ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Class-Specific Feature Selection for One-Against-All Multiclass SVMs

Gaël de Lannoy and Damien François and Michel Verleysen

Université catholique de Louvain
Institute of Information and Communication Technologies,
Electronics and Applied Mathematics
Machine Learning Group
Place du Levant, 3 Louvain-la-Neuve, Belgium

Abstract. This paper proposes a method to perform class-specific feature selection in multiclass support vector machines addressed with the one-against-all strategy. The main issue arises at the final step of the classification process, where binary classifier outputs must be compared one against another to elect the winning class. This comparison may be biased towards one specific class when the binary classifiers are built on distinct feature subsets. This paper proposes a normalization of the binary classifiers outputs that allows fair comparisons in such cases.

1 Introduction

Many supervised classification tasks in a wide variety of domains involve multiclass targets. One frequently used and easy method for solving these problems is to train several off-the-shelf binary support vector machines (SVMs) classifiers and to extend their decision to multiclass targets by using the one-against-one (OAO) or the one-against-all (OAA) approaches. A vast literature exists on the pros and cons of these two approaches, and a comprehensive review can be found for example in [1] and [2].

In the OAA approach, the output value of each competing classifier is used in the decision rule rather than the thresholded class prediction as in the OAO approach. The problem with this OAA decision rule is that every classifier participating to the decision is assumed equally reliable, which is rarely the case. This problem has previously been adressed in [3] where a classifier reliability measure is included in the OAA decision process; experiments show that the performances are improved.

Nevertheless, despite the interesting performance increase, one major drawback of this reliability measure is that the competing classifiers must be trained on the same feature sets to keep the output values comparable. However, the optimal feature subsets might be different for each one-against-all sub-problem, and it is known that spurious features can harm the classifier – even if the latter is able to prune out features intrinsically [4]. In such situations, the feature selection step should rather be made where the training of the model actually happens, and so at the class level rather than at the multiclass level.

In this work, we show how such reliability measure can be modified to overcome this limitation, and therefore allow the feature selection to be made at -

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

and optimized for - the binary classifier level where the training actually happens. The following of this paper is organized as follows. Section 2 provides a short overview of the theoretical background on the methods used in this work. Section 3 introduces the classifier reliability measure and shows how this measure can be included in the OAA decision. Section 4 describes the experiments and the results.

2 One-against-all strategy for multiclass SVMs

SVMs are linear machines that rely on a preprocessing to represent the feature vectors in a higher dimension, typically much higher than the original feature space. With an appropriate non-linear mapping $\varphi(\mathbf{x})$ to a sufficiently high-dimensional space, finite data from two categories can always be separated by a hyperplane. In SVMs, this hyperplane is chosen as the one with the largest margin. SVMs have originally been designed for binary classification tasks [5]. This two-class formulation of SVMs where $y_i \in \{-1, 1\}$ can be extended to solve multiclass problems where $y_i \in \{1, 2, \ldots, M\}$ by constructing M binary classifiers, each classifier being trained with the examples of one class with a positive label and all the other samples with a negative label.

Let $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ be a set of n training samples where $\mathbf{x}_i \in R^p$ is a p-dimensional feature vector and $y_i \in \{-1, 1\}$ is the associated binary class label. In SVMs, the jth classifier yields the following decision function:

$$f_j(\mathbf{x}) = \mathbf{w}_j^T \varphi(\mathbf{x}) + b_j \tag{1}$$

where $\mathbf{w}_j$ and b_j are the parameters of the hyperplane obtained during the training of the jth classifier. Geometrically, $f_j(\mathbf{x})$ corresponds to the distance between $\mathbf{x}$ and the functional margin of the classifier j. At the classification phase, a new observation is then assigned to the class $j*$ which produces the largest output value amongst the M classifiers:

$$j* = \arg \max_{j=1\ldots M} f_j(\mathbf{x}) = \arg \max_{j=1\ldots M} \mathbf{w}_j^T \varphi(\mathbf{x}) + b_j. \tag{2}$$

3 Improving the OAA decision

One major drawback of the OAA approach for solving multiclass problems is that the classifier generating the highest value from its decision function is selected as the winner without considering the reliability of each classifier. The two underlying assumptions behind this approach are first that the classifiers are equally reliable, and second that they have been constructed on the same features. This section first recalls Liu and Zheng's reliability measure [3] associated to a SVM classifier that overcomes the first assumption. Second, we show that this measure can be improved to permit the use of distinct feature sets for each binary classifier. Finally, an improved decision rule for the OAA approach based on the reliability measure is given.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

3.1 Reliability measure

To overcome the first assumption, one would obviously consider the output of a classifier more reliable if the true generalization error $R = E[y \neq sign(f(\mathbf{x})]$ is small. Unfortunately, this value is always unknown and must be estimated from data by the empirical error $\widetilde{R} = 1 - \frac{1}{n}\sum_{i=1}^{n}[y_i = sign(f(\mathbf{x}_i))]$. However, when the number of training samples is relatively small compared to the number of features, it has been shown that a small empirical $\widetilde{R}$ does not guarantee a small R [6].

For this reason, a better classifier reliability measure can be based on an upper bound of R. Indeed, minimizing the SVM objective function has been shown to also minimizing an upper bound on the true generalization error R [6]. Following this idea, the following reliability measure λ has been proposed by [3]:

$$\lambda = \exp\left(-\frac{\frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+}{Cn}\right), \tag{3}$$

where $(z)_+ = z$ if $z > 0$ and 0 otherwise. The Cn denominator is included to cancel the effect of different training sizes and regularization parameter value. In the linearly separable case, the λ reliability measure associated to a classifier is large if its geometrical margin $\frac{2}{\|\mathbf{w}\|^2}$ is large.

3.2 Reliability measure with distinct feature sets

By removing most irrelevant and redundant features from the data, feature selection helps improving the performance of learning models by alleviating the effect of the curse of dimensionality, enhancing generalization capability, speeding up learning process and improving model interpretability. In the OAA approach, one classifier is built to discriminate each class against all the others. Each feature can however have a different discriminative power for each of the binary classifiers and useless features can harm the classifier even if it is able to adapt its weights accordingly during the learning process [4]. This situation is known to happen for example in the classification of heart beats where it has been observed that the duration between successive heart beats discriminates for some cardiac pathologies while it is the morphology of the heart beats that discriminate for other cardiac pathologies [7]. In such situations, the selection of features should thus rather be made at the class level rather than at the global level. Nevertheless, building each classifier in a distinct feature space would make the comparison of the output values unreliable.

To alleviate the effect of dealing with distinct numbers of features, a weighting by the cardinality of $\mathbf{w}$ is inserted into Eq. (3):

$$\beta = \exp\big(-\frac{\frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+}{Cn\left\|\mathbf{w}\right\|_0}\big). \tag{4}$$

The effect of the cardinality is to normalize the squared Euclidean norm of $\mathbf{w}$ with respect to the dimension of the space in which it lives, i.e. the size of the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

selected feature subset. This kind of normalization is rather common in tools aimed at missing data analysis [8].

3.3 Improved OAA decision rule

Assume M classifiers have been trained, each on an optimal subset of features. The reliability measure β_j is also computed for each of the classifiers. Given a new sample $\mathbf{x}$, f_j is evaluated for $1 \leq j \leq M$ according to Eq. (1) and a soft decision function $z_j \in \{-1, 1\}$ is generated:

$$z_j = \text{sign}(f_j(\mathbf{x}))(1 - \exp(-|f_j(\mathbf{x})|)). \tag{5}$$

The output of each classifier is then weighted by the associated reliability measure and the sample $\mathbf{x}$ is assigned to the class j^* according to:

$$j^* = \arg \max_{j=1\ldots M} z_j(\mathbf{x})\beta_j. \tag{6}$$

The weighting of the output of each classifier by its associated β measure penalizes classifiers with a small margin and a poor generalization ability, and also allows every competing classifier to have distinct features, distinct meta-parameters and a distinct number of observations.

4 Experimental results

The proposed weighted OAA decision rule is experimented on three multiclass datasets from the UCI repository[1]. The details of the three datasets are shown in Tab. 1. Five methods are compared in the experiments:

1. OAA without feature selection;
2. OAA with global feature selection;
3. λ-weighted OAA without feature selection;
4. λ-weighted OAA with class-wise feature selection;
5. β normalized OAA with class-wise feature selection.

The selection of features is achieved using a permutation test with the mutual information criterion in a naive ranking approach [9]. The RBF kernel is used in the SVM classifier. The regularization parameter and kernel parameter are optimized on the training set using a 5-fold cross-validation over a wide range of values, and the performances are evaluated on the test set.

The classification error for the five methods are presented in Table 2 together with the percentage of selected features. When the feature selection is achieved at the class level, the average feature selection percentage is reported. The results surprisingly show that the weighting by the λ reliability measure does not

[1] http://archive.ics.uci.edu/ml/

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Name	Training	Test	Classes	Features
Segmentation	210	2100	7	19
Vehicle	676	170	4	18
Isolet	3119	1559	26	617

Table 1: Number of samples, classes and features of the three datasets used in the experiments. For the isolet dataset, only a subsample (50%) of the original training data has been considered.

		Segmentation		Vehicle		Isolet	
Weighting	Selection	Error	Features	Error	Features	Error	Features
none	none	10.0%	100%	17.7%	100%	4.5%	100%
none	global	8.4%	78%	17.7%	100%	4.5%	100%
λ	none	9.8%	100%	18.8%	100%	9.5%	100%
λ	class	8.8%	65%	20.0%	93.0%	7.5%	78%
β	class	6.9%	65%	17.0%	93.0%	3.9%	78%

Table 2: Comparison of the classification error for the five methods. The percentage of selected features is also reported.

always improves the classification performances. However, the best results are achieved by the β weighting and the class-wise feature selection. In particular, the results obtained with the class-wise feature selection and β weighting are better than with the λ weighting and the same class-wise feature selection. This shows the need to include the so-called zero-norm of $\mathbf{w}$ in the computation of the reliability measure when a distinct subset of features are used in each classifier. Furthermore, the results obtained with the class-wise feature selection and the β weighting are better than with the global feature selection. This confirms the benefit from the class level feature selection over the global feature selection.

5 Conclusion

Most methods for multiclass classification assume that there is an optimal subset of features that is common to all classes, while in many applications, it may not be the case. In the one-against-all approach, using distinct feature subsets for each class might however lead to unfair and biased final decision rules. To alleviate this problem, the output of the competing classifiers should be normalized before being compared. The normalization that is proposed in this paper takes into account the number of features used and a measure of reliability of the classifier. On three standard benchmark datasets, the proposed approach, used in conjunction with support vector machines, yields better results than selecting features across all classes.

The class-dependent feature selection methodology allows increasing the performances compared with a feature selection common to all classes. It further-

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

more brings insights about the relationships between the features and the specific target classes.

Acknowledgments

Gaël de Lannoy is funded by a F.R.I.A grant. Computations have been run on the Lemaitre cluster thanks to the "Calcul Intensif et Stockage de Masse" (CISM) of the Université catholique de Louvain.

References

[1] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[2] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[3] Yi Liu and Y.F. Zheng. One-against-all multi-class svm classification using reliability measures. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 2, pages 849 – 854 vol. 2, 31 2005.

[4] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[5] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 1 edition, March 2000.

[6] V. N. Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[7] K.S. Park, B.H. Cho, D.H. Lee, S.H. Song, J.S. Lee, Y.J. Chee, I.Y. Kim, and S.I. Kim. Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function. In *Computers in Cardiology, 2008*, pages 229–232, Sept. 2008.

[8] Pedro J. García-Laencina, José-Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal, and Michel Verleysen. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomput.*, 72(7-9):1483–1493, 2009.

[9] Damien François, Fabrice Rossi, Vincent Wertz, and Michel Verleysen. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing, Elsevier*, 70:1276–1288, 2007.