DISCRETE MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 7
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1982

ВОССТАНОВЛЕНИЕ НЕОПРЕДЕЛЕННЫХ ЗНАЧЕНИЙ ПРИЗНАКОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ОБРАЗОВ

М. МИХАЛЕВИЧ

Институт основ вычислительной техники Польской Академии Наук, Варшава, Польша

Во многих задачах распознавания, классификации или прогноза существует проблема неопределенной информации. Точность всех эвристических процедур, реализующих вышеуказанные задачи, имеет тесную связь с качеством введенных обучающих и контрольных объектов. Отсутствие какой-либо информации об этих объектах, или отсутствие информации об распознаваемых обьектах очень сильно влияет на полученные результаты. Поэтому в существующих алгоритмах либо не допускается неопределенности информации либо априори определяется способы ее трактовки. Главной целью всех таких способов есть минимализация потерей точности полученных результатов.

Следующую работу посвящается описанию традиционных и новых методов трактовки неопределенной информации об объектах в эвристических алгоритмах распознавания.

1. Описание класса распознающих алгоритмов

Предпологаем, что каждый объект описывается значениями данного набора n признаков [2]. Предпологаем далее, что каждый признак количественный и во множестве его значений определена квази-метрика. Это предположение делается лишь только для простоты; все указанные ниже методы нетрудно обобщить на другие типа признаков. Объекты, которые являются „информацией'' для алгоритма, называются *обучающими объектами*. Об каждом из них известно, к которому классу он принадлежит. Других информации о составе классов нет.

Рассмотрим таблицу $T_{n,m,l}$, строки которой отвечают данным m обучающим объектам, столбцы — n признакам; объекты разделены на l классов. В дальнейшем часто отождествляем объекты со строками таблицы $T_{n,m,l}$. Обозначим далее: $S_i = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ — произвольная строка, принадлежащую таблице $T_{n,m,l}$. Пусть $S = (\beta_1, \beta_2, \ldots, \beta_n)$ — произвольная строка, отвечающая распознаваемому объекту. Пусть далее ϱ_q $(q = 1, 2, \ldots, n)$ обозначает данную квази-метрику для пространства значений q-го признака.

Семейство алгоритмов, решающих данную задачу, определяется поэтапно [3]. Все величины, представленные ниже, имеют несколько разных видов [4]. Здесь представим только один вид каждой из этих величин.

1. *Система опорных множеств.* Рассмотрим множество 2^N, где $N = \{1, 2, \ldots, n\}$. Системой опорных множеств алгоритма A (обозначаемой Ω_A) является совокупность всех подмножеств множества 2^N мощности k, $1 \leqslant k \leqslant n-1$.

2. *Функция близости.* Функцию близости $r(\tilde{w}S, \tilde{w}S_l)$, определенной для $\tilde{w}$ — частей ([4]) строк S_l и S, вводим следующим образом: рассмотрим неравенства $\varrho_q(\alpha_q, \beta_q) \leqslant \varepsilon_q$, $q = j_1, j_2, \ldots, j_k$ (т.е. для данной $\tilde{w}$ — части). Обозначим число Q невыполненных неравенств. Тогда:

$$r(\tilde{w}S, \tilde{w}S_l) = \begin{cases} 1, & \text{если} \quad Q \leqslant \varepsilon, \\ 0, & \text{если} \quad Q > \varepsilon. \end{cases}$$

Параметры $\varepsilon, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ участвуют вместе с введенным ранее параметром в задании семейства алгоритмов.

3. *Оценка $\Gamma^j_{\tilde{w}}(S, S_l)$.* Введем параметры $p_1, p_2, \ldots, p_n$, соответствующие отделным столбцам таблицы $T_{n,m,l}$, и параметры $\gamma_{i_1}, \gamma_{i_2}, \ldots, \gamma_{i_l}$, соответствующие i-той строке данной таблицы, $i = 1, 2, \ldots, m$. Введенные параметры тоже описывают свойства алгоритма.

$$\Gamma^j_{\tilde{w}}(S, S_l) = \gamma_{lj}(p_{j_1} + p_{j_2} + \ldots + p_{j_k}) \cdot r(\tilde{w}S, \tilde{w}S_l).$$

4. *Оценка $\Gamma_j(S)$. Имеем*

$$\Gamma_j(S) = \frac{1}{N_j} \sum_{S_l \in K_j} \sum_{\Omega_{\tilde{w}} \in \Omega_A} \Gamma^j_{\tilde{w}}(S, S_l).$$

Здесь $1/N_j$ — нормирующий множитель, $j = 1, 2, \ldots, l$. Этим способом мы сопоставили каждому классу K_j таблицы $T_{n,m,l}$ и объекту S некоторые оценки $\Gamma_j(S)$. Нашей задачей является в данном случае указать номера классов, к которым пбинадлежит объект S.

Вводится решающее правило следующим образом: алгоритм A (определенный конкретными значениями указанных выше параметров и параметров δ_{ij}, выступающих далее) относит S к классу K_j, если $\Gamma_j(S) \cdot \left[\sum_{i=1}^{l} \Gamma_i(S)\right]^{-1} \geqslant \delta_{1j}$. A не относит S к классу K_j, если $\Gamma_j(S) \cdot \left[\sum_{i=1}^{l} \Gamma_i(S)\right]^{-1} \leqslant \delta_{2j}$. A отказывается от распознавания S относительно K_j в оставшихся случаях. Представленный метод позволяет определить удобные, эффективные процедуры вычисления оценок $\Gamma_j(S)$. Обозначим через $Q^l = \{q_1^l, q_2^l, \ldots, q_{r_l}^l\}$ множество номеров тех столбцов, для которых $\varrho_u(\alpha_u, \beta_u) \leqslant \varepsilon_u$; $u \in Q^l$.

Обозначим далее:

$$N - Q^l = N^l; \quad \sum_{u \in Q^l} p_u = P_Q^l; \quad \sum_{u=1}^{n} p_u - P_Q^l = P_N^l.$$

Тогда мы имеем следующую теорему:

Теорема 1 (за [3]). *Имеем*

$$\Gamma_j(S) = \frac{1}{N_j} \sum_{S_l \in K_j} \gamma_{lj} \sum_{t=0}^{\varepsilon} \left(c_{r_l-1}^{k-t-1} \cdot c_{n-r_l}^{t} \cdot P_Q^l + c_{n-r_l-1}^{t-1} \cdot c_{r_l}^{k-t} \cdot p_N^l\right).$$

Кроме таблицы $T_{n,m,l}$ вводится т.н. таблица контроля составлена из некоторого числа контрольных объектов, разделеных на l классов и описанных значениями n признаков (этих самых, что объекты из таблицы $T_{n,m,l}$). Это позволяет сформулировать некоторый критерий определения точности распознавания [2]. Такой критерий представляется в форме функционала качества φ, определенного на семействе алгоритмов. Функционал φ, вообще говоря, оценивает точность распознавания на данной таблице контроля [1]. Алгоритм, реализующий экстремальное значение функционала качества называем *оптимальным* и считаем решением задачи оптимализации в данном семействе алгоритмов распознавания.

II. Неопределенные значения признаков

В задачах распознавания довольно часто выступают неопределенные значения признаков. Они выступают так в таблицах $T_{n,m,l}$ и $T'_{n,m',l}$, как и в формальном описании распознаваемого объекта S.

Существуют две самые типичные методы трактовки неопределенных значений признаков (т.н. далее „прочерков"). Все неравенства $\varrho(\alpha, \beta)$, в которых хотя бы одно из значений является прочерком — по определению считаем выполнеными или невыполнеными. Ниже мы укажем ряд других методов.

1. Игнорирование прочерков. Игнорирование прочерков состоит в том, что все элементы системы опорных множеств Ω_A, которые для данной строки S_l и строки S содержат хотя бы один столбец с прочерком, не рассматриваются при конструкции оценки $\Gamma_j(S)$. Здесь возможные два разных подхода к игнорированию прочерков:

(а) лишние элементы системы Ω_A удаляются во время вычисления оценки $\Gamma_j(S)$. Ω_A задана для всей таблицы,

(б) система опорных множеств задана отдельно для каждой строки; систему опорных множеств для строки S_l обозначаем Ω_A^l. Оценку для каждой строки (класса) нормируется потому, что мощности различных Ω_A^l для разных i — различны.

Эти два подхода отличаются между собой только формально. Можно доказать, что оценка $\Gamma_j(S)$ в этих двух случаях является одинаковой. Можно также показать эффективную формулу для вычисления оценок.

2. Метод усреднения прочерков. Метод усреднения прочерков состоит в том, что каждому прочерку приписываем среднее значение соответствующего признака. Это может быть его среднее значение для данного класса объектов (так, как каждый прочерк в таблице $T_{n,m,l}$ связан с конкретным объектом), или среднее значение для данной таблицы. Кроме того здесь можно применять много разных определений самого среднего значения. Однако независимо от принятого определения в результате прочерк становится конкретным числовым значением, сравнение которого с любыми другими значениями признаков в данных квази-метрических пространствах вполне возможно. Сохраняя эффективность процедур для вычислений оценок $\Gamma_j(S)$ можно употреблять больше чем один параметров ε (например разные значения ε в зависимости от значений (прочерк или нет) признаков для пар: объект из таблицы $T_{n,m,l}$, распознаваемый объект).

3. Оптимизация прочерков. Для этой группы методов приписывается прочерком такие числовые значения, чтобы для них отыскивалась наибольшая точность распознавания для объектов из таблицы контроля $T'_{n,m',l}$, т.е. отыскивалось экстремальное значение функционала качества φ. Тогда значения прочерков, т.е. всех прочерков в данном столбце, всех прочерков в данном столбце и классе, или просто всех прочерков, являются новыми параметрами, определяющими множество алгоритмов. Для задач распознавания большой размерности указанный групп методов имеет лишь только теоретическое значение из-за большого числа операции на ЭВМ.

4. Изменения в определении множества алгоритмов. Здесь можно указать два метода: изменения в определению функции близости и изменения в характере учета параметров ϱ_l в оценке $\Gamma^J_{\tilde{w}}(S, S_l)$. Для примера подробно опишем первый из этих методов.

Предположим, что два неравенства $\varrho_i(\alpha_i, \beta_i) \leqslant \varepsilon_i$ и $\varrho_j(\alpha_j, \beta_j) \leqslant \varepsilon_j$, где хотя бы одно из значений α, β в каждом из указанных выше неравенств не определено, мы считаем за одно выполненное неравенство. Это предположение влияет на изменение определения функции близости. Пусть:

$$\eta_i = \begin{cases} 0 & \text{если} \quad \varrho_i(\alpha_i, \beta_i) \leqslant \varepsilon_i \text{ и } \alpha_i, \beta_i \neq \text{„}-\text{"}, \\ 1 & \text{если} \quad \varrho_i(\alpha_i, \beta_i) > \varepsilon_i \text{ и } \alpha_i, \beta_i \neq \text{„}-\text{"}, \\ 1/2 & \text{если} \quad \alpha_i = \text{„}-\text{"} \quad \text{или} \quad \beta_i = \text{„}-\text{"} \end{cases}$$

(„$-$" обозначает прочерк), Введем обозначение $\eta = \sum_{l=1}^{k} \eta_{j_l}$, где, как прежде, набор $\{j_1, j_2, \ldots, j_k\}$ обозначает любой элемент $\Omega_{\tilde{w}}$ системы опорных множеств Ω_A. Определяем следующую функцию близости:

$$r(\tilde{w}S, \tilde{w}S_l) = \begin{cases} 1 & \text{если} \quad \eta \leqslant \varepsilon, \\ 0 & \text{если} \quad \eta > \varepsilon. \end{cases}$$

Пусть $V^l = \{v_1^l, v_2^l, \ldots, v_{z_l}^l\}$ обозначает множество номеров этих столбцов, для которых в i-той строке таблицы $T_{n,m,l}$ значение соответствующего признака не определено. Для строки S аналогичное множество обозначим V^0. Пусть $V^{0l} = V^0 \cup V^l$; $\|V^{0l}\| = z_{0l}$. Сохраняем из I обозначения Q^l и P_Q^l. Введем другие обозначения:

$$\sum_{u \in V^{0l}} = p_u = P_V^{0l}, \quad N - (V^{0l} \cup Q^l) = N^l,$$

$$n - r_l - z_{0l} = n_l, \quad \sum_{u=1}^{n} p_u - P_Q^l - P_V^{0l} = P_N^l.$$

Принимаем далее:

$$\sum_{p=0}^{2\varepsilon-2t} C_{r_l-1}^{k-t-p-1} \cdot C_{z_{0l}}^{p} = C_{1,0}^{t},$$

$$\sum_{p=0}^{2\varepsilon-2t} C_{r_l}^{k-t-p} \cdot C_{z_{0l}-1}^{p-1} = C_{0,1}^{t},$$

$$\sum_{p=0}^{2\varepsilon-2t} C_{r_l}^{k-t-p} \cdot C_{z_{0l}}^{p} = C_{0,0}^{t}.$$

Докажем тогда следующую теорему:

Теорема 2. *Имеем*

$$\Gamma_j(S) = \frac{1}{N_j} \sum_{S_l \in K_j} \gamma_{lj} \sum_{t=0}^{\varepsilon} [P_Q^l \cdot C_{1,0}^t \cdot C_{n_l}^t + P_V^{0l} \cdot C_{0,1}^t \cdot C_{n_l}^t + P_N^l \cdot C_{0,0}^t \cdot C_{n_l-1}^{t-1}].$$

Доказательство. Выясним, какие наборы из k столбцов вносят ненулевой вклад в величину $\Gamma_j(S)$. Предположим, что мы взяли $k-t$ столбцов из множества $Q^l \cup V^{0l}$ и t столбцов из множества оставшихся n_l столбцов. Наборы с ненулевым вкладом мы будем искать, меняя t от 0 до ε. Предположим, следовательно, что среди выбранных $k-t$ столбцов p столбцов принадлежит множеству V^l, а оставшихся $k-t-p$ столбцов принадлежит множеству Q^l. Заметим, что этих последних надо взять не меньше, чем $k-2\varepsilon+t$, что вытекает из решения неравенства $x+1/2(k-t-x) \geqslant k-\varepsilon$. Оттуда следует, что $k-t-p \geqslant k-2\varepsilon+t$. Ясно тогда, что p меняется от 0 до $2\varepsilon-2t$.

Каждый столбец из числа вошедших в множество Q^l входит ровно в $\left(\sum_{p=0}^{2\varepsilon-2t} C_{r_l-1}^{k-t-p-1} \cdot C_{z_{0l}}^{p}\right) \cdot C_{n_l}^t = C_{1,0}^t \cdot C_{n_l}^t$ наборов, вносящих ненулевой вклад в величину $\Gamma_j(S)$. Рассмотрим столбец с номером u, $u \in Q^l$. Вклад такого столбца в оценку $\Gamma_j(S)$ очевидно равен:

$$\frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \cdot p_u \cdot C^t_{1,0} \cdot C^t_{n_l}.$$

Аналогично, вклады столбцов принадлежащих множествам V^{0l} и N^l соответственно равны:

$$\frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \cdot p_v \cdot C^t_{0,1} \cdot C^t_{n_l},$$

и

$$\frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \cdot p_\tau \cdot C^t_{0,0} \cdot C^{t-1}_{n_l-1},$$

где $v \in V^{0l}$, $C \in N^l$. Суммарный вклад при суммировании по всем множествам $\Omega_{\tilde{w}}$ составленным из $k-t$ столбцов из множества $Q^l \cup V^{0l}$ и t столбцов из множества N^l, при сравнении строк S и S_l ($S_l \in K_j$), есть:

$$\frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj}\Big[\sum_{u \in Q^l} p_u \cdot C^t_{1,0} \cdot C^t_{n_l} + \sum_{v \in V^{0l}} p_v \cdot C^t_{0,1} \cdot C^t_{n_l} + \sum_{\tau \in N^l} p^\tau \cdot C^t_{0,0} \cdot C^{t-1}_{n_l-1}\Big] =$$

$$= \frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj}[P^l_Q \cdot C^t_{1,0} \cdot C^t_{n_l} + P^{0l}_V \cdot C^t_{0,1} \cdot C^t_{n_l} + P^l_N \cdot C^t_{0,0} \cdot C^{t-1}_{n_l-1}] = \frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \cdot B^t_l.$$

Для того, чтобы получить общий вклад в оценку $\Gamma_j(S)$ при сравнении S и S_l ($S_l \in K_j$), достаточно взять:

$$\sum_{t=0}^{\varepsilon} \frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \cdot B^t_l = \frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \sum_{t=0}^{\varepsilon} B^t_l =$$

$$= \frac{1}{N_j}\sum_{S_l \in K_j} \gamma_{lj} \sum_{t=0}^{\varepsilon} [P^l_Q \cdot C^t_{1,0} \cdot C^t_{n_l} + P^{0l}_V \cdot C^t_{0,1} \cdot C^t_{n_l} + P^l_N + C^t_{0,0} \cdot C^{t-1}_{n_l-1}].$$

Теорема доказана.

Следует заметить, что предположение о счете двух неравенств с прочерками за одно выполненное неравенство без всякой трудности можно обобщить на случай произвольного числа неравенств.

5. Метод исправления прочерков. Идеей этого метода является приписывание прочеркам в таблице обучения таких числовых значений, чтобы независимо от процесса оптимизации параметров алгоритма получить наибольшую точность распознавания для объектов из таблицы контроля. Рассмотрим этот метод сперва в случае упрощенной модели распознавания, а затем

в общем случае. Описываемый метод требует более точных обозначений, чем методы введенные до сих пор.

5.1. Рассмотрим таблицу обучения $T_{n,m,l}$ и объект $S = (\beta_1, \beta_2, \ldots, \beta_n)$. Выбираем любой объект $S_l \in T_{n,m,l}$; $S_l = (\alpha^l_1, \alpha^l_2, \ldots, \alpha^l_n)$. Рассмотрим далее множество неравенств:

$$(1) \qquad \varrho_i(\beta_j, \alpha^l_j) \leqslant \varepsilon_j, \quad j = 1, 2, \ldots, n.$$

Предполагаем, что эти неравенства выполнены для столбцов с номерами $i_1, i_2, \ldots, i_k$, причем все неравенства, содержавшие прочерки, трактуем как невыполненные (в данной модели это равнозначно с игнорированием прочерков). Введем следующую функцию близости:

$$r(S, S_l) = \gamma(S_l) \cdot (p_{i_1} + p_{i_2} + \ldots + p_{i_{k_l}}).$$

Оценку $\Gamma_j(S)$, $j = 1, 2, \ldots, l$, определяем следующим способом:

$$\Gamma_j(S) = \sum_{S_l \in K_j} \gamma(S_l)(p_{i_1} + p_{i_2} + \ldots + p_{i_{k_l}}).$$

Вычислим теперь насколько метод игнорирования неопределенных значений признаков в таблице $T_{n,m,l}$ уменьшает оценки $\Gamma_j(S)$. Рассмотрим класс K_j и любой объект $S_l \in K_j$. Пусть далее $S_l = (\alpha^l_1, \alpha^l_2, \ldots, \alpha^l_n)$, $S = (\beta_1, \beta_2, \ldots, \beta_n)$. Предположим что $\alpha^l_u =$ „$-$", причем $\beta_u \neq$ „$-$". Нетрудно заметить, что если α^l_u являлось бы числовым значением, расстояние которого от β_u не больше ε_u, то оценка $\Gamma_j(S)$ увеличилась бы на число $\gamma(S_l) \cdot p_u$.

Теперь вычислим, насколько потенциально уменьшают оценку $\Gamma_j(S)$ все прочерки из таблицы обучения. Нетрудно заметить, что эти убытки описывает величина:

$$x_S(K_j) = \sum_{S_l \in K_j} \sum_{u:\, \alpha^l_u = „-"} \gamma(S_l) \cdot p_u.$$

Конечно, это не все убытки оценок $\Gamma_j(S)$. Прочерки в объекте S тоже вызывают их уменьшение. Для того, чтобы оценить убытки, введены игнорированием прочерков объекта S, необходимая некоторая конструкция.

Рассмотрим u-ый столбец таблицы $T_{n,m,l}$ и все значения соответствующего признака в j-ом классе. Если множество M_u значений этого признака конечное, то запоминаем все значения из j-го класса вместе с их кратностями, т.е. с числами, определяющими, сколько раз выступило данное значение в $T_{n,m,l} \cup K_j$ для u-го признака. Если M_u бесконечное, выбираем все значения из $T_{n,m,l} \cup K_j$ и устанавливаем в растущий ряд, запоминая их кратности. К каждому значению α^l_u этого ряда добавляем два новых значения: $\alpha^l_u - \varepsilon_u$, $\alpha^l_u + \varepsilon_u$ и выбираем те из них, которые входят в данный отрезок M_u. Таким образом, мы разбили отрезок M_u на несколько подотрезков. Середины этих подотрезков определяют множество, которое обозначаем как K^u_j. Для каждого элемента ряда K^u_j мы можем теперь оценить, сколько неравенств (1) выполнено для этого элемента (учитывая, конечно, кратности значений) и какая величина, полученная из суммирования соответствующих $\gamma(S_i)\cdot p_u$ отвечает данному элементу, причем все прочерки в $T_{n,m,l}$ трактуем как не выполняющие неравенства (1). Наибольшую из таких величин обозначаем как $x_j(\beta_u)$. Пусть:

$$x_j(S) = \sum_{u:\beta_u=,,-''} x_j(\beta_u).$$

Определенная этим способом величина $x_j(S)$ описывает уменьшение оценки $\Gamma_j(S)$ игнорированием прочерков объекта S.

Вообще говоря, метод игнорирования прочерков, потенциально уменьшает оценку $\Gamma_j(S)$ на величину $x_s(K_j)+x_j(S)$, $j = 1, 2, \ldots, l$. Рассмотрим таблицу обучения $T_{n,m,l}$ и таблицу контроля $T'_{n,m',l}$. Каждому объекту $S'_t \in T'_{n,m',l}$ приписываем две величины:

$$\Gamma_j(S'_t) \quad \text{и} \quad \Gamma_j(S'_t)+x_{S'_t}(K_j)+x_j(S'_t), \qquad j = 1, 2, \ldots, l.$$

Имея решающее правило, мы можем сразу узнать, дают ли эти две величины разные результаты распознавания объекта S'_t для классов $K_1, K_2, \ldots, K_l$, или нет. Если объект S'_t дает разные результаты распознавания для класса K_j, назовем его *вариантным* для этого класса; в противном случае *инвариантным*. Таким образом мы сможем каждому объекту S'_t из таблицы контроля сопоставить ряд $K_1^t, K_2^t, \ldots, K_{j_t}^t$ классов, для которых S'_t является вариантным объектом. Наоборот, можно получить оттуда для каждого класса множество его вариантных объектов. Класс, который содержит хотя бы один вариантный объект, назовем *вариантным классом*.

Как мы уже заметили раньше, нашей целью является приписывание неопределенным значениям признаков в $T_{n,m,l}$ таких числовых значений, чтобы получить наибольшую точность распознавания. Оказывается, что это возможно только для прочерков из вариантных классов.

Рассмотрим класс $K_j \in T'_{n,m',l}$ и обозначим как $S'_{j_1}, S'_{j_2}, \ldots, S'_{j_{t_j}}$ вариантные для класса K_j объекты из $T'_{n,m',l}$. Множество этих объектов можно разделить на два подмножества: A_j^1 и A_j^2; к множеству A_j^1 принадлежат те объекты из $S'_{j_1}, S'_{j_2}, \ldots, S'_{j_{t_j}}$ которые в таблице $T'_{n,m',l}$ принадлежали j-му классу; к множеству A_j^2 принадлежат оставшиеся объекты.

Рассмотрим, следовательно, произвольный объект $S_i \in T_{n,m,l}$, $S_i = (\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i)$ и такое значение α_u^i, чтобы $\alpha_u^i = ,,-''$. Пусть $\{K\}^i$ обозначает множество классов, к которым принадлежит объект S_i (множество $\{K\}^i$ определяют единицы классификационного вектора объекта S_i ([3])). Рассмотрим класс K_j, $K_j \in \{K\}^i$. Пусть $(A_j^1)^u$ и $(A_j^2)^u$ обозначают соответственно те подмножества A_j^1 и A_j^2, для которых объекты не содержат прочерков в u-том столбце. Мощности этих подмножеств обозначим как $(a_j^1)^u$ и $(a_j^2)^u$.

Пусть далее $S'_t = (\beta_1^t, \beta_2^t, \ldots, \beta_n^t)$ обозначает произвольный объект из таблицы контроля $T'_{n,m',l}$. Введем функцию g, определенную на значениях

объектов из $T_{n,m,l}$ и $T'_{n,m',l}$ следующим образом:

$$g(\alpha_u^i, \beta_u^t) = \begin{cases} 1 & \text{если} \quad \varrho_u(\alpha_u^i, \beta_u^t) \leqslant \varepsilon_u,\ \alpha_u^i, \beta_u^t \neq ,,-'', \\ 0 & \text{во всех остальных случаях.} \end{cases}$$

Обозначим как $\tilde{\alpha}_u^i$ числовое значение, которое приписываем прочерку α_u^i. Введем следующий функционал качества значения $\tilde{\alpha}_u^i$ прочерка α_u^i:

$$\psi_1(\tilde{\alpha}_u^i) = \sum_{K_j\in\{K\}^i}\left[\left(\sum_{S'_t\in(A_j^1)^u}\frac{g(\tilde{\alpha}_u^i, \beta_u^t)}{(a_j^1)^u}\right)^{\mu} - \left(\sum_{S'_t\in(A_j^2)^u}\frac{g(\tilde{\alpha}_u^i, \beta_u^t)}{(a_j^2)^u}\right)^{\mu}\right],$$

где $\mu > 1$ — некоторый параметр. Выбираем такое числовое значение $\tilde{\alpha}_u^i$, чтобы значение функционала ψ_1 достигало своего максимума.

Идея представленного выше метода выбора значения $\tilde{\alpha}_u^i$ состоит в том, что выбираются такие значения $\tilde{\alpha}_u^i$, чтобы они были ближе к значениям вариантных объектов, которые в таблице $T'_{n,m',l}$ принадлежат классам из множества $\{K\}^i$ и возможно дальше от значений вариантных объектов вне множества $\{K\}^i$.

Заметим, что среди всех неопределенных значений признаков, принадлежащих вариантным классам, могут существовать такие, которые не попали в вышеуказанную процедуру. Это возможно только в одном случае. Допустим, например, что $\alpha_u^i = ,,-''$. Видно, что, рассматривая функционал ψ_1, мы не сумели приписать прочерку α_u^i числового значения, если:

(2) $\qquad (A_j^1)^u = \varnothing \quad$ и $\quad (A_j^2)^u = \varnothing \quad$ для всех j таких, что $K_j \in \{K\}^i$.

Как мы заметили прежде, каждому неопределенному значению β_u^t объекта $S'_t \in T'_{n,m',l}$ мы можем приписать некоторое множество величин $x_j(\beta_u^t)$, характеризующих размеры убытков оценок $\Gamma_j(S'_t)$, вызванных игнорированием прочерков. Затем для оставшихся неопределенных значений α_u^i в таблице обучения, т.е. исполняющих (2), приписываем такое числовое значение, чтобы достигало максимума значение следующего функционала ψ_2:

$$\psi_2(\tilde{\alpha}_u^i) = \sum_{K_j\in\{K\}^i}\left[\sum_{S'_t\in A_j^1} x_j(\beta_u^t) - \sum_{S'_t\in A_j^2} x_j(\beta_u^t)\right].$$

5.2. Рассмотрим теперь общий случай, т.е. модель распознавания, определенную, как в 5.1. Рассмотрим произвольный объект $S_i \in T_{n,m,l}$, $S_i = (\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i)$ и произвольный объект $S'_t \in T'_{n,m',l}$, $S'_t = (\beta_1^t, \beta_2^t, \ldots, \beta_n^t)$. Рассмотрим далее неравенства $\varrho_p(\alpha_p^i, \beta_p^t) \leqslant \varepsilon_p$, $p = 1, 2, \ldots, n$. Предположим, что r_{it} из этих неравенств выполнены. Множество $\{q_1^{it}, q_2^{it}, \ldots, q_{r_{it}}^{it}\}$ столбцов, для которых эти неравенства выполнены, обозначим как Q^{it}.

Пусть $V^i = \{v_1^i, v_2^i, \ldots, v_{z_i}^i\}$ обозначает множество столбцов, в которых для объекта S_i находятся прочерки. Как $\tilde{V}^t$ обозначим аналогичное множество для объекта S'_t; $\tilde{V}^t = \{\tilde{v}_1^t, \tilde{v}_2^t, \ldots, \tilde{v}_{z_t}^t\}$.

Сумму $V^i\cup\tilde{V}^t$ множеств V^i и $\tilde{V}^t$ обозначим как V^{it}, мощность множества

V^{it} обозначим как z_{it}. Введем кроме того дополнительные обозначения:

$$N-(Q^{it}\cup V^{it}) = N^{it} \quad \text{и} \quad n-r_{it}-z_{it} = n_{it}.$$

Предположим, что при трактовке произвольного неравенства, которое содержит хотя бы один прочерк, как невыполненное, мы получили оценки $\Gamma_j(S_t')$, $j = 1, 2, \ldots, l$. Нашей целью является ответ на вопрос, насколько факт пропускания неравенств с прочерками потенциально уменьшил эти оценки.

Рассматриваемый общий случай не позволяет селекционировать объекты в классе, если в каком-то столбце объекта S_t' находится прочерк. В связи с этим все неравенства, в которых сответственное значение β_u^t равно „—", трактуем как выполненное. Введем обозначения:

$$\sum_{u\in Q^{it}} p_u = P_Q^{it}; \quad \sum_{u\in N^{it}} p_u = P_N^{it}; \quad \sum_{u\in V^{it}} p_u = P_V^{it}.$$

Тогда мы имеем следующую теорему:

Теорема 3.2. *Оценка $\Gamma_j(S_t')$ уменьшена потенциально на следующую величину $x_j(S_t')$:*

$$x_j(S_t') = \sum_{S_l\in K_j} \gamma(S_l) \sum_{t=1}^{z_{lt}} (P_Q^{it}\cdot C_{r_{lt}-1}^{k-\varepsilon-\tau-1}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}}^{\tau} +$$

$$+ C_{n_{lt}-1}^{\varepsilon-1}\cdot P_N^{it}\cdot C_{n_{lt}-1}^{k-\varepsilon-\tau}\cdot C_{z_{lt}}^{\tau} + P_V^{lt}\cdot C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}-1}^{\tau-1}).$$

Доказательство. При вычислении оценок $\Gamma_j(S_t')$ мы суммировали параметры p_u по всем таким наборам, для которых было выполено не менее $k-\varepsilon$ неравенств. Теперь требуется учитывать те наборы, для которых не выполнено соответственно $\varepsilon+1, \varepsilon+2, \ldots, \varepsilon+z_{lt}$ неравенств, причем эти невыполненные неравенства должны содержать хотя бы одно неопределенное значение признака.

Допустим, что для объекта S_t' мы взяли τ таких признаков из множества V^{lt}. В связи с этим мы должны взять $k-\varepsilon-\tau$ признаков из множества Q^{lt} и ε признаков из множества N^{lt}. Это даст вместе $C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}}^{\tau}$ разных наборов.

Параметр p_u, соответствующий признаку из множества Q^{lt}, выступает в $C_{r_{lt}-1}^{k-\varepsilon-\tau-1}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}}^{\tau}$ наборах, затем учет этого параметра в величине $x_j(S_t')$ есть: $p_u\cdot C_{r_{lt}-1}^{k-\varepsilon-\tau-1}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}}^{\tau}$. Аналогично, учеты параметров p_s и p_t, соответственно для признаков из множеств V^{lt} и N^{lt}, заносят следующие учеты до величины $x_j(S_t')$:

$$p_S\cdot C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}-1}^{\tau-1} \quad \text{и} \quad p_t\cdot C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}-1}^{\varepsilon-1}\cdot C_{z_{lt}}^{\tau}.$$

Оттуда видно, что прочерки объектов S_l и S_t' уменьшили оценку $\Gamma_j(S_t')$ на следующую величину:

$$(3) \qquad x_j(S_l, S_t') = \gamma(S_l) \sum_{\tau=1}^{z_{lt}} (P_Q^{lt}\cdot C_{r_{lt}-1}^{k-\varepsilon-\tau-1}\cdot C_{n_{lt}}\cdot C_{z_{lt}}^{\tau} +$$

$$+ P_N^{lt}\cdot C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}-1}^{\varepsilon-1}\cdot C_{n_{lt}}^{\tau} + P_V^{lt}\cdot C_{r_{lt}}^{k-\varepsilon-\tau}\cdot C_{n_{lt}}^{\varepsilon}\cdot C_{z_{lt}-1}^{\tau-1}).$$

Чтобы получить величину перемены оценки $\Gamma_j(S_t')$ для j-го класса, т.е. $x_j(S_t')$, требуется суммировать выражение (3) по всем объектам, принадлежащим j-му классу. Теорема доказана.

Каждому объекту $S_l \in T_{n,m,l}$ и объекту $S_t' \in T'_{n,m',l}$ припишем теперь некоторый параметр $W_j(S_l, S_t')$ следующим образом. Предположим, что $S_l \in K_j$; тогда:

$$W_j(S_l, S_t') = \frac{x_j(S_l, S_t')}{\Gamma_j(S_t')+x_j(S_l, S_t')}\cdot\frac{1}{z_{lt}}.$$

Введенная величина характеризует вес каждого из прочерков для признаков из множества V^{lt} для j-го класса.

Дальнейшая часть метода вполне аналогична указанной выше в 5.1. В сущности вводим идентичное определение вариантных и инвариантных объектов и классов и определение функции g. Надо одновременно заметить, что если существует (2), тогда не имеем возможности дифференцировать объекты из таблицы обучения для прочерков из объектов таблицы контроля. Оттуда следует, что мы не рассматриваем функционала φ_2. Принимаем здесь:

$$\psi_1(\tilde\alpha_u^l) = \sum_{K_j\in\{K\}^l} \left[\left(\sum_{S_t'\in(A_j^1)^u} \frac{g(\tilde\alpha_u^l, \beta_u^t)}{(a_j^1)^u}\cdot W_j(S_l, S_t')\right)^{\mu} - \left(\sum_{S_t'\in(A_j^2)^u} \frac{g(\tilde\alpha_u^i, \beta_u^t)}{(a_j^2)^u}\cdot W_j(S_l, S_t')\right)\right]^{\mu}.$$

Выбираем такое числовое значение $\tilde\alpha_u^l$ прочерка $\tilde\alpha_u^i$, чтобы значение функционала ψ_1 достигало своего максимума.

Следует заметить, что представленный метод выбора оптимальных в смысле точности распознавания значений прочерков совсем не зависит от оптимизации параметров, задающих семейство алгоритмов. Указанная процедура дополняет процесс оптимизации, ее результаты по крайней мере не ухудшают точности распознавания.

Литература

[1] Ю. И. Журавлев, *Экстремальные задачи, возникающие при обосновании эвристических процедур*, Сб. *Проблемы прикладной математики и механики*, Наука, Москва 1971.

[2] Ю. И. Журавлев, М. М. Камилов, Ш. Е. Туляганов, *Алгоритмы вычисления оценок и их применение*. Фан, Ташкент 1974.

[3] Ю. И. Журавлев, М. Михалевич, *Алгоритмы распознавания, основанные на вычислении оценок для задач с пересекающимися классами*, Труды ВЦ ПАН 145 (1974).
[4] Ю. И. Журавлев, В. В. Никифоров, *Алгоритмы распознавания, основанные на вычислении оценок*, Кибернетика 3 (1971).

Presented to the Semester
Discrete Mathematics
(February 15–June 16, 1977)

DISCRETE MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 7
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1982

СИНТЕЗ ЛОГИЧЕСКИХ СЕТЕЙ ДЛЯ РЕАЛИЗАЦИИ КЛАССОВ БУЛЕВЫХ МАТРИЦ С ДАННЫМ ЧИСЛОМ УГЛОВЫХ КЛЕТОК

И. ГАВЕРЛИК

Кафедра теоретической кибернетики, Университет им. Коменского, Братислава, Чехословакия

В картографии при автоматическом строении карт решаются проблемы строения разного типа изолиний. Из изходных данных конструируются базовые — первичные, и из них вторичные [1], [2], [3]. При этом образуются разного типа плоские фигуры. Одна из задач состоит в том, чтобы узнать, входит ли точка с заданными координатами в фигуру или нет. Важный вопрос состоит в определении сложности реализации такого типа характеристических функций. При этом используется некоторое геометрическое представление булевых матриц.

В более общем случае в теории распознавания образов рассматриваются и проблемы восприятия и обработки зрительной информации, в частности плоских объектов. В связи с обработкой зрительной информации Ф. Этниф [8] высказал предположение, что в распознавании формы наиболее важную роль играют те точки, в которых контурные линии меняют свое направление или обрываются. Исходя из этого тезиса введем понятие их дискретного аналога в том частном случае, когда имеется дело с булевыми матрицами.

Геометрическая постановка задачи рассматриваемой в этой работе состоит в следующем: Имеется решетка размера $N \times N$, т.е. N строк и в каждой строке N клеток (или $N+1$ строк и в каждой строке $N+1$ узлов). Строки и столбцы клеток решетки занумерованы в естественном порядке числами от 0 до $N-1$. Каждой клетке решетки приписаны координаты: номер строки и номер столбца решетки, в которых она находится.

В естественном порядке занумерованы и строки и столбцы узлов решетки числами от 0 до N. Узлу решетки приписаны координаты: номер строки и номер столбца (в нумерации строк и столбцов), в которых он находится.

Пусть клеткам решетки приписаны значения из $\{0, 1\}$. Значения клеток тогда образуют некоторую квадратную булеву матрицу порядка N. Класс всех булебых квадратных матриц порядка N обозначим через $\mathfrak{N}_N$.