Learning Concepts from Large-Scale Data Sets by Pairwise Coupling with Probabilistic Outputs

Feng Zhou and Bao-Liang Lu* *Senior Member, IEEE*

***Abstract*— This paper considers the problems of learning concepts from large-scale data sets. The way we take is completely classification algorithm independent. Firstly, the original problem is decomposed into a series of smaller two-class sub-problems which are easier to be solved. Secondly we present two principles, namely the shrink and expansion principles, to restore the global solution from the intermediate results learned from the sub-problems. In the theoretical analysis, this procedure of integration is described as a statistical inference of a posteriori probability and is degraded as the min-max principles in the special case considering 0-1 outputs. We also propose a revised approach which reduces the computational complexity of the training and testing stage to a linear level . Finally, experiments on both the synthetic and text-classification data are demonstrated. The results indicate that our methods are effective to large scale problems.**

I. Introduction

The problems we face today are often large-scale ones, which are difficult sometimes even impossible to handle directly. For instance, the computational cost of Support Vector Machines (SVMs) scales quadratically with respect to the number of examples. As the size of data grows, it is impossible to store the whole kernel matrix in the memory and the time required in training becomes unacceptable in many practical applications. Recently various methods have been proposed to overcome this computational inability. In [1], the author uses a mixture of SVMs, each of them trained only on a part of the data set. The samples would be reassigned according to the prediction. Another strategy is to eliminate non-support vectors early during the optimization process [2]. These methods could provide substantial saving in computation, however, they depend strongly on classification methods.

On the other hand, multi-class classification problems are usually decomposed into learning a series of two-class classifiers. One-against-all and one-against-one are two popular strategies of decomposition. As pointed in [3], the run-time complexity of the latter one is below that of the former one, although it casts a k-class problem into $\binom{k}{2}$ two-class subproblems and the number of modules increases. A common way to combine the pairwise comparison is by voting [4]. It assumes that the pairwise classifiers output 0-1 values and selects the class with the most winning two-class decisions. In most cases, however, a 0-1 coding for the outcome of pairwise classifier is not so reasonable as a probabilistic one which could express more complex relationship between instances and classes. In the past few years, several authors [5] [6] have proposed probability estimates by combing the pairwise class probabilities. The min-max-modular (M^3) network [7] differs from the above work that it further decomposes a two-class problem, which is still too large to fit into particular models, into smaller ones. The main advantage of M^3 is that the original problem could be handled as simple as we expect. Its effectiveness has been provided by tasks involving with multi-class and imbalanced data [8]. In the integration stage, it uses min-max principles to combine the outcome of classifiers. The voting policy has been compared with min-max principles in [9] that the former one achieve the highest accuracy, while the latter one earns the highest performance.

In this paper, we adopt a decomposition method similar to M^3 network to divide a large two-class data set. The main problem we concern is to obtain a global concept from intermediate ones, which are in probabilistic forms reported by classifiers. Then we give a theoretical justification of the procedure of integration which could be categorized into the shrink and expansion principles. Furthermore, we implement this procedure in a more efficient way while considering the over complete set of classifiers.

This paper is organized as follows. In Section II, M^3 network is introduced briefly. In Section III, we analyze the problem from a view of statistical learning, then propose two algorithms, and compare them with min-max principles respectively. Several experiments are presented in Section IV. Finally, conclusions of our work are outlined in Section V.

II. Min-Max Modular Network

Given a complex two-class problem, the positive and negative training sets could be described as follows:

$$\mathcal{T} = \mathcal{X}^+ \cup \mathcal{X}^-$$
$$\text{where } \mathcal{X}^+ = \{x_k^+, \omega^+\}_{k=1}^{l^+} \text{ and } \mathcal{X}^- = \{x_k^-, \omega^-\}_{k=1}^{l^-}$$

where $\{x_k^+, x_k^-\} \in \mathbb{R}^d$, $\{\omega^+, \omega^-\}$ are the class labels, and $\{l^+, l^-\}$ denote the number of samples.

The first step M^3 takes is to divide the training samples

Asterisk indicates corresponding author. This work was supported in part by the National Natural Science Foundation of China under the grants NSFC 60375022 and NSFC 60473040, and the Microsoft Laboratory for Intelligent Computing and Intelligent Systems of Shanghai Jiao Tong University. F. Zhou and B. L. Lu are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240 China. E-mail: {zhfe99; bllu}@sjtu.edu.cn.

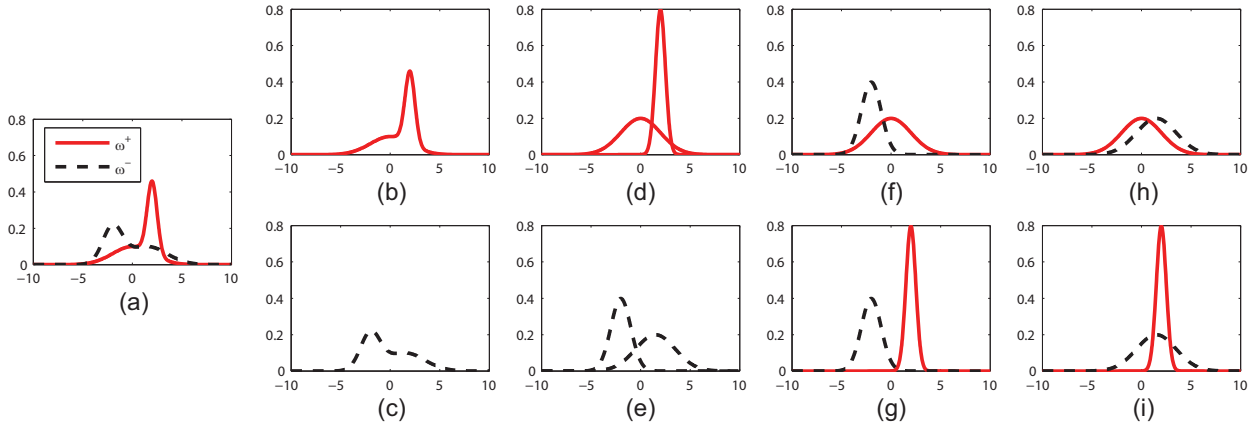

Fig. 1. The decomposition of two complex distributions. (a) mixture of two complex probability density functions. (b)(c) separation in positive and negative parts. (d)(e) decomposition into gaussian probability densities. (f)-(i) combination of a positive gaussian probability density and a negative one.

into a series of small subsets:

$$\mathcal{X}^+ = \bigcup_{i=1}^{n^+} \mathcal{X}_i^+, \qquad \mathcal{X}_i^+ = \{x_k^+, \omega_i^+\}_{k=1}^{n_i^+}$$

and $$\mathcal{X}^- = \bigcup_{i=1}^{n^-} \mathcal{X}_i^-, \qquad \mathcal{X}_i^- = \{x_k^-, \omega_i^-\}_{k=1}^{n_i^-}$$

in which, $\{n^+, n^-\}$ are the number of positive and negative subsets, respectively; $\{n_i^+, n_i^-\}$ are the sizes of subsets, and $\{\omega_i^+, \omega_i^-\}$ are the class labels for subsets. Instead of training a huge classifier, it is turned to train $n^+ \times n^-$ sub-classifiers. For each classifier c_{ij}, the training set is

$$\mathcal{T}_{ij} = \mathcal{X}_i^+ \cup \mathcal{X}_j^-$$

In the testing stage, the values reported by these small classifiers are integrated according to the min-max principles [7]. Suppose that a new instance x_0 is introduced, and all the outcomes of classifiers are arranged in a n^+-by-n^- matrix M, whose element M_{ij} comes from the one trained from $\mathcal{T}_{ij}$.

First in the Min step, the minimum value in each row is selected to represent the n^- classifiers in the same row. These n^+ values are the inputs in Max step, in which the maximum one is the winner. This two-layer process can be described here

$$\hat{q} = \max_{i=1}^{n^+} \hat{q}_i$$
$$\hat{q}_i = \min_{j=1}^{n^-} M_{ij} \quad \text{where } 0 \le M_{ij} \le 1 \tag{1}$$

III. The Integration of Probabilistic Outputs

A. A Statistical View

From the viewpoint of the statistical learning, the task of a classification problem is to deduce the underlying probability distribution from the limited available data. If all the relevant probabilities are known, we could decide its class label according to the well-known Bayesian decision theory [10].

Specifically, let $p(x|\omega^+)$ and $p(x|\omega^-)$ denote the class-conditional probability density functions for positive and negative parts respectively. The prior probability for positive and negative classes are $p(\omega^+)$ and $p(\omega^-)$. Then given a new instance x, the optimal class label is:

$$\widehat{\omega} = arg \max_{\omega=\{\omega^+,\omega^-\}} p(\omega|x)$$
$$\text{where } p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x|\omega^+)p(\omega^+) + p(x|\omega^-)p(\omega^-)}$$

In many cases (Figures 1. (a)-(c)), the distribution of samples is not easily learnable or the parametric form is quite complex. However, the large picture may be interpreted as the result of coincidental influence from some simpler distributions (Figures 1. (d)(e)). If the decomposition method is effective enough, the description of a small area is an easy task.

B. Shrink and Expansion Principle

After decomposition on the training samples, it is possible to design classifiers based on the combinations of one positive subset and one negative subset (Figures 1 (f)-(i)).

Thus, each outcome M_{ij} predicted form the classifier c_{ij} could be inferred as the the probability of the event that the test sample x belongs to ω_i^+, relative to the hypothesis that it belongs to either ω_i^+ or ω_j^-

$$\begin{aligned} M_{ij} &= p(\omega_i^+ | x, \omega_i^+ \cup \omega_j^-) \\ &= \frac{p(x, \omega_i^+ \cap (\omega_i^+ \cup \omega_j^-))}{p(x, \omega_i^+ \cup \omega_j^-)} \\ &= \frac{p(x, \omega_i^+)}{p(x, \omega_i^+) + p(x, \omega_j^-)} \end{aligned} \tag{2}$$

At first, we calculate the probability that x belongs to ω_i^+

$$\begin{aligned} q_i(x) &= p(\omega_i^+|x, \omega_i^+ \cup \omega^-) \\ &= \frac{p(x, \omega_i^+)}{p(x, \omega_i^+) + p(x, \omega^-)} \quad (3) \\ &= \frac{p(x, \omega_i^+)}{p(x, \omega_i^+) + \sum_{j=1}^{n^-} p(x, \omega_j^-)} \\ &= \frac{1}{1 + \sum_{j=1}^{n^-} \frac{p(x, \omega_j^-)}{p(x, \omega_i^+)}} \quad (4) \end{aligned}$$

From Equation (2), we could get

$$\frac{p(x, \omega_j^-)}{p(x, \omega_i^+)} = \frac{1}{M_{ij}} - 1 \quad (5)$$

Substituting this into Equation (4), we obtain

$$q_i(x) = \frac{1}{\sum_{j=1}^{n^-} \frac{1}{M_{ij}} - (n^- - 1)} \quad (6)$$

Then we calculate the probability that x belongs to ω^+

$$\begin{aligned} q(x) &= p(\omega^+|x) \\ &= 1 - p(\omega^-|x, \omega^- \cup \omega^+) \end{aligned}$$

Refering to $\omega^+ = \cup_{i=1}^{n^+} \omega_i^+$ and Equation (4), we alternatively have

$$q(x) = 1 - \frac{1}{1 + \sum_{i=1}^{n^+} \frac{p(x, \omega_i^+)}{p(x, \omega^-)}} \quad (7)$$

Rewrite Equation (3)

$$\frac{p(x, \omega_i^+)}{p(x, \omega^-)} = \frac{1}{1 - q_i(x)} - 1$$

Substituting this into Equation (7), we finally get

$$q(x) = 1 - \frac{1}{\sum_{i=1}^{n^+} \frac{1}{1 - q_i(x)} - (n^+ - 1)} \quad (8)$$

We call Equation (6) as "Shrink Principle" and Equation (8) as "Expansion Principle". Based on these principles, we obtain our first integration method which is summarized in Algorithm 1.

C. Analysis of Algorithm I

1) Comparison with Min-Max Principles: Comparing Equation (1) with Equation (6), we can obtain the following equation

$$\hat{q}_i = q_i = M_{ij}$$

if and only if

$$M_{ik} = 1, \text{ for all } k \neq j$$

Alternatively, we have

$$\hat{q} = q = q_i$$

if and only if

$$q_k = 0, \text{ for all } k \neq i$$

Algorithm 1 A New Integration Method

Input:
Number of positive subsets: n^+
Number of negative subsets: n^-
Classifiers set: $\mathcal{C} = \{c_{ij}\}_{i=1,j=1}^{n^+,n^-}$
Test instance: x
Output: Probability $q(x) = p(\omega^+|x)$

```
Form matrix M ∈ R^{n+ × n-}, where M_ij = c_ij(x)
for i = 1 to n+ do
  if M_ij = 0, for some j = 1, ··· , n- then
    q_i(x) = 0
  else
    q_i(x) = 1 / (Σ_{j=1}^{n-} 1/M_ij − (n- − 1))
  end if
end for
if q_i = 1, for some i = 1, ··· , n+ then
  q(x) = 1
else
  q(x) = 1 − 1 / (Σ_{i=1}^{n+} 1/(1 − q_i) − (n+ − 1))
end if
```

The equations above show that both the minimization and maximization principles are special cases of the probabilistic outputs.

From Equation (6), we can see that q_i is only dependent on the M_{ij}s with the same i. Adding a new M_{ij} would cause the posterior $p(\omega_i^+|x)$ shrinking from the former value. This is also where the name "*shrink principle*" comes from. Moreover, the minimum M_{ij}^{min} has the greatest influence on the value of q_i. As the gap between the M_{ij}^{min} and other M_{ij}s is growing larger, it would get closer to q_i coincidentally. In the extreme case, if all other c_{ij}s except one believe x is a ω_i^+, we only need to follow the suggestion of the left one.

The above situation is quite similar to *expansion principle*. The effect of adding a new non-zero q_i would expand the value of q that means the evidence of x is a ω^+ is getting stronger.

Refering to the previous example (Figures 1), here we show the procedure in Figures 2. If the proportion of the positive samples in the area near x is not 100% in both the training set (Figures 2 (a) (c)), then the proportion must be smaller in the combining training set (Figures 2 (e)). If one of training set contains nearly nothing about the discriminating information ($x = 2$ in Figures 2 (b)), then the final result (Figures 2 (f)) is mainly decided by the other one (Figures 2 (d)).

Overall, the min-max principles choose the one which contains the most information of discrimination from a set of classifiers. This is similar to the idea of PCA which tries to find the most expressive subsets of features to describe the whole distribution of samples. On the contrary, shrink-expansion principles take information provided by all the classifiers, and assign different weights on them. If the noise happens rarely, the decision made under shrink-expansion

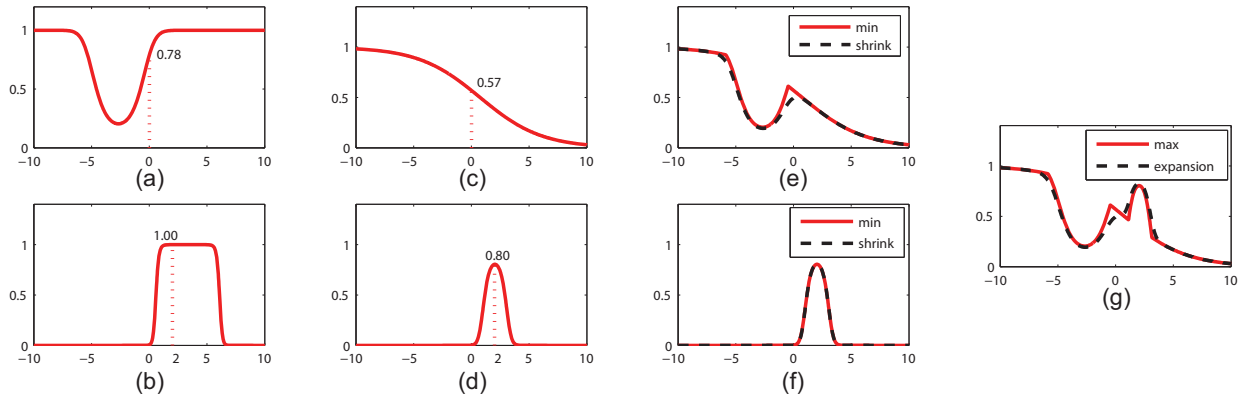

Fig. 2. The integration of the outcome from small classifiers. (a)-(d) outcomes of classifiers which output the partial posteriors corresponding to Figures 1 (f)-(i). (e) the integration of (a) and (b). (f) the integration of (c) and (d). (g) final outcome integrated from (e) and (f)

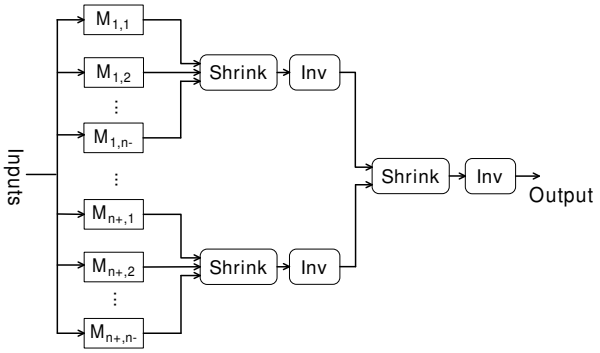

Fig. 3. The data flow in out system which take the partial posteriors matrix as inputs, and a posterior probability $p(\omega^+|x)$ comes out at the end.

principles would be optimal.

2) Implementation of the System: Closely investigating on the Equation (6) and Equation (8), we find there are mainly two operations involved

$$Shrink(x) = \frac{1}{\sum_{i=1}^{d} \frac{1}{x_i} - (d-1)}, \quad x \in \mathbb{R}^d$$
$$Inv(x) = 1 - x, \quad x \in \mathbb{R}$$

The expansion principle (Equation (8)) can be instead calculated with the combination of the above two operators. The structure of our system is designed in Figures 3.

3) Computational Complexity: It needs $O(n^+n^-)$ float operations in the shrink procedure. And in the next expansion step, it runs $O(n^+)$ times. Therefore, the overall computational time is bounded by the size of the partial posteriors matrix M, which is $O(n^+n^-)$.

4) Extension to Multi-class Problems: Until now, we have illustrated a method to solve large-scale two-class problems. For a multi-class problem, as pointed in the introduction, we could firstly reduce the original problem into a series of two-class sub-problems with various decomposition methods. Based on the results of these sub-problems, specific integration algorithm would proceed to restore the final result.

If the one-against-all strategy of decomposition is adopted, the class with the maximum possibility would be the winner in the integration stage. And in the one-against-one case, voting based on the possibilities from the underlying sub-problems would be carried out then. In addition, many probabilistic frameworks [5] [6] have been proposed to combine the probabilistic outputs. Actually, Equation (6) could be interpreted in another way: the ith positive sub-class plus other n^- negative sub-classes are $n^- + 1$ different classes, and it calculates the posterior of one class.

Thus we could easily adopt the above strategies to extend our method to multi-class problems. And the original complex problem would be finally solved in parallel simple ones.

D. Revised Shrink and Expansion Principles

Algorithm 1 is built on the $n^+ \times n^-$ binary classifiers which are actually not over complete. We could obtain additional classifiers trained from any two positive samples subsets or two negative ones.

Here we define that "homo-pairwise" classifier is trained from a pair of sample subsets with same type of class label. For instance, a positive homo-pairwise classifier c'_{ij} is trained form $\mathcal{X}_i^+ \cup \mathcal{X}_j^+$. Thus there are more $\binom{n^+}{2}$ positive homo-pairwise classifiers and $\binom{n^-}{2}$ negative ones we could obtain from the train data.

We calculate q in a different way

$$\begin{aligned} q &= p(\omega^+|x) \\ &= p(\omega^+|x, \omega^+ \cup \omega^-) \\ &= p(\cup_i \omega_i^+ | x, \cup_i \omega_i^+ \bigcup \cup_j \omega_j^-) \\ &= \frac{\sum_i p(\omega_i^+, x)}{\sum_i p(\omega_i^+, x) + \sum_j p(\omega_j^-, x)} \\ &= \frac{1 + \sum_{i \neq k} \frac{p(\omega_i^+, x)}{p(\omega_k^+, x)}}{1 + \sum_{i \neq k} \frac{p(\omega_i^+, x)}{p(\omega_k^+, x)} + \sum_j \frac{p(\omega_j^-, x)}{p(\omega_k^+, x)}} \end{aligned} \tag{9}$$

The outcome from c'_{ij} is

$$M'_{ij} = p(\omega_i^+ | x, \omega_i^+ \cup \omega_j^+)$$

Comparing with Equation (5), we get

$$\frac{p(x, \omega_j^+)}{p(x, \omega_i^+)} = \frac{1}{M'_{ij}} - 1$$

Substituting this and Equation (5) into Equation (9), we have

$$q = \frac{\sum_{i \neq k} \frac{1}{M'_{ki}} - (n^+ - 2)}{\sum_{i \neq k} \frac{1}{M'_{ki}} + \sum_j \frac{1}{M_{kj}} - (n^+ + n^- - 2)}$$

Based on the above equation, the second algorithm is designed in Algorithm 2.

Algorithm 2 A Revised Integration Method

Input:
Number of positive subsets: n^+
Number of negative subsets: n^-
Classifiers set 1: $\mathcal{C} = \{c_{kj}\}_{j=1}^{n^-}$
Classifiers set 2: $\mathcal{C}' = \{c'_{ki}\}_{i=1, i \neq k}^{n^+}$
Test instance: x
Output: Probability $q(x) = p(\omega^+ | x)$

Form the vector $M \in \mathbb{R}^{n^-}$, where $M_j = c_{kj}(x)$
Form the vector $M' \in \mathbb{R}^{n^+}$, where $M'_{i(i \neq k)} = c'_{ki}(x)$
if $M_j = 0$, for some $j = 1, \cdots, n^-$ **then**
  $q(x) = 0$
else if $M'_i = 0$, for some $i = 1, \cdots, k-1, k+1, \cdots, n^+$ **then**
  $q(x) = 1$
else

$$q(x) = \frac{\sum_{i \neq k} \frac{1}{M'_{ki}} - (n^+ - 2)}{\sum_{i \neq k} \frac{1}{M'_{ki}} + \sum_j \frac{1}{M_{kj}} - (n^+ + n^- - 2)}$$

end if

E. Analysis of Algorithm II

1) The Selection of k: For any $k = 1, 2, \ldots, n^+$, we could obtain a q_k. If the classifiers are "ideal" ones, the value of these q_ks would be the same as the posterior probability $p(\omega^+|x)$. However, it is more unstable than the Algorithm 1 in a noisy system. The first reason is that it uses fewer classifiers than the former one, and the result would be sensitive to the disturbance of several classifiers. Secondly, the k-th positive subset is as a central subset that it makes contribution for each used classifiers. How to select the best k is an important factor for the right decision.

The best k is the indicator of the positive sample subset whose relationship with other positive and negative subsets could effectively and correctly depict the information of discrimination between positive and negative classes. According to this interpretation, we can apply some effective decomposition methods (e.g. K-Means, Spectral Clustering [11], PCA [8], etc) to divide the original samples into well-separated subsets. Then the indicator of the subset with the most samples is an appropriate candidate for k. Other methods such as cross-validation could be used to search the best k.

2) Computational Complexity: The most appealing of this algorithm is its efficiency. During the training stage, it is only needed to train $n^+ + n^-$ classifiers. And in the testing stage, the number of operations would be linable to the number of classifiers. Overall, both the space and the time complexities are $O(n^+ + n^-)$.

TABLE I
ACCURACY ON THE GAUSSIAN AND UNIFORM DATA WITH DIFFERENT INTEGRATION METHODS

Case ID	Gaussian Accuracy (%)			Uniform Accuracy (%)		
	Min-Max	SE	SE^r	Min-Max	SE	SE^r
1	73.00	**75.00**	**75.00**	60.00	**80.00**	**80.00**
2	78.00	**83.00**	**83.00**	59.00	**79.00**	**79.00**
3	72.00	**75.00**	**75.00**	57.00	**78.00**	**78.00**
4	66.00	**69.00**	**69.00**	61.00	**75.00**	**75.00**
5	67.00	**68.00**	**68.00**	61.00	**73.00**	**73.00**
6	67.00	**74.00**	**74.00**	55.00	**74.00**	**74.00**
7	74.00	**77.00**	**77.00**	57.00	**68.00**	**68.00**
8	70.00	**72.00**	**72.00**	57.00	**78.00**	**78.00**
9	64.00	**68.00**	**68.00**	59.00	**74.00**	**74.00**
10	**69.00**	65.00	65.00	61.00	**72.00**	**72.00**

SE^r denotes the revised shrink-expansion principles.

IV. EXPERIMENTS

In this section, we present two experiments to illustrate the effectiveness of our methods.

A. Synthetic Data

In order to provide some intuitions on our algorithms' behavior without influences from the specific classifier, we show the first example in an ideal situation. The two complex distribution are composed from four 2-D gaussian distribution and four 2-D uniformed distribution respectively (Figures 4). We simulate the experiment by 10 times, and 25 samples for each distribution are randomly generated.

For each sample, the outcome matrix M is calculated according to the known distribution. Min-max, shrink-expansion principles and the revised one are used in the integration stage.

The result (Table I) shows that using min-max principles to estimate posteriors is not accurate enough. And if the pairwise classifiers could express the underlying distribution perfectly, the revised one could achieve the almost high accuracy as the shrink-expansion principles get.

B. Text Categorization

Text categorization is a task to assign a thematic label for a document based on its content. The data set we used is the 20 Newsgroups corpus [12], which is a collection of 20 different newsgroups and approximately 20000 documents. The version in our experiment is the sorted-by-date one in which duplicates and some headers are removed. In addition,

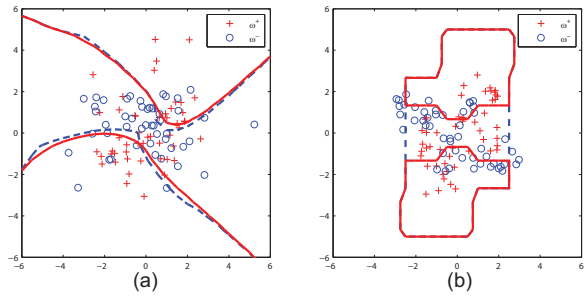

Fig. 4. Decision boundaries yielded by min-max(dashed line) and shrink-expansion(solid line) principles. a) a gaussian sample. b) a uniform sample.

TABLE II
THE F1 VALUES WITH DIFFERENT INTEGRATION METHODS

Newsgroups	$F_1(\%)$			
	Voting	Min-Max	SE	SE'
alt.atheism	71.70	74.16	**76.28**	75.00
comp.graphics	**65.17**	62.70	63.24	60.17
comp.os.ms-windows.misc	0.50	0.50	0.50	**1.51**
comp.sys.ibm.pc.hardware	60.21	61.96	**62.17**	61.01
comp.sys.mac.hardware	73.29	73.00	**75.18**	69.67
comp.windows.x	72.68	**74.09**	73.23	72.79
misc.forsale	69.70	69.21	**71.04**	68.49
rec.autos	84.91	83.61	**85.07**	82.48
rec.motorcycles	88.94	90.07	**90.71**	89.60
rec.sport.baseball	90.86	**92.35**	92.31	90.68
rec.sport.hockey	95.58	**95.82**	95.43	93.98
sci.crypt	88.42	87.77	**88.89**	86.46
sci.electronics	69.66	68.74	**70.46**	67.08
sci.med	83.49	82.79	**83.72**	82.51
sci.space	88.24	**88.89**	88.40	87.17
soc.religion.christian	85.98	86.57	**87.44**	84.61
talk.politics.guns	77.76	78.25	78.40	**78.59**
talk.politics.mideast	**88.89**	88.37	88.86	86.22
talk.politics.misc	64.43	**65.40**	64.49	61.63
talk.religion.misc	57.28	**57.86**	57.71	57.68

stemming and stop word removal are performed, and all words occurring less than 30 times in the training data are also removed. Finally, the dictionary contains 9208 different kinds of words.

The basic classifier we used is the novel naive Bayes classifier, which simply assumes that all the words are independent with others. However, the posterior estimated by naive Bayes tends to be extremely close to 0 or 1 [13]. We instead learn a sigmoidal function

$$p(\omega|x) = \frac{1}{1 + e^{Ax+B}}$$

where x is the outcome of naive Bayes classifier, and A and B's values are searched in an iterative way [14].

We respectively use four methods to handle this large-scale multi-class problem, and their performance (Table II) are compared based on the F_1 values [15].

Firstly, we simply adopt the one-against-one policy to decompose it into various binary problems, and then choose the class which receives the most votes as the document's label. The other three methods work in a same procedure as the first one in the stage of extending binary solutions to multi-class results, however, for each two-class problem, both the positive and negative training sets are further randomly partitioned into three subsets. Then these three methods do the integration in different principles.

From the results, we can see that shrink-expansion principles have better performance than min-max principles do. For some newsgroups, however, min-max principles achieve higher accuracy which mainly ascribed to two aspects. Firstly, naive Bayes could not describe the true underlying distribution accurate enough. On the other hand, as the existence of noise, min-max principles would preserve the major factor while dropping the noisy parts.

V. CONCLUSIONS

In this paper, we have built a structure of classifiers based on probabilistic formulas to solve the large-scale problems. Another contribution of our work is a statistical explanation to the min-max principle. In addition, an efficient approach is also given to reduce the complexity of computation.

ACKNOWLEDGMENT

The authors would like to give their thankfulness to HOU Xiaodi and YANG yang for their useful advices.

REFERENCES

[1] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," *Advances in Neural Information Processing Systems*, 2002.
[2] T. Joachims, "Making large-scale svm learing practical," *Advances in Kernel Methods,* MIT Press, 1998.
[3] J. Fürnkranz, "Round robin classification," *Journal of Machine Learning Research,* Vol.2, pp.721-747, 2002.
[4] J. Friedman. "Another approach to polychotomous classification," Technical report, Department of Statistics, Stanford University, 1996.
[5] D. Price, S. Knerr, L. Personnaz, and G. Dreyfus "Pairwise neural network classifiers with probabilistic outputs," *Advances in Neural Information Processing Systems*, 1995.
[6] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The Annals of Statistics,*, 26(1), pp.451-471, 1998.
[7] B. L. Lu and M. Ito, "Task decomposition and module combination based on class relations: a modular neural network for pattern classification," *IEEE Trans. Neural Networks,* Vol.10, pp.1244-1256, 1999.
[8] K. Chen, B. L. Lu, and J. T. Kwok, "Efficient classification of multi-label and imbalanced data using min-max modular classifiers," *IEEE International Joint Conference on Neural Networks*, pp.1770-1775, 2006.
[9] H. Zhao, B. L. Lu, "On efficient selection of binary classifiers for min-max modular classifier," *IEEE International Joint Conference on Neural Networks*, Vol.5, pp.3186-3191, 2005.
[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification (2nd Edition)*, Wiley-Interscience, 2000.
[11] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering analysis and an algorithm", *Advances in Neural Information Processing Systems*, 2001.
[12] http://people.csail.mit.edu/jrennie/20Newsgroups/
[13] P. N. Bennett, "Assessing the calibration of naive Bayes' posterior estimates", In Technical Report CMU-CS-00-155, Computer Science Department, School of Computer Science, Carnegie Mellon University, 2000.
[14] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood Methods", *Advances in Large Margin Classifiers*, MIT Press, 1999.
[15] D. D. Lewis, "Evaluating and optimzing autonomous text classification systems," *Proc. The 18th Annual International ACM SIGIR Conference on Research adn Development in Information Retrieval (SIGIR 95)*, pp.246-254. 1995.