PAC-Bayesian Majority Vote for Late Classifier Fusion

Emilie Morvant
Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France
firstname.name@lif.univ-mrs.fr

Stéphane Ayache
Univ. of St-Etienne, Lab. Hubert Curien, CNRS, UMR 5516, F-42000, St-Etienne, France
amaury.habrard@univ-st-etienne.fr

May 2, 2014

Abstract

A lot of attention has been devoted to multimedia indexing over the past few years. In the literature, we often consider two kinds of fusion schemes: The early fusion and the late fusion. In this paper we focus on late classifier fusion, where one combines the scores of each modality at the decision level. To tackle this problem, we investigate a recent and elegant well-founded quadratic program named MinCq coming from the Machine Learning PAC-Bayes theory. MinCq looks for the weighted combination, over a set of real-valued functions seen as voters, leading to the lowest misclassification rate, while making use of the voters’ diversity. We provide evidence that this method is naturally adapted to late fusion procedure. We propose an extension of MinCq by adding an order-preserving pairwise loss for ranking, helping to improve Mean Averaged Precision measure. We confirm the good behavior of the MinCq-based fusion approaches with experiments on a real image benchmark.

Keywords: Machine Learning, Multimedia fusion, Multi-modality search, Ranking and re-ranking

1 Introduction

Combining multimodal information is an important issue in Multimedia and a lot of research effort has been dedicated to this problem (see Atrey et al. (2010) for a survey). Indeed, the fusion of multimodal inputs can bring complementary information, from various sources, useful for improving the quality of any multimedia analysis method such as for semantic concept detection, audio-visual event detection, object tracking, etc. The different modalities correspond generally to a relevant set of features that can be grouped into different views. For example, classical visual or textual features commonly used in multimedia are based on TF-IDF, bag of words, texture, color, SIFT, spatio-temporal descriptors, etc. Once these features have been extracted, another step consists in using machine learning methods in order to build classifiers able to discriminate a given concept.

Two main schemes are generally considered Snoek et al. (2005). In the early fusion approach, all the available data/features are merged into one feature vector before the learning and classification steps. This can be seen as a unimodal classification. However, this kind of approach has to deal with heterogeneous data or features which are sometimes difficult to combine. The late fusion model works at the decision level by combining the prediction scores available for each modality. This is usually called multimodal classification or classifier fusion. Late fusion may not always outperform unimodal classification. Especially when one modality provides significantly better results than others or when one has to deal with imbalanced input features. However, late fusion scheme tends to give better results for learning semantic concepts in case of multimodal video Snoek et al. (2005). Several methods based on a fixed decision rule have been proposed for combining classifiers such as max, min, product, sum, etc. Kittler et al. (1998). Other approaches, often referred to as stacking Wolpert (1992), need of an extra learning step.

In this paper, we address the problem of late multimodal fusion at the decision level with stacking. Let $h_i$ be the classifier that gives the score associated with the $i^{th}$ modality for any instance $x$. A classical method consists in looking for a weighted linear combination of the different scores,

$$H(x) = \sum_{i=1}^{n} q_i h_i(x),$$
where \(q_i\) represents the weight associated with \(h_i\). It is usually required that \(0 \leq q_i \leq 1\) and \(\sum_{i=1}^{n} q_i = 1\). This linear weighting scheme can be seen as a majority vote. This approach is widely used because of its robustness, simplicity and scalability due to small computational costs [Ayache et al. 2010]. It is also more appropriate when there exist dependencies between the views [Wu et al. 2004]. An important issue is then to find an optimal way to combine the scores. One solution is to use machine learning methods to assess the weights [Ayache et al. 2010].

From a machine learning standpoint considering a set of classifiers with a high diversity is generally a desirable property [Dietterich 2000]. One illustration is given by the algorithm AdaBoost [Freund and Schapire 1996], frequently used as a multimodal fusion method. AdaBoost weights the classifiers according to different distributions of the training data, introducing some diversity, but requires at least weak classifiers to perform well. Another recent approach based on the portfolio theory [Wang and Kankanhalli 2010] proposes a fusion procedure trying to minimize some risks over the different modalities and a correlation measure. While it is well-founded, it needs to define some appropriate functions and is not completely fully adapted to the classifier fusion problem since it does not directly take into account the diversity between the outputs of the classifiers.

We propose to study a new machine learning method, namely MinCq, introduced in [Laviolette et al. 2011]. It proposes a quadratic program for learning a weighted majority vote over real-valued functions called voters (such as score functions of classifiers). The algorithm is based on the minimization of a generalization bound that takes into account both the risk of committing an error and the diversity of the voters, offering strong theoretical guarantees on the learned majority vote. In this article, our aim is to show the interest of this algorithm for classifier fusion. We provide evidence that MinCq is able to find good linear weightings but also very performing non-linear combination based on the portfolio theory. Wang and Kankanhalli (2010) proposes a fusion procedure trying to minimize some risks over the different modalities and a correlation measure. While it is well-founded, it needs to define some appropriate functions and is not completely fully adapted to the classifier fusion problem since it does not directly take into account the diversity between the outputs of the classifiers.

The paper is organized as follows. Section 2 deals with the theoretical framework of MinCq. We extend MinCq as a late fusion method in Section 3. Before concluding in Section 5, we evaluate empirically the MinCq late fusion in Section 4.

## 2 PAC-Bayesian MinCQ

In this section we present the algorithm MinCq of Laviolette et al. [Laviolette et al. 2011] for learning a \(Q\)-weighted majority vote of real-valued functions (e.g. classifier scores). This method is based on the PAC-Bayes theory [McAllester 1999]. We first recall the setting of MinCq.

We consider binary classification tasks over a feature space \(X \subseteq \mathbb{R}^d\) of dimension \(d\). The label space is \(Y = \{-1, 1\}\). The training sample is \(S = \{(x_i, y_i)\}_{i=1}^{n}\) where each example \((x_i, y_i)\) is drawn i.i.d. from a fixed — but unknown — probability distribution \(D\) defined over \(X \times Y\). We consider a space of real-valued voters \(\mathcal{H}\), such that \(\forall h_i \in \mathcal{H}, h_i : X \rightarrow \mathbb{R}\). Given a voter \(h_i\), the predicted label of \(x \in X\) is given by \(\text{sign}[h_i(x)]\), where \(\text{sign}[a] = 1\) if \(a \geq 0\) and \(-1\) otherwise. Then, the learner aims at choosing a distribution \(Q\) over \(\mathcal{H}\) — the weights \(q_i\) — leading to the \(Q\)-weighted majority vote \(B_Q\) with the lowest risk. \(B_Q\) is defined by,

\[
B_Q(x) = \text{sign}[H_Q(x)],
\]

\[
H_Q(x) = \sum_{i=1}^{|\mathcal{H}|} q_i h_i(x).
\]

The associated true risk \(R_D(B_Q)\) is defined as the probability that the majority vote misclassifies an example drawn according to \(D\),

\[
R_D(B_Q) = P_{(x,y) \sim D} (B_Q(x) \neq y).
\]

In the case of MinCq, \(\mathcal{H}\) has to be a finite auto-complemented family of \(2n\) real-valued voters \(\mathcal{H} = \{h_1, \ldots, h_{2n}\}\) such that,

\[
\forall x \in X, \forall i \in \{1, \ldots, n\}, h_{i+n}(x) = -h_i(x).
\]

Moreover, the algorithm considers quasi-uniform distributions \(Q\) over \(\mathcal{H}\), i.e. the sum of the weight of a voter and its opposite is \(\frac{1}{n}\),

\[
\forall i \in \{1, \ldots, n\}, Q(h_i) + Q(h_{i+n}) = q_i + q_{i+1} = \frac{1}{n}.
\]
Due to the auto-complemented (1) and quasi-uniformity (2) assumptions, the algorithm can be expressed as a linear combination of the empirical first and second moments of the distributions D over X × Y, i.e. distribution Q, on H and any distribution D over X × Y, if E_{(x,y) ~ D}[yH_Q(x)] > 0 then R_D(B_Q) ≤ C_Q.

Finally, the Q-weighted majority vote learned by MinCq is then

\[ B_Q(x) = \text{sign}[H_Q(x)], \]

with \( H_Q(x) = \sum_{i=1}^{n} \left( 2q_i - \frac{1}{n} \right) h_i(x). \)

### 3 MinCq as a Late Fusion Method

PAC-Bayesian MinCq has been proposed in the particular context of binary classification where the objective is to minimize the classification error of the Q-weighted majority vote by taking into account the diversity of the voters. From a multimedia indexing standpoint, MinCq thus appears to be a natural way for late classifiers fusion to combine the predictions of classifiers separately trained from different modalities.

Concretely, given a training sample of size 2nm, we split it randomly into two subsets \( S' \) and \( S = \{(x_j, y_j)\}_{j=1}^{nm} \) of the same size. Let n be the number of modalities. For each modality \( i \), we train a classifier \( h_i \) from \( S' \). Let
\( \mathcal{H} = \{h_1, \ldots, h_n, -h_1, \ldots, -h_n\} \) be the set of the \( n \) associated prediction functions and their opposites. At this step, the fusion is achieved by MinCq: We learn from \( S \) the \( Q \)-weighted majority vote over \( \mathcal{H} \) with the lowest risk. However, in many applications, such as multimedia document retrieval, people are interested in performance measures related to precision or recall. Since a low-error vote is not necessarily a good ranker, we propose an adaptation of MinCq to improve the popular Mean Averaged Precision (MAP).

We first recall the definition of the MAP measured on \( S \) for a given real-valued function \( h \). Let \( S^+ = \{(x_j, y_j) : (x_j, y_j) \in S \land y_j = 1\} \) be the set of the \( m^+ \) positive examples from \( S \) and \( S^- = \{(x_j, y_j) : (x_j, y_j) \in S \land y_j = -1\} \) be the set of the \( m^- \) negative examples from \( S \). For evaluating the MAP, one ranks the examples in descending order of the scores. The MAP of \( h \) is,

\[
\text{MAP}_S(h) = \frac{1}{|m^+|} \sum_{j : y_j = 1} \text{Prec}@j,
\]

where \( \text{Prec}@j \) is the percentage of positive examples in the top \( j \). The intuition behind this definition is that we prefer positive examples with a score higher than negative ones. To achieve this goal, we propose to learn with pairwise preference [Fünkranz and Hüllermeier (eds) (2010)] on pairs of positive-negative instances. Indeed, pairwise methods are known to be a good compromise between accuracy and more complex performance measure like MAP. Especially, the notion of order-preserving pairwise loss was introduced in [Zhang (2004)] in the context of multiclass classification. Following this idea, Yue et al. [Yue et al. (2007)] have proposed a SVM-based method with a hinge-loss relaxation of a MAP-loss. In our specific case of MinCq for multimedia fusion, we design an order-preserving pairwise loss for correctly ranking the positive examples. Actually, for each pair \((x_j^+, x_j^-) \in S^+ \times S^-\), we want: \( H_Q(x_j^+ > H_Q(x_j^-) \Leftrightarrow H_Q(x_j^-) - H_Q(x_j^+) < 0 \). This can be forced by minimizing (according to the weights \( q_i \)) the following hinge-loss relaxation of the previous equation,

\[
\frac{1}{m^+ m^-} \sum_{j^+ = 1}^{m^+} \sum_{j^- = 1}^{m^-} \left[ H_Q(x_{j^+}) - H_Q(x_{j^-}) \right]_+,
\]

where \([a]_+ = \max(a, 0)\) is the hinge-loss. In the setting of MinCq, with \( H \) auto-complemented (Eq.(1)) and \( Q \) quasi-uniform (Eq.(2)), we reduce the term (3) to,

\[
\frac{1}{m^+ m^-} \sum_{j^+ = 1}^{m^+} \sum_{j^- = 1}^{m^-} \left[ \sum_{i=1}^{n} \left(2q_i - \frac{1}{n}\right) \left(h_i(x_{j^+}) - h_i(x_{j^-})\right) \right]_+ ,
\]

To deal with the hinge-loss of (4), we consider \( m^+ \times m^- \) additional slack variables \( \xi_{s^+ \times s^-} = (\xi_{j^+ j^-})_{1 \leq j^+ \leq m^+, 1 \leq j^- \leq m^-} \) weighted by a parameter \( \beta > 0 \). We make a little abuse of notation to highlight the difference with (MinCq). Since \( \xi_{s^+ \times s^-} \) appear only in the linear term, we simply add (4) after the (MinCq) formulation. We obtain the quadratic program (MinCqPW),

\[
\arg\min_Q \xi_{s^+ \times s^-} \quad Q^T S Q - A^T S Q + \beta \text{Id}^T \xi_{s^+ \times s^-},
\]

s.t. \( m^+ Q = \frac{\mu}{2} + \frac{1}{2nm} \sum_{j=1}^{n} y_j h_i(x_j) \),

\[\forall j^+ \in \{1, \ldots, m^+\}, \forall j^- \in \{1, \ldots, m^-\}, \xi_{j^+ j^-} \geq 0,\]

\[\xi_{j^+ j^-} \geq \frac{1}{m^+ m^-} \sum_{i=1}^{n} \left(2q_i - \frac{1}{n}\right) \left(h_i(x_{j^+}) - h_i(x_{j^-})\right),\]

and \( \forall i \in \{1, \ldots, n\}, 0 \leq q_i \leq \frac{1}{n}, \)
In this section, we show empirically the interest of MinCq, and our extension, as a late fusion method with stacking parameters leading to the lowest risk, we select the ones leading to the best MAP.

Finally, for tuning the hyperparameters ($\mu$, $\beta$) we use a cross-validation process (CV). Instead of selecting the parameters leading to the lowest risk, we select the ones leading to the best MAP.

### 4 Experiments

In this section, we show empirically the interest of MinCq, and our extension, as a late fusion method with stacking (implemented with MOSEK solver). We experiment the MinCq-based approaches on the PascalVOC’07 benchmark [Everingham et al. (2007)], where the goal is a list of 20 visual concepts to identify in images. The corpus is constituted of 5000 training and 5000 test images. In general, the ratio between positive and negative examples is less than 10%. For each concept, we generate a training sample constituted of all the training positive examples and negative examples independently drawn such that the positive ratio is 10%.

Our objective is not to provide the best results on this benchmark but rather to evaluate if the MinCq-based methods could be helpful for the late fusion step in multimedia indexing. To do so, we split the training sample into two subsets, $S^+\text{ and } S^−$, of the same size. We consider 9 different visual features: 1 SIFT, 1 LBP, 1 Percepts, 2 HOG, 2 Local Color Histograms and 2 Color Moments. Then, we train from $S^+$ a SVM-classifier for each visual feature (with the LibSVM library [Chang and Lin (2001)] and a rbf kernel with parameters tuned by CV). The final classifier fusion is learned from $S$.

In a first series of experiments, the set of voters $H$ is constituted by the 9 SVM-classifiers (MinCq also considers the opposites). We compare the 3 linear MinCq methods ($\text{MinCq}$, $\text{MinCq}_{\text{PWF}}$, $\text{MinCq}_{\text{PWF}}$) to the following 4 baseline fusion approaches.
The best classifier of $H$:

$$h_{\text{best}} = \arg\max_{h_i \in H} MAP_S(h_i).$$

The one with the highest margin:

$$\text{best}(x) = \arg\max_{h_i \in H} |h_i(x)|.$$

The sum of the classifiers (unweighted vote):

$$\Sigma(x) = \sum_{h_i \in H} h_i(x).$$

The MAP-weighted vote:

$$\Sigma_{MAP}(x) = \sum_{h_i \in H} \frac{MAP_S(h_i)}{\sum_{h_{i'} \in H} MAP_S(h_{i'})} h_i(x).$$

In a second series, we propose to introduce non-linear information with a rbf kernel layer. We represent each example by the vector of its scores of the 9 SVM-classifiers, $H$ being the set of kernels over the sample $S$; Each $x \in S$ is seen as a voter $k_l(\cdot, x)$. We then compare our method to stacking with SVM tuned by CV (SVM$^{rbf}$). Note that we do not report the results of (MinCq) in this context, because the computational cost is much higher and the performance is lower. The full pairwise version implies too many variables which may penalize the resolution of (MinCq).

In either case, the hyperparameters of MinCq-based methods are tuned with a grid search by a 5-folds CV. The MAP-performances are reported on Table 1, we can make the following remarks.

- On the right, for the first experiments, we clearly see that the linear MinCq-based algorithms outperform on average the linear baselines. At least one MinCq-based method produces the highest MAP, except for “boat” for which $h_{\text{best}}$ is the best. We note that the order-preserving hinge-loss is not really helpful: The classical (MinCq) shows the best MAP. In fact, this can be explained by the limited number of voters.

- On the left, with a kernel layer, at least one MinCq-based method achieves the highest MAP and for $17/20$ both are better than SVM. Moreover, MinCq$_{\text{rbf}^{P\text{Vav}}}$ with the averaged pairwise preference is the best for 17 concepts, showing the order-preserving loss is a good compromise between improving the MAP and keeping a reasonable computational cost.

- Globally, kernel-based MinCq methods outperform the other methods. Moreover, at least one MinCq-based approach is the best for each concept showing PAC-Bayesian MinCq is a good alternative for late classifiers fusion.

5 Conclusion

We propose in this paper to make use of a well-founded learning quadratic program called MinCq as a novel multimedia late fusion method. PAC-Bayesian MinCq was originally developed for binary classification and aims at minimizing the error rate of the weighted majority vote by considering the diversity of the voters [Laviolette et al. (2011)]. In the context of multimedia indexing, we claim that MinCq thus appears naturally appropriate for late classifier fusion in order to combine the predictions of classifiers trained from different modalities. Our experiments show that MinCq is a very competitive alternative for classifier fusion. Moreover, the incorporation of average order-preserving constraints is sometimes able to improve the MAP-performance measure. Beyond these results, such PAC-Bayesian methods open the door to define other theoretically well-founded frameworks to design new algorithms in many multimedia tasks such as multi-modality indexing, multi-label classification, ranking, etc.

References

Atrey, P. K., Hossain, M. A., El-Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimedia Syst., 16(6):345–379.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines.
Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15.

Everingham, M., Van Gool, L., Williams, C. K. L., Winn, J., and Zisserman, A. (2007). The PASCAL visual object classes challenge 2007 (VOC2007) results.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proc. of ICML*, pages 148–156.

Fürnkranz, J. and Hüllermeier (eds), E. (2010). *Preference Learning*. Springer-Verlag.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *PAMI*, 20:226–239.

Laviolette, F., Marchand, M., and Roy, J.-F. (2011). From PAC-Bayes bounds to quadratic programs for majority votes. In *Proceedings of ICML*.

McAllester, D. A. (1999). PAC-bayesian model averaging. In *Proceedings of COLT*, pages 164–170.

Snoek, C., Worring, M., and Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of ACM Multimedia*, pages 399–402.

Wang, X. and Kankanhalli, M. S. (2010). Portfolio theory of multimedia fusion. In *Proceedings of ACM Multimedia*, pages 723–726.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Wu, Y., Chang, E. Y., Chang, K. C.-C., and Smith, J. R. (2004). Optimal multimodal fusion for multimedia data analysis. In *Proceedings of ACM Multimedia*, pages 572–579.

Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of SIGIR*, pages 271–278.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *JMLR*, 5:1225–1251.