Probabilistic Multi-classifier by SVMs from voting rule to voting features

Anh Phuc TRINH, David BUFFONI, Patrick GALLINARI*

*Laboratoire d'Informatique de Paris 6
104, avenue du Président Kennedy, 75016 Paris.
{anh-phuc.trinh,david.buffoni,patrick.gallinari}@lip6.fr,

1 Probabilistic multi-classifier by SVMs

Definition of the posterior probabilities for multiclass problem

*Let $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ be a set of m training examples. We assume that each example $\mathbf{x}_i$ is drawn from a domain $X \in \mathbb{R}^n$ and each class y_i is an integer from the set $Y = \{1, \ldots, k\}$ with $k > 2$. The posterior probabilities of multiclass problem is a conditional probability of each class $y \in Y$ given an instance **x***

$$P(y = i|\mathbf{x}) = p_i \tag{1}$$

subject to

$$\sum_{i=1}^{k} p_i = 1 \; p_i > 0 \; \forall i \tag{2}$$

There are two approaches, either one-vs-one or one-vs-rest, in solving the multi-class problem by SVMs. Following the setting of the one-vs-one approach, we have the voting method proposed by (Tax, 2002) using decision values $f_{ij}(\mathbf{x})$ of SVMs to estimate the posterior probabilities. Another method of (Wu T-F, 2004) obtains p_i from the pairwise probability of (Platt, 2000).

2 From voting rule to voting features

Definition of the voting features

Suppose that $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ is the set of m training examples drawn from an independent and identical distribution. A voting feature representation $\Theta : C(f_{ij}(\mathbf{x})) \times Y \rightarrow \mathbb{B}^d$ is a function Θ that maps a configuration of decision values $c(f_{ij}(\mathbf{x})) \subset C(f_{ij}(\mathbf{x}))$ and a class $y_i \in Y$ to a d-dimensional feature vector, thus the set of voting features is denoted by $\mathbb{VF}$.

The posterior probabilities definied on the set of voting features $\mathbb{VF}$ $p_i = P(y = i|\mathbf{x}, \lambda) = \frac{exp(\sum_{l=1}^{d} \lambda_l \times \Theta_l(\mathbf{x}, y=i))}{\sum_{y=1}^{k} exp(\sum_{l=1}^{d} \lambda_l \times \Theta_l(\mathbf{x}, y))}$ is estimated in maximizing the logarithm of the conditional likelihood (Nigam et McCallum, 1999) and is solved by unconstrained optimization problem.

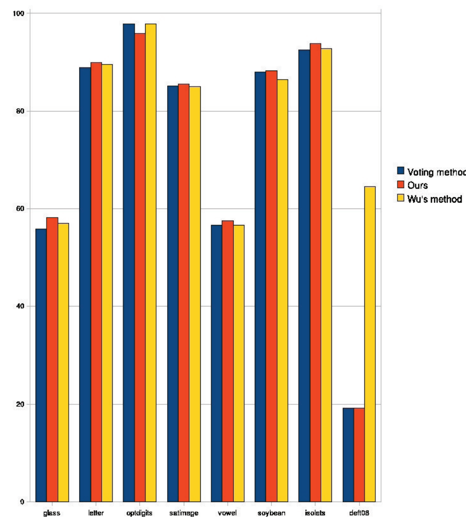

FIG. 1 – *Accuracy rates of three different methods on seven UCI and deft08 test datasets are obtained using the polynominal kernel ; The voting rule, our and Wu's methods are figured respectively by violet, red and yellow columns*

3 Experiments

To compare the performance of our method with others, we selected seven datasets from the UCI learning data repository [1], and the DEFT08 dataset [2] .

Références

Nigam, K. J. L. et A. McCallum (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering 1*, 61–67.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers 14*, 61–74.

Tax, D. R. P. W. D. (2002). Using two-class classifiers for multiclass classification. *International Conference on Pattern Recognition 2*, 124– 127.

Wu T-F, Chih-Jen Lin, R. C. W. (2004). Probability estimates for multi-class classification by pairwise coupling. *International Conference on Pattern Recognition 5*, 975–1005.

[1] `http://mlearn.ics.uci.edu/MLSummary.html`
[2] `http://deft08.limsi.fr/`