Теорія Ймовір. та Матем. Статист.
Вип. 70, 2004

Theor. Probability and Math. Statist.
No. 70, 2005, Pages 41–51
S 0094-9000(05)00629-0
Article electronically published on August 5, 2005

GIBBS CLASSIFIERS

UDC 519.21

B. A. ZALESKY

Abstract. New statistical classifiers for dependent observations with Gibbs prior distributions of the exponential or Gaussian type are presented. It is assumed that the observations are characterized by feature functions that assume values in finite sets of rational numbers. The distributions of observations are either Gibbs exponential or Gibbs Gaussian. Arbitrary neighborhoods on a completely connected graph are considered instead of local neighborhoods of the nearest observation.

The models studied in this paper can be used for some problems of the classification of random fields, in statistical physics, and for image processing.

A method of finding an *optimal Bayes decision rule* is described. The method is based on the reduction of the problem to the evaluation of the minimal cut of an appropriate graph. The method can be used for the fast evaluation of optimal Bayes decision rules for large samples.

1. Introduction

The Gibbs estimation is a branch of the Bayes statistics that is of interest for the theory and is intensively developing nowadays in view of its extending practical applications. The Gibbs *maximal a posteriori estimators* have several useful properties (say, they have a distribution of the exponential type) that are helpful when solving various applied problems. In particular, maximal posterior estimators are widely used for the image processing and in statistical physics. However, the evaluation of such estimators (even the evaluation of approximations of such estimators) is a time consuming procedure, and until recently the evaluation of maximal posterior estimator was impossible. Different approximations have been used instead, and one of the main methods to evaluate approximate maximal posterior estimator is the Simulated Annealing. This simple and practically convenient method is computationally expensive, namely its computational time exponentially depends on the size of a sample. Moreover (and what is even more important) the Simulated Annealing not always leads to satisfactory estimators, especially in problems where the quality of estimators plays the most important role (say, in problems of image processing).

A significant progress in the Gibbs estimation is achieved over the last decade. In particular, the methods of the discrete optimization are developed for the approximation of Gibbs estimators for high dimensions (see [5]–[9]). There are also developed methods for the exact evaluation of maximal posterior estimators for models with Boolean spaces of features.

In this paper, we describe models of the cluster analysis for dependent observations with Gibbs prior distributions and methods for the evaluation of optimal Bayes decision rules that minimize the probability of a wrong classification. There are many papers

2000 *Mathematics Subject Classification.* Primary 60G60, 62H30; Secondary 90C27.
Supported by a grant of MNTC B–517.

devoted to various problems of the cluster analysis (see, for example, [10]–[13]). Some of them [12, 13] deal with methods of the classification of dependent observations. We consider the problem of classification of observations $y_1, y_2, \ldots, y_n$ represented by a feature vector whose coordinates assume values in an arbitrary finite set of rational numbers. The clusters are identified with the help of labels that also assume only a finite number of rational values. The labels are dependent and have a Gibbs distribution. The models mentioned above appear when classifying nonhomogeneous random fields, in statistical physics, in the image processing, etc. The Gibbs field of labels is defined on an directed completely connected graph whose vertices correspond to observations. Until recently only graphs with nearest connected vertices were considered; moreover, they were regular lattices).

In what follows we avoid the language of Markov fields, since there is a nonlocal dependence of observations (the latter property is due to the complete connectedness of the graph).

Below we present methods allowing one to evaluate exactly the optimal Bayes decision rule for the problem of the Gibbs classification. These methods reduce the problem of the Bayes classification to the problem of integer optimization and the problem of the minimal cut of an appropriate graph. Minimal cuts of graphs can be computed with the help of fast algorithms [9, 14] that allow execution in concurrent mode.

The results obtained below solve, in particular, a problem posed by Greig, Porteous, and Seheult [4] in 1989.

2. Description of the models

The energy functions of the Gibbs distributions studied in this paper are homogeneous in all their arguments. This allows one to reduce the problem to the case of integer-valued feature functions.

Let the feature function f of observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ assume values in the set $Z_L = \{0, 1, \ldots, L\}$ and put $f_i = f(y_i)$.

Remark 1. Feature functions related to observations allow one to classify not only numerical data but also nonnumerical data, say functions or sets.

Let

$$\mathcal{M} = \{m(0), m(1), \ldots, m(k)\}, \qquad 0 \le m(0) < m(1) < \cdots < m(k) \le L,$$

be the set of possible values of cluster labels. We assume that the observations

$$y_1, y_2, \ldots, y_n$$

may belong to one of the $k+1$ clusters (we use the number $k+1$ instead of k to simplify the notation). Each cluster corresponds to a number j, $0 \le j \le k$, and a fixed integer-valued label $m_j \in \mathcal{M}$.

The observations $y_1, y_2, \ldots, y_n$ are treated as random variables, while labels are assumed to be dependent random variables forming a Gibbs field. The Gibbs random field of labels $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is defined on a directed and completely connected graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices of the graph and

$$E = \{(i, j) \mid i, j \in V\}$$

is the set of its directed arcs. Vertices of the graph correspond to observations. The values of $\mathbf{x}$ belong to the set $\mathcal{M}^V$, that is, $x_i \in \mathcal{M}$, $i \in V$. The distribution of the Gibbs field $\mathbf{x}$ is either exponential or Gaussian, that is, its density is either

$$p_{\text{exp}}(\mathbf{m}) = c \cdot \exp\Big\{-\sum_{(i,j)\in E} \beta_{i,j}|m_i - m_j|\Big\}, \quad \text{or}$$

$$p_{\text{gaus}}(\mathbf{m}) = c \cdot \exp\Big\{-\sum_{(i,j)\in E} \beta_{i,j}(m_i - m_j)^2\Big\}$$

where c is a normalizing constant (in what follows we denote different constant by the same symbol), $\mathbf{m} = \{m_1, m_2, \dots, m_n\}$, and $\beta_{i,j} \geq 0$.

Remark 2. The distribution $p_{\text{exp}}(\mathbf{m})$ is more convenient than $p_{\text{gaus}}(\mathbf{m})$ in some applied problems because of a small blurring effect in the case of p_{exp}. Moreover the evaluation of the optimal Bayes decision rule is more computationally efficient in this case.

Given $\mathbf{x}$, the feature functions $f_1, f_2, \dots, f_n$ are conditionally independent and the conditional distribution is

$$p_i^m(l) = \mathsf{P}(f_i = l \mid \mathbf{x} = \mathbf{m}), \qquad i \in V,\ l \in Z_L,\ \mathbf{m} \in \mathcal{M}^V.$$

The problem of the Bayes classification is to find the optimal Bayes decision rule

$$\widehat{\mathbf{m}} = \widehat{\mathbf{m}}(\mathbf{f}) = (\widehat{m}_1, \widehat{m}_2, \dots, \widehat{m}_n)$$

that estimates the unknown vector of labels $\mathbf{x}$ and minimizes the Bayes risk

$$R(\widehat{\mathbf{m}}) = \mathsf{P}(\widehat{\mathbf{m}} \neq \mathbf{x}).$$

The optimal Bayes decision rule for an arbitrary prior distribution coincides with the maximum a posteriori probability estimator [13].

Theorem. *Let*

$$\mathbb{Q}(\mathbf{m}), \qquad \mathbf{m} \in \mathcal{M}^V,$$

be an arbitrary prior distribution of clusters. Then the optimal Bayes decision rule minimizing the risk $R(\widehat{\mathbf{m}})$ is given by

$$\widehat{\mathbf{m}} = \operatorname*{argmax}_{\mathbf{m}\in\mathcal{M}^V} \left\{\sum_{i\in V} \ln p_i^m(f_i) + \ln \mathbb{Q}(\mathbf{m})\right\}. \tag{1}$$

In this paper, $\mathbb{Q}(\mathbf{m})$ equals either $p_{\text{exp}}(\mathbf{m})$ or $p_{\text{gaus}}(\mathbf{m})$. In each of these cases we obtain an efficient method for the evaluation of optimal Bayes decision rule for nonidentically distributed f_i, $i \in V$, if the distribution is either exponential

$$p_{i,\text{exp}}^m(f_i) = c_i \cdot \exp\{-\lambda_i|f_i - m_i|\}, \qquad \lambda_i \geq 0,$$

or Gaussian

$$p_{i,\text{gaus}}^m(f_i) = c_i \cdot \exp\{-\lambda_i(f_i - m_i)^2\}, \qquad \lambda_i \geq 0.$$

Both these distributions are widely used in applied statistics. The optimal Bayes decision rules for these types of distributions are given by the energy functions

$$\widehat{\mathbf{m}}_{\text{exp}} = \operatorname*{argmin}_{m\in\mathcal{M}^V} \Big\{\lambda_i \sum_{i\in V} |f_i - m_i| + \sum_{(i,j)\in E} \beta_{i,j}|m_i - m_j|\Big\} \tag{2}$$

and

$$\widehat{\mathbf{m}}_{\text{gaus}} = \operatorname*{argmin}_{m\in\mathcal{M}^V} \Big\{\sum_{i\in V} \lambda_i(f_i - m_i)^2 + \sum_{(i,j)\in E} \beta_{i,j}(m_i - m_j)^2\Big\}, \tag{3}$$

respectively. The optimal Bayes decision rule can be treated as a measure of discrepancy between the feature function $\mathbf{f}$ and the vector of labels $\mathbf{m}$.

Despite its easy formulation the problem of the evaluation of $\widehat{\mathbf{m}}_{\text{exp}}$ and $\widehat{\mathbf{m}}_{\text{gaus}}$ is complicated for large samples. For example, computer images which often are objects for the classification contain up to 2^{18} variables. Nevertheless computational algorithms for both these classifiers require a polynomial time of execution; moreover the run-time is ckn^3 for the first case and $c(kn)^3$ for the second. For many applied problems the run-time is of linear order $O(nk)$ even in the nonconcurrent mode. It is also possible to compute optimal Bayes decision rules for mixed models with exponential conditional distributions $p^m_{i,\text{exp}}(f_i)$ and Gaussian prior distributions $p_{\text{gaus}}(\mathbf{m})$ and vice versa, that is, with Gaussian conditional distributions and exponential prior distributions.

3. Evaluation of optimal classifiers

If there are only two clusters, that is, $k = 2$, then the optimal Bayes decision rules $\widehat{\mathbf{m}}_{\text{exp}}$ and $\widehat{\mathbf{m}}_{\text{gaus}}$ coincide, since $b^2 = |b|$ for any Boolean variable b. The optimal Bayes decision rules can be evaluated in this case with the help of methods developed for finding the minimal cut of a graph [4, 15]. In 1989 Greig, Porteous, and Seheult [4] proposed a heuristic network flow algorithm that is especially efficient for the evaluation of approximate optimal Bayes decision rules. A multiresolution network flow algorithm for constructing the minimal cut in a graph was recently developed in [9]. That algorithm allows one to evaluate exactly the Boolean classifiers in the case of $k = 2$. The problem of the evaluation of the classifier in the case of $k > 2$ is posed in [4]. The results below solve this problem.

Evaluation of $\widehat{\mathbf{m}}_{\text{exp}}$. Let

$$U_1(\mathbf{m}) = \sum_{i\in V} \lambda_i |f_i - m_i| + \sum_{(i,j)\in E} \beta_{i,j} |m_i - m_j|, \qquad \mathbf{m} \in \mathcal{M}^V.$$

The idea of the method is to represent the vector of labels $\mathbf{m}$ with the help of an integer-valued combination of Boolean vectors and then reduce the problem of the integer optimization of $U_1(\mathbf{m})$ to a problem of Boolean optimization. The latter problem can be solved successfully by using methods of the combinatorial optimization.

Let μ and ν be integers and the indicator function $\mathbb{1}_{(\mu\geq\nu)}$ equal one if $\mu \geq \nu$ and equal zero otherwise. For all $\mu \in \mathcal{M}$ and all Boolean variables $x(l) = \mathbb{1}_{(\mu\geq m(l))}$ such that

$$x(1) \geq x(2) \geq \cdots \geq x(k)$$

we have

$$\mu = m(0) + \sum_{l=1}^{k} (m(l) - m(l-1)) x(l). \tag{4}$$

Conversely, any nonincreasing sequence of Boolean variables

$$x(1) \geq x(2) \geq \cdots \geq x(k)$$

determines a label $\mu \in \mathcal{M}$ according to the rule (4). Thus relation (4) is a one-to-one correspondence between the sequences of nonincreasing Boolean variables and the labels.

Similarly, the feature functions can be represented as sums

$$f_i = \sum_{\tau=1}^{L} f_i(\tau)$$

of nonincreasing sequences of Boolean variables

$$f_i(1) \geq f_i(2) \geq \cdots \geq f_i(L).$$

Let $\mathbf{x}(l) = (x_1(l), x_2(l), \ldots, x_n(l))$ be a Boolean vector for $l = 1, \ldots, k$, the coordinates of the vector $\mathbf{z}(l) = (z_1(l), \ldots, z_n(l)$ be such that

$$z_i(l) = \frac{1}{m(l) - m(l-1)} \sum_{\tau=m(l-1)+1}^{m(l)} f_i(\tau), \qquad i = 1, \ldots, n,$$

and

$$|\mathbf{x}| = \sum_{i=1}^{n} |x_i|$$

be the norm of the vector $\mathbf{x}$.

In what follows we use the following result.

Proposition 1. *For all integers $\nu \in \mathcal{M}$ and all $f_i \in Z_L$*

$$|\nu - f_i| = \left| m(0) - \sum_{\tau=1}^{m(0)} f_i(\tau) \right| + \sum_{l=1}^{k} \left| (m(l) - m(l-1))\, \mathbb{1}_{(\nu \geq m(l))} - \sum_{\tau=m(l-1)+1}^{m(l)} f_i(\tau) \right| + \sum_{\tau=m(k)+1}^{L} f_i(\tau).$$

Thus the function $U_1(\mathbf{m})$ can be represented in the following form:

$$U_1(\mathbf{m}) = \sum_{l=1}^{k} (m(l) - m(l-1)) u(l, \mathbf{x}(l)) \tag{5}$$

for all feature functions $\mathbf{f} \in Z_L^V$ and for all Boolean vectors

$$\mathbf{x}(l) = \left(\mathbb{1}_{(m_1 \geq l)}, \mathbb{1}_{(m_2 \geq l)}, \ldots, \mathbb{1}_{(m_n \geq l)} \right)$$

where

$$u(l, \mathbf{b}) = \sum_{i \in V} \lambda_i |z_i(l) - b_i| + \sum_{(i,j) \in E} \beta_{i,j}\, |b_i - b_j|$$

for a given n-dimensional Boolean vector $\mathbf{b}$.

Denote by

$$\check{\mathbf{x}}(l) = \operatorname*{argmin}_{\mathbf{b}} u(l, \mathbf{b}), \qquad l = 1, \ldots, k, \tag{6}$$

the Boolean solutions that minimize the functions $u(l, \mathbf{b})$. For vectors $\mathbf{v}$ and $\mathbf{w}$ we write $\mathbf{v} \geq \mathbf{w}$ if $v_i \geq w_i$ for all $i = 1, \ldots, n$. Analogously, we write $\mathbf{v} \not\geq \mathbf{w}$ if there are at least two pairs of coordinates (v_i, v_j) and (w_i, w_j) such that $v_i \geq w_i$ and $v_j < w_j$.

Note that

$$\mathbf{z}(1) \geq \mathbf{z}(2) \geq \cdots \geq \mathbf{z}(k).$$

Moreover

$$z_i(1) = z_i(2) = \cdots = z_i(\nu - 1) = 1, \qquad 0 \leq z_i(\nu) \leq 1,$$
$$z_i(\nu + 1) = \cdots = z_i(k) = 0$$

for some integer $1 \leq \nu \leq k$. In general, it is easy to see that solutions $\check{\mathbf{x}}(l)$ are unordered. Nevertheless, there always exists at least one nonincreasing sequence $\check{x}(l)$ of solutions of problem (6).

Theorem 2. *There exists a nonincreasing sequence*

$$\check{\mathbf{x}}(1) \geq \check{\mathbf{x}}(2) \geq \cdots \geq \check{\mathbf{x}}(k)$$

of solutions of problem (6).

Proof. Let $D \in V$ be a set of vertices and $\mathbf{b}_D = (b_i)_{i \in D}$ be the restriction of the vector $\mathbf{b}$. Put

$$u_D(l, \mathbf{b}) = \sum_{i \in D} \lambda_i |z_i(l) - b_i| + \sum_{(i,j) \in (D \times D) \cup (D \times D^c) \cup (D^c \times D)} \beta_{i,j} |b_i - b_j|.$$

Assume that Boolean vectors $\mathbf{z}(l') \geq \mathbf{z}(l'')$ are ordered for $1 \leq l' \leq l'' \leq k$, and let $\check{\mathbf{x}}(l') \ngeq \check{\mathbf{x}}(l'')$ be two unordered solutions of problem (6). Let

$$D = \{ i \mid \check{x}_i(l') = 0,\ \check{x}_i(l'') = 1 \}$$

and put $D^c = V \setminus D$. Note that the restrictions $\check{\mathbf{x}}_{D^c}(l') \geq \check{\mathbf{x}}_{D^c}(l'')$ are ordered. The functions

$$v_D(l, \mathbf{b}) = u(l, \mathbf{b}) - u_D(l, \mathbf{b})$$

do not depend on $\mathbf{b}_D$. Thus the restriction $\check{\mathbf{x}}_D(l)$ minimizes the function $u_D(l, \mathbf{b})$ for a fixed $\check{\mathbf{x}}_{D^c}(l)$ if $\check{\mathbf{x}}(l)$ is a solution of problem (6). This implies the following two inequalities:

$$\begin{aligned} u_D(l', \check{\mathbf{x}}(l')) &= \sum_{i \in D} \lambda_i z_i(l') + \sum_{(i,j) \in (D \times D^c)} (\beta_{i,j} + \beta_{j,i}) \check{x}_j(l') \\ &\leq \sum_{i \in D} \lambda_i (1 - z_i(l')) + \sum_{(i,j) \in (D \times D^c)} (\beta_{i,j} + \beta_{j,i})(1 - \check{x}_j(l')) \\ &= u_D(l', \mathbf{w}(l')), \end{aligned} \tag{7}$$

where

$$\mathbf{w}(l') = (\mathbb{1}_D, \check{\mathbf{x}}_{D^c}(l')) = (\check{\mathbf{x}}_D(l''), \check{\mathbf{x}}_{D^c}(l')),$$

and

$$\begin{aligned} u_D(l'', \check{\mathbf{x}}(l'')) &= \sum_{i \in D} \lambda_i (1 - z_i(l'')) + \sum_{(i,j) \in (D \times D^c)} (\beta_{i,j} + \beta_{j,i})(1 - \check{x}_j(l'')) \\ &\leq \sum_{i \in D} \lambda_i z_i(l'') + \sum_{(i,j) \in (D \times D^c)} (\beta_{i,j} + \beta_{j,i}) \check{x}_j(l'') \\ &= u_D(l'', \mathbf{w}(l'')) \end{aligned} \tag{8}$$

for the vector

$$\mathbf{w}(l'') = (0_D, \check{\mathbf{x}}_{D^c}(l'')) = (\check{\mathbf{x}}_D(l'), \check{\mathbf{x}}_{D^c}(l'')).$$

We have

$$u_D(l', \mathbf{w}(l')) \leq u_D(l'', \mathbf{w}(l'')), \qquad u_D(l', \check{\mathbf{x}}(l')) \geq u_D(l'', \mathbf{w}(l'')),$$

since $\mathbf{z}(l') \geq \mathbf{z}(l'')$ and $\check{\mathbf{x}}_{D^c}(l') \geq \check{\mathbf{x}}_{D^c}(l'')$. Therefore all the inequalities in (7) and (8) are, in fact, equalities. Thus the ordered Boolean vectors $\mathbf{w}(l') \geq \mathbf{w}(l'')$ are solutions of problem (6) for l' and l'', respectively. □

Some properties of the structure of solutions $\check{\mathbf{x}}(l)$ are described in the following result.

Corollary 3. *If $1 \leq l' < l'' \leq k$ are integers and a sequence of vectors is such that*

$$\mathbf{z}(1) \geq \mathbf{z}(2) \geq \cdots \geq \mathbf{z}(k),$$

then

(i) *if $\check{\mathbf{x}}(l')$ is an arbitrary solution of problem* (6)*, then there exists a solution $\check{\mathbf{x}}(l'')$ such that $\check{\mathbf{x}}(l') \geq \check{\mathbf{x}}(l'')$. Conversely, if $\check{\mathbf{x}}(l'')$ is an arbitrary solution of problem* (6)*, then there exists a solution $\check{\mathbf{x}}(l')$ such that $\check{\mathbf{x}}(l') \geq \check{\mathbf{x}}(l'')$;*

(ii) *given $1 \leq l \leq k$, the set of solutions $\{\check{\mathbf{x}}(l)\}$ has a minimal $\underline{\mathbf{x}}(l)$ and maximal $\overline{\mathbf{x}}(l)$ elements;*

(iii) *the sets of minimal and maximal elements are ordered, that is, $\underline{\mathbf{x}}(1) \geq \cdots \geq \underline{\mathbf{x}}(k)$ and $\overline{\mathbf{x}}(1) \geq \cdots \geq \overline{\mathbf{x}}(k)$.*

Statement (i) follows explicitly from Theorem 2. Statement (ii) is derived from (i) for $l' = l''$, while statement (iii) is obtained from (ii) and the definition of minimal and maximal elements.

Let $\mathbf{b}'$ and $\mathbf{b}''$ be Boolean vectors. Let the coordinates of the vectors $\underline{\mathbf{b}} = \mathbf{b}' \wedge \mathbf{b}''$ and $\overline{\mathbf{b}} = \mathbf{b}' \vee \mathbf{b}''$ are $\underline{b}_i = \min\{b'_i, b''_i\}$ and $\overline{b}_i = \max\{b'_i, b''_i\}$, respectively. If $\check{\mathbf{x}}(1), \dots, \check{\mathbf{x}}(k)$ is an arbitrary unordered sequence of solutions of problem (6), then an ordered sequence of solutions

$$\mathbf{x}_{(1)} \geq \mathbf{x}_{(2)} \geq \cdots \geq \mathbf{x}_{(k)}$$

can be obtained by applying the operators $\wedge$ and $\vee$.

Further, it is easy to check that the sum of ordered solutions $\check{\mathbf{x}}(l)$ is a solution of (1).

Proposition 4. *If $\check{\mathbf{x}}(1) \geq \check{\mathbf{x}}(2) \geq \cdots \geq \check{\mathbf{x}}(k)$ is an arbitrary sequence of ordered solutions of problem* (6)*, then the sum*

$$\widehat{\mathbf{m}} = m(0) + \sum_{l=1}^{k} (m(l) - m(l-1))\check{\mathbf{x}}(l)$$

is an optimal Bayes decision rule that minimizes $U_1(\mathbf{m})$.

The problem of finding the Boolean solutions $\check{\mathbf{x}}(l)$ is known in discrete optimization [4]. It is equivalent to the problem of finding a minimal cut of an appropriate graph. Note that there are fast algorithms of finding minimal cuts [9, 14]. In Section 3 we exhibit an application of the method considered in this paper.

Evaluation of $\widehat{\mathbf{m}}_{\text{gaus}}$. We describe an algorithm for the minimization of the function

$$U_2(\mathbf{m}) = \sum_{i \in V} \lambda_i (f_i - m_i)^2 + \sum_{(i,j) \in E} \beta_{i,j} (m_i - m_j)^2, \qquad \mathbf{m} \in \mathcal{M}^V.$$

To determine the solution $\widehat{\mathbf{m}}_{\text{gaus}}$ that minimizes the function $U_2(\mathbf{m})$ we again use the representation of the vector of labels $\mathbf{m}$ as an integer combination of Boolean vectors. Then the problem of the integer minimization of the function $U_2(\mathbf{m})$ is reduced to a problem of the Boolean optimization. Now we use a single (large) graph with $kn+2$ nodes instead of k graphs with $n+2$ nodes each, used in the above discussion.

Put

$$g_i = f_i - m(0), \quad i \in V, \qquad a_l = m(l) - m(l-1), \quad l = 1, \dots, k.$$

According to (4) the vector $\mathbf{m}$ can be represented as the linear combination

$$\mathbf{m}(0) + \sum_{l=1}^{k} a_l \mathbf{x}(l)$$

of Boolean vectors $\mathbf{x}(l) = (x_1(l), x_2(l), \dots, x_n(l))$, while the function U_2 can be written as follows:

$$U_2(\mathbf{m}) = \sum_{i \in V} \lambda_i \left(g_i - \sum_{l=1}^{k} a_l x_i(l) \right)^2 + \sum_{(i,j) \in E} \beta_{i,j} \left(\sum_{l=1}^{k} a_l \big(x_i(l) - x_j(l) \big) \right)^2.$$

Let $d_{l,\tau} = a_l a_\tau$, $l, \tau = 1, \dots, k$, and $\beta_{i,i} = 0$, $i \in V$. Then

$$U_2(\mathbf{m}) = \sum_{i \in V} \lambda_i g_i^2 + P(\mathbf{x}(1), \dots, \mathbf{x}(k)) \tag{9}$$

where the polynomial of the Boolean variables

$$
\begin{aligned}
&P(\mathbf{x}(1),\dots,\mathbf{x}(k))\\
&\quad=\sum_{i\in V}\lambda_i\Big[\sum_{l=1}^{k}\big(a_l^2-2g_ia_l\big)x_i(l)\Big]\\
&\qquad-\sum_{(i,j)\in E}\beta_{i,j}\sum_{1\le\tau<l\le k}d_{l,\tau}\big(x_i(l)+x_j(\tau)+x_j(l)+x_i(\tau)\big)\\
&\qquad+\sum_{i\in V}\lambda_i\sum_{l\ne\tau}d_{l,\tau}x_i(l)x_i(\tau)+\sum_{(i,j)\in E}\beta_{i,j}\sum_{l\ne\tau}d_{l,\tau}(x_i(l)x_i(\tau)+x_j(l)x_j(\tau))\\
&\qquad+\sum_{(i,j)\in E}\beta_{i,j}\Big[\sum_{l=1}^{k}a_l^2\big(x_i(l)-x_j(l)\big)^2\\
&\qquad\qquad\qquad+\sum_{1\le\tau<l\le k}d_{l,\tau}\big[\big(x_i(l)-x_j(\tau)\big)^2+(x_j(l)-x_i(\tau))^2\big]\Big]
\end{aligned}
$$

can be rewritten as follows:

$$
\begin{aligned}
&P(\mathbf{x}(1),\dots,\mathbf{x}(k))\\
&\quad=\sum_{i\in V}\sum_{l=1}^{k}\Big[\lambda_ia_l^2-2\lambda_ig_ia_l-a_l(m(k)-m(0)-a_l)\sum_{j\in V}(\beta_{i,j}+\beta_{j,i})\Big]x_i(l)\\
&\qquad+2\sum_{i\in V}\Big[\lambda_i+\sum_{j\in V}(\beta_{i,j}+\beta_{j,i})\Big]\sum_{1\le\tau<l\le k}d_{l,\tau}x_i(\tau)x_i(l)\\
&\qquad+\sum_{(i,j)\in E}\beta_{i,j}\Big[\sum_{l=1}^{k}a_l^2\big(x_i(l)-x_j(l)\big)^2\\
&\qquad\qquad\qquad+\sum_{l\ne\tau}d_{l,\tau}\big[\big(x_i(l)-x_j(\tau)\big)^2+(x_j(l)-x_i(\tau))^2\big]\Big].
\end{aligned}
$$

Note that the latter polynomial has the same points of minimum as $U_2(\mathbf{m})$. Consider another polynomial of the Boolean variables:

$$
\begin{aligned}
&Q(\mathbf{x}(1),\dots,\mathbf{x}(k))\\
&\quad=\sum_{i\in V}\sum_{l=1}^{k}\Big[\lambda_ia_l^2-2\lambda_ig_ia_l-a_l(m(k)-m(0)-a_l)\sum_{j\in V}(\beta_{i,j}+\beta_{j,i})\Big]x_i(l)\\
&\qquad+2\sum_{i\in V}\Big[\lambda_i+\sum_{j\in V}(\beta_{i,j}+\beta_{j,i})\Big]\sum_{1\le\tau<l\le k}d_{l,\tau}x_i(l)\\
&\qquad+\sum_{(i,j)\in E}\beta_{i,j}\Big[\sum_{l=1}^{k}a_l^2\big(x_i(l)-x_j(l)\big)^2\\
&\qquad\qquad\qquad+\sum_{l\ne\tau}d_{l,\tau}\big[\big(x_i(l)-x_j(\tau)\big)^2+(x_j(l)-x_i(\tau))^2\big]\Big].
\end{aligned}
$$

Note that $Q(\mathbf{x}(1),\dots,\mathbf{x}(k))\ge P(\mathbf{x}(1),\dots,\mathbf{x}(k))$.

By

$$(\mathbf{q}^*(1),\mathbf{q}^*(2),\dots,\mathbf{q}^*(k))=\operatorname*{argmin}_{\mathbf{x}(1),\mathbf{x}(2),\dots,\mathbf{x}(k)}Q(\mathbf{x}(1),\dots,\mathbf{x}(k))$$

we denote any family of Boolean vectors that minimizes $Q(\mathbf{x}(1),\dots,\mathbf{x}(k))$. Now we check that $\mathbf{q}^*(1)\ge\mathbf{q}^*(2)\ge\dots\ge\mathbf{q}^*(L-1)$. This property allows one to represent the solution of the initial problem in the form $\widehat{\mathbf{m}}_{\mathrm{gaus}}=\sum_{l=1}^{L-1}\mathbf{q}^*(l)$.

Theorem 5. *Any family* $(\mathbf{q}^*(1),\mathbf{q}^*(2),\dots,\mathbf{q}^*(L-1))$ *that minimizes* Q *is a nonincreasing sequence.*

The polynomials P *and* Q *have the same ordered solutions, and therefore every solution* $\widehat{\mathbf{m}}_{\mathrm{gaus}}$ *is given by* $\widehat{\mathbf{m}}_{\mathrm{gaus}}=\sum_{l=1}^{k}\mathbf{q}^*(l)$.

Proof. We represent Q as follows:

$$Q(\mathbf{x}(1),\dots,\mathbf{x}(k)) = P(\mathbf{x}(1),\dots,\mathbf{x}(k))+2\sum_{i\in V}\left[\lambda_i+\sum_{j\in V}(\beta_{i,j}+\beta_{j,i})\right]\sum_{1\le\tau<l\le k}d_{l,\tau}(1-x_i(\tau))x_i(l).$$

Assume that there is an unordered family

$$(\mathbf{q}^*(1),\mathbf{q}^*(2),\dots,\mathbf{q}^*(L-1))$$

that minimizes Q. Then

$$\sum_{1\le\tau<l\le k}d_{l,\tau}(1-q_i^*(\tau))q_i^*(l)>0$$

for at least one index $i\in V$ (recall that $d_{l,\tau}>0$ for all τ). Thus

$$Q(\mathbf{q}^*(1),\dots,\mathbf{q}^*(k))>P(\mathbf{q}^*(1),\dots,\mathbf{q}^*(k)).$$

It follows from equality (9) that the polynomial P depends on the sum $\bar{\mathbf{m}}=\sum_{l=1}^{k}\mathbf{q}^*(l)$ only. Consider another family $(\bar{\mathbf{x}}(1),\dots,\bar{\mathbf{x}}(k))$ whose coordinates are $\bar{x}_i(l)=\mathbb{1}_{(\overline{m_i}\ge l)}$. This is a nondecreasing family such that $\overline{\mathbf{m}}=\sum_{l=1}^{k}\overline{\mathbf{x}}(l)$, whence we obtain that

$$P(\bar{\mathbf{x}}(1),\dots,\bar{\mathbf{x}}(k))=P(\mathbf{q}^*(1),\dots,\mathbf{q}^*(k)).$$

Since this family is monotone,

$$P(\bar{\mathbf{x}}(1),\dots,\bar{\mathbf{x}}(k))=Q(\bar{\mathbf{x}}(1),\dots,\bar{\mathbf{x}}(k)).$$

Hence the unordered family $(\mathbf{q}^*(1),\dots,\mathbf{q}^*(k))$ cannot minimize Q.

The polynomials P and Q have the same set of ordered solutions, since P and Q coincide at all ordered sequences of Boolean variables $(\mathbf{x}(1),\dots,\mathbf{x}(k))$. Therefore,

$$\widehat{\mathbf{m}}_{\mathrm{gaus}}=\sum_{l=1}^{k}\mathbf{q}^*(l). \qquad \square$$

Theorem 5 allows one to evaluate the classifier $\widehat{\mathbf{m}}_{\mathrm{gaus}}$ with the help of the Boolean optimization of the polynomial Q. Unlike the case of P, the polynomial Q can be minimized by using the algorithms for finding the minimal cuts of graphs (see [4, 9]). The appropriate graph is constructed in [8].

4. Applications

To exhibit the methods described above we give an example of the evaluation of optimal Bayes decision rule $\widehat{\mathbf{m}}_{\mathrm{exp}}$ for a grey scale image distorted by a random noise.

The original grey scale ultrasound image of a thyroid gland shown in Figure 1(a) is distorted by a strong random noise. The problem is to discriminate the boundary of the thyroid body. Figure 1(b) shows the boundary of the thyroid body drawn by an expert. The optimal Bayes decision rule $\widehat{\mathbf{m}}_{\mathrm{exp}}$ for the exponential Gibbs model is shown in Figures 1(c) and 1(d). The optimal Bayes decision rule serves in this case to

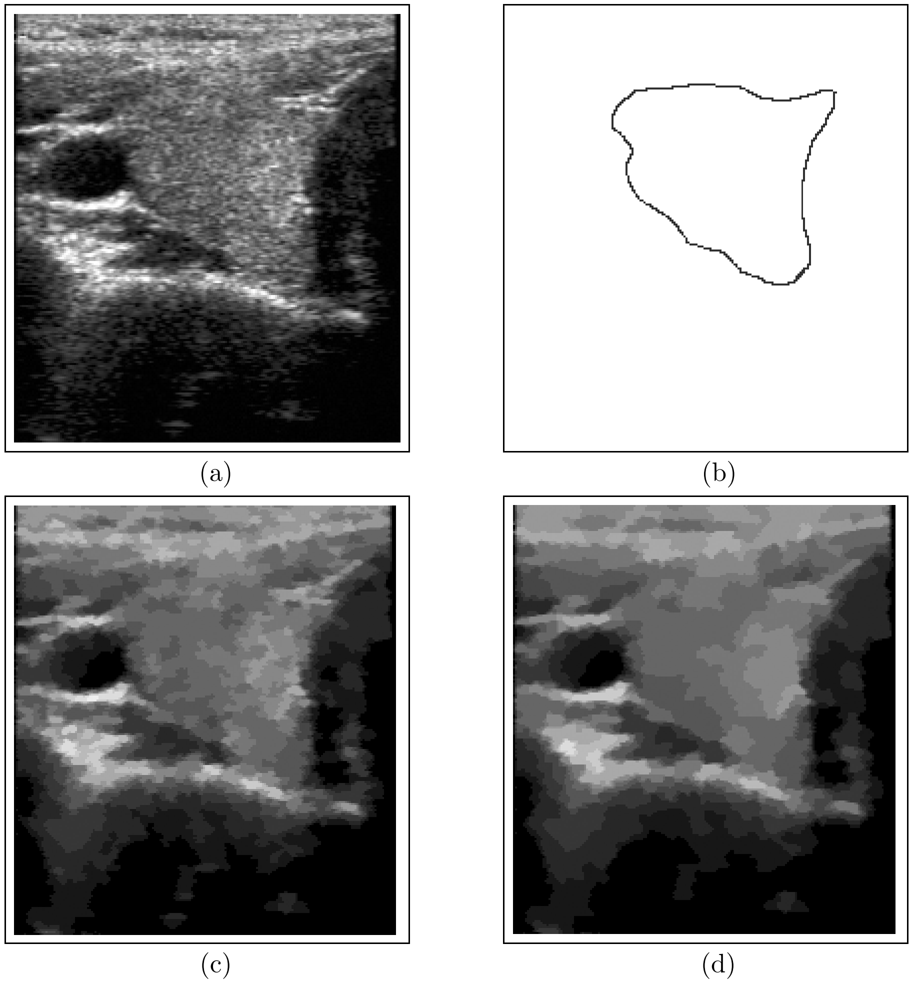

Figure 1.

discriminate the boundary of the thyroid body by a computer. It can be seen from the figures that the boundary drawn by the expert is near the boundaries of the clusters.

Bibliography

1. S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intel. PAMI-6 **6** (1984), 721–741.
2. B. Gidas, *Metropolis Type Monte Carlo Simulation Algorithm and Simulated Annealing*, Topics in Contemp. Probab. Appl., Stochastics Ser., CRC, Boca Raton, Fl, 1995. MR1410538 (98b:60128)
3. B. A. Zalesky, *Sthochastic relaxation for buildind some classes of piecewise linear regression functions*, Monte-Carlo Meth. Appl. **6** (2000), no. 2, 141–157. MR1773375
4. D. M. Greig, B. T. Porteous, and A. H. Seheult, *Exact maximum a posteriori estimation for binary images*, J. R. Statist. Soc. B. **58** (1989), 271–279.
5. Yu. Boykov, O. Veksler, and R. Zabih, *Fast approximate energy minimization via graph cuts*, IEEE Trans. Pattern Anal. Machine Intel. **23** (2001), no. 11, 1222–1239.
6. V. Kolmogorov and R. Zabih, *What Energy Functions Can Be Minimized Via Graph Cuts?*, Cornell CS Technical Report, TR2001-1857 (2001).

7. B. A. Zalesky, *Computation of Gibbs estimates of gray-scale images by discrete optimization methods*, Proceedings of the Sixth International Conference PRIP'2001 (Minsk, May 18–20, 2001), pp. 81–85.
8. B. A. Zalesskiĭ, *Efficient minimization of quadratic polynomials in integer variables with quadratic monomials* $b_{i,j}^2(x_i - x_j)^2$, Dokl. Nats. Akad. Nauk Belarusi **45** (2001), no. 6, 9–11. (Russian) MR1984687 (2003m:90073)
9. B. A. Zalesky, *Network Flow Optimization for Restoration of Images*, Preprint AMS, Mathematics ArXiv, math.OC/0106180.
10. S. A. Aivazyan, B. M. Bukhshtaber, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Classification and Dimensionality Reduction*, "Finansy i statistika", Moscow, 1989. (Russian)
11. G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992. MR1190469 (94a:62091)
12. V. V. Mottl' and I. B. Muchnik, *Hidden Markov Models in Structural Analysis of Signals*, Fizmatlit, Moscow, 1999. (Russian) MR1778152 (2001m:94014)
13. E. E. Zhuk and Yu. S. Kharin, *Stability in the Cluster Analysis of Multivariate Data*, Belgosuniversitet, Minsk, 1998. (Russian)
14. B. V. Cherkassky and A. V. Goldberg, *On Implementing Push-Relabel Method for the Maximum Flow Problem*, Technical Report STAN-CS-94-1523, Department of Computer Science, Stanford University, 1994.
15. J. C. Picard and H. D. Ratliff, *Minimal cuts and related problems*, Networks **5** (1975), 357–370. MR0391960 (52:12778)

Joint Institute of Problems in Informatics, National Academy of Sciences, Surganov Street 6, Minsk 220012, Belarus

E-mail address: `zalesky@newman.bas-net.by`

Received 4/APR/2002

Translated by THE AUTHOR