J. C., Cristianini N., and Shawe-Taylor J., 2000. Large margin DAG's for multi-class classification. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol. 12 pp. 547-553.

Raudys, S., 1998. Evolution and generalization of a single neuron. I. SLP as seven statistical classifiers. *Neural Networks* 11: 283–96.

Raudys, S., 2001. Statistical and Neural Classifiers: An integrated approach to design. Springer-Verlag, NY.

Raudys., S, Kybartas R., Zavadskas, E. K., 2010. Multicategory Nets of Single-layer perceptrons: Complexity and Sample Size Issues, *IEEE Transactions on Neural Networks*, vol. 2, no. 5, 784 – 795.

Skurichina, M., Raudys, S., Duin, R. P. W, 2000. K-NN directed noise injection in multilayer perceptron training, *IEEE Trans. on Neural Networks*, 11(2): 504–511.