On an Efficient Approximation Algorithm for Minimal Committee Machine Learning[1]

M. Yu. Khachai
Institute of Mathematics and Mechanics, Ural Division, Russian Academy of Sciences,
ul. S. Kovalevskoi 16, Yekaterinburg, 620219 Russia
e-mail: mkhachay@imm.uran.ru

Abstract—The combinatorial problem of the minimal committee of an inconsistent system of constraints arising at the stage of construction of the committee decision rule with a small number of variables is discussed. It is demonstrated that, in the general case, the problem is intractable. An approximation algorithm for a particular case of the problem (for a system of linear inequalities) is proposed; its computational complexity and estimate of accuracy are considered.

INTRODUCTION

Committee machine learning algorithms (see, e.g., the survey in [1]) construct aggregate decision rules from the elements of the basic class using the majority voting logic. Due to several objective reasons, most interesting are the algorithms that, for each particular task determined by a learning sample and a class of basic rules, construct decision rules with a minimal or close to minimal number of elements (so-called *minimal committees*). It is known that the problem of committee machine learning and the problem of finding a generalized committee solution (or just a committee) for an appropriate system of constraints (expressed, as a rule, as a system of algebraic inequalities and equations) are closely connected. In this report, a new approximation algorithm for solving the problem of the minimal committee for an inconsistent system of linear inequalities is discussed.

PROBLEM FORMULATION

Let X be an arbitrary nonempty set and let the set of its subsets $D_1, D_2, \ldots, D_m$ be given. Consider the system of inclusions

$$x \in D_j \quad (j \in \mathbb{N}_m = \{1, 2, \ldots, m\}), \tag{1}$$

not necessarily consistent, where $\bigcap_{j \in \mathbb{N}_m} D_j$ may be equal to an empty set. As usual, (see, e.g., [1]), a committee of the majority of system (1) from q elements (or just a committee) we call the finite sequence $Q = (x^1, x^2, \ldots, x^q)$ of elements of X, where

$$\left|\{i \in \mathbb{N}_q : x^i \in D_j\}\right| > q/2, \quad (j \in \mathbb{N}_m).$$

A minimal committee problem is the following combinatorial problem. An arbitrary nonempty set X and a set of its subsets $D_1, D_2, \ldots, D_m$ are given. A committee of system (1) with a minimal possible number of elements should be found.

Theorem 1. Let X, D_1, D_2,…, D_m be the finite sets. Then, the minimal committee problem is NP-hard.

Until now, the question about the polynomial or exponential computational complexity of the minimal committee problem for a system of linear inequalities in $\mathbb{R}^n$

$$(a_j, x) > 0 \quad (j \in \mathbb{N}_m) \tag{2}$$

remained open. This problem is, obviously, a particular case of the above problem (here, $X = \mathbb{R}^n$ and $D_j = \{x \in \mathbb{R}^n : (a_j, x) > 0\}$).

RESULTS

Let us impose some additional restrictions on system (2).

(i) $m > n$ and any subsystem of n inequalities is consistent;
(ii) $\|a_j\|_2 = 1$ for any $j \in \mathbb{N}_m$;
(iii) $m = 2k + n - 1$ for some natural k.

The last condition is introduced only for convenience of calculating the estimations (in the case of $m = 2k + n$, one can obtain the estimations by analogy).

Below, we present an approximation algorithm (in the sense of [3]) for solving the problem of the minimal committee of system (2) and an additional restriction on the system which enables the algorithm to find an accurate solution to the problem.

Let the sets

$$J_>(x) = \{j \in \mathbb{N}_m : (a_j, x) > 0\},$$
$$J_<(x) = \{j \in \mathbb{N}_m : (a_j, x) < 0\},$$
$$J_=(x) = \{j \in \mathbb{N}_m : (a_j, x) = 0\}$$

correspond to an arbitrary vector $x \in \mathbb{R}^n$.

[1] The work was supported by the Russian Foundation for Basic Research, project nos. 01-01-96454, 00-15-96041, 01-01-00563, and 02-01-06391.

Received October 29, 2002

Algorithm.

Step 1. Find any nontrivial solution z_1 of the system

$$(a_j, z) = 0 \; (j \in \mathbb{N}_{n-1})$$

and consider the sets $J_>(z^1)$, $J_<(z^1)$, and $J_=(z^1)$. As x^1 select any solution of the subsystem with index set J_1 of system (2) where

$$J_1 = \begin{cases} J_>(z^1) \cup J_=(z^1), & \text{if } \left|J_>(z^1)\right| > \left|J_<(z^1)\right| \\ J_<(z^1) \cup J_=(z^1), & \text{otherwise.} \end{cases}$$

Set $J = \mathbb{N}_m \backslash J_1$ and $i = 1$.

Step 2. If $J = \varnothing$, then the procedure ends and the sequence $(x^1, x^2, \ldots, x^i)$ is a committee of system (2).

Step 3. Take any subset $L \subseteq J$: $|L| = \min\{|J|, n - 1\}$ and then find the nontrivial solution z^{i+1} of the system

$$(a_j, z) = 0 \quad (j \in L)$$

and the solutions x^{i+1}, x^{i+2} of subsystems with index sets $J_>(z^{i+1}) \cup J_=(z^{i+1})$ and $J_<(z^{i+1}) \cup J_=(z^{i+1})$ of system (2).

Step 4. Set $J = J \backslash L$, $i = i + 2$ and go to Step 2.

Note that this algorithm is a simplified version of the algorithm described in [2] and is also based on the ideas of committee construction introduced by Vl. Mazurov [4]. However, unlike the mentioned algorithm, the proposed method is a polynomial, its complexity does not depend on the number of maximally consistent subsystems of system (2). Most computational time is spent on finding the solution for the two consistent subsystems of linear inequalities at each iteration. As is known, this problem is polynomially solvable.

Theorem 2.

(i) The number of iterations of the algorithm is no greater than $\left\lceil \dfrac{k}{n-1} \right\rceil$.

(ii) Let the cardinality of any maximal consistent subsystem of system (2) have the number $k + (n - 1) + t$ as an upper bound; then, the approximation ratio α of the algorithm satisfies the following condition:

$$1 \le \alpha \le \frac{2\left\lceil \dfrac{k}{n-1} \right\rceil + 1}{2\left\lceil \dfrac{k-t}{2t + (n-1)} \right\rceil + 1} \approx 1 + \frac{2t}{n-1}. \qquad (3)$$

To obtain the estimation, we use the traditional approach (see, for example, [3]). It consists in analyzing the solvable pair of dual problems of linear programming. One of them is a real relaxation of integer programming equivalent to the minimal committee problem. The numerator in the fraction estimates the number of elements of the resulting committee from above, and the denominator coincides with the value of the objective function of the dual problem calculated for some of the allowable decisions of the problem, i.e., is a lower estimation of the optimum solution to both problems.

Let us notice that, for any n, there is an infinite class of (2)-type systems for which the algorithm finds an exact solution (minimal committee). It is a class of the so-called uniformly distributed (according to D. Gale) systems of inequalities.

Definition. The system of inequalities (2) for $m = 2k + (n - 1)$ is called uniformly distributed (according to Gale) if the condition $|J_>(x)| \ge k$ is fulfilled for any $x \in \mathbb{R}^n$.

The following criterion of uniform distribution of a system is valid.

Theorem 3. System (2) is uniformly distributed if and only if both of the following conditions are fulfilled:

(i) any n vectors from $a_1, a_2, \ldots, a_m$ are linearly independent and

(ii) if L is the index set of any maximal consistent by inclusion of the subsystem of system (2), then $|L| = k + n - 1$.

As the corollary of Theorems 2 and 3 we have the following theorem.

Theorem 4. The minimal committee of uniformly (according to Gale) distributed system (2) has $2\left\lceil \dfrac{k}{n-1} \right\rceil + 1$ elements.

Therefore, we conclude that the minimal committee problem in the class of uniformly (according to Gale) distributed systems has polynomial complexity.

One can use estimate (3) to describe the class of systems for which the algorithm finds an "almost minimal" committee with the number of elements differing from minimal by no more than $2p$ (for a given natural p).

It can be expressed by the following inequality:

$$\left\lceil \frac{k-t}{2t + (n-1)} \right\rceil > \left\lceil \frac{k}{n-1} \right\rceil - p.$$

Theorem 5. Let condition (ii) of Theorem 2 be fulfilled and

$$0 < t < \frac{(p-1)(n-1)^2}{m - 2(p-1)(n-1)}, \quad p \in \mathbb{N}.$$

Then, the number of elements for the generated committee differs from the minimal by more than $2p$.

REFERENCES

1. Mazurov, Vl.D. and Khachai, M.Yu., Committee Constructions, *Izv. Ural'skogo Gos. Univ.*, Mathematics and Mechanics, 1999, issue 2(14), pp. 77–108.
2. Khachai, M.Yu. and Rybin, A. I., A New Estimate of the Number of Members in a Minimum Committee of a System of Linear Inequalities, *Pattern Recognit. Image Anal.*, 1998, vol. 8, no. 4. pp. 491–496.
3. Williamson, D.P., The Primal-Dual Method for Approximation Algorithms, *Math. Programming*, 2002, vol. 91, no. 3, Ser. B, pp. 447–478.
4. Mazurov, Vl.D., On Construction of a Committee of a System of Convex Inequalities, *Kibernetika*, 1967, no. 2, pp. 56–59.