О приложениях некоторых линейных процедур распознавания

Азар Шайеб

В классе метрических алгоритмов распознавания выделяется подкласс линейных метрических алгоритмов. Типичным представителем этого класса является алгоритм H [6], основанный на метрике Хэмминга в пространстве бинарных признаков.

Усиленным вариантом такого алгоритма является предлагаемый в статье алгоритм $\mathcal{A}$. Помимо самой процедуры распознавания, этот алгоритм также включает в себя процедуру проверки и формирования свойств слабой и сильной $\mathcal{A}$-отделимости эталонных множеств. Он определяет число ошибок, при которых распознавание остается устойчивым, а также включает в себя процедуру сокращения исходного пространства признаков.

Введение

Создание вычислительных машин и связанное с этим ускоренное развитие математических теорий, в том числе математической кибернетики и дискретной математики позволило ставить и решать новые задачи, до недавнего времени находившиеся исключительно в компетенции человека. Одной из таких фундаментальных задач является рассматриваемая в настоящей работе задача распознавания образов.

В общем виде эта задача может быть сформулирована следующим образом: необходимо отнести предъявленный объект, определяемый некоторой совокупностью своих признаков, к одному из нескольких непересекающихся классов–образов. В том или ином виде данная задача решается человеком практически во всех сферах его деятельности.

Первые математические работы по данной задаче и реализованные на их основе технические системы появились во второй половине XX века и с тех пор активно используются во многих областях науки и техники, таких как геология, медицина, военное дело, социально-политические исследования и многое другое.

В работах по теории распознавания образов рассматриваются различные подходы к этой задаче. В частности, распознающие системы могут делиться по тому, доступны ли системе примеры объектов, принадлежащих к тому или иному классу (такие системы называются системами распознавания с обучением) или нет (системы распознавания без обучения). Другим критерием, по которому можно классифицировать такие системы, является принцип построения решающего правила. Исторически одними из первых и интуитивно наиболее понятных распознающих систем являлись линейные распознающие системы. В таких системах каждый объект представляется как точка в некотором многомерном пространстве, а решающее правило представляется в виде совокупности поверхностей, отделяющих области этого пространства, соответствующие различным классам. Простейшим случаем такой разделяющей поверхности является гиперплоскость. Системы, основанные на таком подходе, в литературе принято называть линейными распознающими системами.

В настоящей работе рассматривается линейная распознающая система с обучением.

Дальнейшее изложение будет организовано следующим образом.

В §1 будут изложены необходимые определения и обозначения. В §2 будет коротко описан базовый алгоритм распознавания H. В §3 будет предложен усовершенствованный алгоритм $\mathcal{A}$, позволяющий, помимо прочего, производить сокращение пространства признаков, не ухудшающее качество распознавания. Завершает работу описание приложений данного алгоритма, направления дальнейших исследований и список литературы.

1. Основные определения и обозначения

Введем некоторые обозначения, которые понадобятся нам для дальнейшего изложения:

$E = \{0,1\}$; $D = \{-1,1\}$; $\bar{D} = [-1,1]$; E^n, D^n, $\bar{D}^n$ — соответствующие n-мерные кубы в пространстве $\mathbb{R}^n$;

$\rho(\tilde{\alpha},\tilde{\beta})$ — метрика, порожденная нормой $|\bar{x}| = \sum\limits_{i=1}^{n} x_i$, $\bar{x} = (x_1,\dots,x_n) \in \mathbb{R}^n$;

$\|\bar{x}\|$ — евклидова норма для $\bar{x} \in \mathbb{R}^n$;

$(\bar{x},\bar{y})$ — скалярное произведение векторов $\bar{x}$ и $\bar{y}$ в $\mathbb{R}^n$.

$|M|$ — мощность множества M.

Сформулируем теперь необходимые определения. Всюду далее в этом разделе, если не оговаривается противное, предполагается, что $\tilde{\alpha} \in E^n$, $\tilde{\beta} \in E^n$, $M, M_1, M_2 \subseteq E^n$.

Величина

$$d(M) = \max_{\tilde{\alpha},\tilde{\beta} \in M} \rho(\tilde{\alpha},\tilde{\beta})$$

называется диаметром множества M.

Величина

$$\rho(M_1,M_2) = \min_{\tilde{\alpha} \in M_1, \tilde{\beta} \in M_2} \rho(\tilde{\alpha},\tilde{\beta})$$

называется расстоянием между множествами M_1 и M_2.

Пусть $\tilde{\alpha},\tilde{\beta} \in E^n$. Положим

$$\gamma(\tilde{\alpha},\tilde{\beta}) = n - \rho(\tilde{\alpha},\tilde{\beta}).$$

Очевидно, что $\gamma(\tilde{\alpha},\tilde{\beta})$ равно числу координат, в которых наборы $\tilde{\alpha},\tilde{\beta}$ совпадают.

Величина

$$\gamma(\tilde{\alpha},M) = \sum_{\tilde{\beta}_i \in M} \gamma(\tilde{\alpha},\tilde{\beta}_i)$$

называется числом голосов набора $\tilde{\alpha}$ относительно множества M.

Величина

$$\Gamma(\tilde{\alpha},M) = \frac{\gamma(\tilde{\alpha},M)}{|M|}$$

называется средним числом голосов набора $\tilde{\alpha}$ относительно множества M.

Величина

$$\rho(\tilde{\alpha}, M) = \sum_{\tilde{\beta}_i \in M} \rho(\tilde{\alpha}, \tilde{\beta}_i)$$

называется расстоянием между набором $\tilde{\alpha}$ и множеством M.

Величина

$$\mathcal{P}(\tilde{\alpha}, M) = \frac{\rho(\tilde{\alpha}, M)}{|M|}$$

называется средним расстоянием от набора $\tilde{\alpha}$ до множества M.

Очевидно, что при $\tilde{\alpha} \in E^n$ И $M \subseteq E^n$ имеет место

$$\Gamma(\tilde{\alpha}, M) = n - \mathcal{P}(\tilde{\alpha}, M).$$

2. Постановка задач и основные алгоритмы

Перейдем теперь к формальной постановке задачи классификации с эталонами в рамках комбинаторно-логического подхода к решению задач типа распознавания [1], [2].

2.1. Задача распознавания

Пусть K_1 и K_2 — два класса строк вида $(x_1, \ldots, x_n)$, $x_i \in E$, $i = 1, \ldots, n$. Номер i величины x_i будем называть признаком.

Пусть $L_1 = \{\tilde{\alpha}_1^{(1)}, \ldots \tilde{\alpha}_{m_1}^{(1)}\} \subseteq K_1$, $\quad L_2 = \{\tilde{\alpha}_1^{(2)}, \ldots \tilde{\alpha}_{m_2}^{(2)}\} \subseteq K_2$. Множества L_1 и L_2 называем материалом обучения.

Требуется, исходя из множеств L_1 и L_2, указать решающее правило $\to$, такое, что при $\tilde{\alpha} \in K_1 \cup K_2$ выполнено

$$\tilde{\alpha} \to \begin{cases} K_1, & \text{если } \tilde{\alpha} \in K_1; \\ K_2, & \text{если } \tilde{\alpha} \in K_2. \end{cases}$$

Эта задача называется задачей распознавания с обучением.

Если $\tilde{\alpha} \to K_1$, но $\tilde{\alpha} \notin K_1$ или $\tilde{\alpha} \to K_2$, но $\tilde{\alpha} \notin K_2$, говорят, что произошла ошибка распознавания.

2.2. H-отделимость

Рассмотрим алгоритм H, задаваемый следующим правилом:

$$\tilde{\alpha} \to \begin{cases} K_1, & \text{если } \Gamma(\tilde{\alpha}, L_1) > \Gamma(\tilde{\alpha}, L_2) \\ K_2, & \text{если } \Gamma(\tilde{\alpha}, L_1) < \Gamma(\tilde{\alpha}, L_2) \\ \text{не определено}, & \text{если } \Gamma(\tilde{\alpha}, L_1) = \Gamma(\tilde{\alpha}, L_2) \end{cases} \qquad (*)$$

Эти условия, соответственно, эквивалентны таким

$$\begin{aligned} &\mathcal{P}(\tilde{\alpha}, L_1) < \mathcal{P}(\tilde{\alpha}, L_2); \\ &\mathcal{P}(\tilde{\alpha}, L_1) > \mathcal{P}(\tilde{\alpha}, L_2); \\ &\mathcal{P}(\tilde{\alpha}, L_1) = \mathcal{P}(\tilde{\alpha}, L_2). \end{aligned}$$

Множества K_1 и K_2 называются слабо H-разделимыми (посредством L_1 и L_2), если для $L_1 \subseteq K_1$ и $L_2 \subseteq K_2$ алгоритм H все элементы из множества $L_1 \cup L_2$ правильно относит к L_1 и L_2.

Множества K_1 и K_2 называются сильно H-разделимыми, если для любых непустых $L_1 \subseteq K_1$, $L_2 \subseteq K_2$ множества K_1 и K_2 слабо H-разделимы (посредством L_1 и L_2).

Можно доказать следующее утверждение.

Теорема 1 ([6]). *Если $M_1, M_2 \subseteq E^n$, $|M_i| = m_i$, то*

а) M_1 и M_2 слабо H-разделимы, при

$$\frac{d(M_i)}{\rho(M_1, M_2)} < \frac{m_i}{m_i - 1}, \qquad i = 1, 2;$$

б) M_1 и M_2 сильно H-разделимы, при

$$\frac{d(M_i)}{\rho(M_1, M_2)} < 1, \qquad i = 1, 2.$$

2.3. Линейная отделимость

Пусть в E^n задана гиперплоскость G вида $(\tilde{\alpha}, x + \tilde{\beta}) = 0$, $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}^n$.

Будем говорить, что $\tilde{\gamma} \in E^n$ лежит в верхнем полупространстве, если $(\tilde{\alpha}, \tilde{\gamma} + \tilde{\beta}) > 0$; в нижнем полупространстве, если $(\tilde{\alpha}, \tilde{\gamma} +$

$\tilde{\beta}) < 0$. В том случае, если $(\tilde{\alpha}, \tilde{\gamma} + \tilde{\beta}) = 0$, считаем, что положение $\tilde{\gamma}$ не определено.

Говорят, что множества $M_1, M_2 \in E^n$ линейно отделимы, если существует гиперплоскость G, такая что M_1 и M_2 лежат в разных полупространствах по отношению к G, то есть эта гиперплоскость определяет решающее правило для точек из этих полупространств.

Замечание. В линейной отделимости без ограничения общности можно ограничиться рассмотрением только гиперплоскостей, проходящих через начало координат, то есть гиперплосткостей вида $(\tilde{\alpha}, x) = 0$.

Справедливо следующее утверждение.

Теорема 2 ([6]). *Если для $M_1, M_2 \in E^n$ выполнено правило (*), то M_1 и M_2 линейно отделимы.*

2.4. Сокращение поля признаков

Наряду с n-мерным кубом E^n нам будет удобно рассматривать n-мерный куб $D^n = \{-1, 1\}^n$.

Для множеств $M_1, M_2 \subseteq D^n$ положим $M = M_1 \cup (-M_2)$. Этому множеству можно сопоставить (-1,1)-матрицу, строками которой являются вектора из множества M. В дальнейшем мы будем обозначать ее той же буквой, что и соответствующее множество.

Пусть $M_1 = \{\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{m_1}\}$, $M_2 = \{\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_{m_2}\}$; $\tilde{\alpha}_i, \tilde{\beta}_j \in \bar{D}^n$.

Вектор

$$\overline{p}(M) = \frac{1}{|M|} \sum_{\tilde{\alpha}_i \in M} \tilde{\alpha}_i$$

называется информационным вектором множества M.

Вектор

$$\overline{q}(M_1, M_2) = \frac{\overline{p}(M_1) - \overline{p}(M_2)}{2}$$

называется характеризационным вектором множеств M_1 и M_2.

Будем предполагать, что множества M_1 и M_2 слабо H-разделимы, то есть существует такая разделяющая гиперплоскость G вида $(\overline{q}, \overline{x}) = 0$, что при $\tilde{\alpha} \in M_1$ и $\tilde{\beta} \in M_2$ выполнено $(\overline{q}, \tilde{\alpha}) > 0$ и $(\overline{q}, \tilde{\beta}) < 0$.

Другими словами, для всех векторов $\tilde{\alpha}$ из эталонного множества M выполняются неравенства $(\overline{q}, \tilde{\alpha}) > 0$.

Не ограничивая общности, можно считать, что все компоненты q_i вектора $\overline{q}$ положительные, $i = 1, 2, \ldots, n$. Кроме того, достаточно рассмотреть лишь случай $|M_1| = |M_2|$ (при этом вектор $\overline{q}$ коллинеарен весовому вектору $\overline{d} = \sum\limits_{\tilde{\alpha}_i \in M} \tilde{\alpha}_i$).

Величина

$$B = \min_{\tilde{\alpha} \in M} \quad (\tilde{\alpha}, \overline{d})$$

называется порогом распознавания для эталонного множества M.

Пусть $\Pi = \{1, 2, \ldots, n\}$ — исходное множество признаков.

Если $\Pi' = \{i_1, i_2, \ldots, i_k\} \subseteq \Pi$ есть некоторое (упорядоченное по возрастанию) подмножество признаков, то для $\tilde{\alpha} = \{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n\} \in D^n$ через $\tilde{\alpha}' \in D^k$ обозначим вектор, составленный из компонент $i_1, \ldots, i_k$ вектора $\tilde{\alpha}$.

Будем говорить, что вектор $\tilde{\alpha}'$ получен из $\tilde{\alpha}$ путем проектирования (удаления компонент, или удаления признаков). Аналогично, для множества векторов $M = \{\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_m\}$ через M' будем обозначать множество векторов $\{\tilde{\alpha}'_1, \tilde{\alpha}'_2, \ldots, \tilde{\alpha}'_m\}$, полученных указанным проектированием. Будем также говорить, что множество M' получено проектированием множества M на множество признаков Π'.

Задача сокращения поля признаков может быть сформулирована следующим образом.

Для данного эталонного множества M с порогом распознавания $B > 0$ и некоторого фиксированного числа $c \in \mathbb{R}$, $\quad 0 \leqslant c \leqslant B$, найти минимальное (по мощности) подмножество признаков Π' такое, что порог распознавания для M' (проекции M на Π') будет не менее, чем c.

Можно доказать следующее утверждение.

Теорема 3 ([7], [8]). *Задача сокращения поля признаков является NP-полной [10].*

3. Алгоритм распознавания с обучением $\mathcal{A}$

Здесь мы рассмотрим новый алгоритм распознавания $\mathcal{A}$. Этот алгоритм, помимо собственно процедуры распознавания, включает в себя процедуру проверки и формирования свойств слабой и сильной разделимости эталонных множеств, а также процедуру сокращения исходного множества признаков.

Опишем подробнее этапы алгоритма $\mathcal{A}$.

3.1. Подготовительный этап

Заданные эталонные множества M_1 и M_2 ($|M_1| = |M_2| = m$) представляются в виде $(-1, 1)$-матриц размера $m \times n$, из которых образуется матрица $M = M_1 \cup (-M_2)$ размера $2m \times n$ и вычисляется ее информационный вектор $\bar{q}(M) = (q_1, \dots, q_n)$ (он совпадает с характеризационным вектором для эталонных множеств M_1 и M_2).

В случае $q_i = 0$ из матрицы M удаляется (посредством массива $M_{asst}(n)$) столбец с номером i, при этом сам номер запоминается в массиве $M_{zap}(n)$.

Если для некоторого i имеем $q_i < 0$, то такой номер i также запоминается в массиве $M_{zap}(n)$, а в матрице M все знаки в столбце с номером i заменяются на противоположные [7]. Все операции также синхронно проводятся с вектором $\bar{q}(M)$, в результате чего оставшиеся его компоненты становятся положительными. Эти компоненты вектора $\bar{q}(M)$ выписываются в порядке неубывания.

Соответствующая перестановка столбцов матрицы M приводит к разбиению ее столбцов на группы, на каждой из которых значение компоненты вектора $\bar{q}(M)$ постоянно.

3.2. Проверка (формирование) слабой $\mathcal{A}$-разделимости

После умножения матрицы M на вектор $\bar{q}(M)$ определяются величины

$$SV(I) = \min_{1 \leqslant i \leqslant m} (\bar{q}(M), a_i), \qquad SN(J) = \min_{m+1 \leqslant j \leqslant 2m} (\bar{q}(M), a_j),$$

где a_i, a_j — строки матрицы M, а I, J — номера строк в M, на которых достигается соответствующий минимум. Если теперь $\min(SV(I), SN(J)) > 0$, то множества M_1 и M_2 слабо $\mathcal{A}$-разделимы; в противном случае удаляем из M строки с номерами I и J (возможно несколько раз), до тех пор, пока не придем к выполнению условия слабой $\mathcal{A}$-разделимости.

Будем называть распознавание устойчивым на матрице M, если удаление строки из M не уменьшает порога распознавания.

На этом шаге мы также можем вычислить число ошибочных строк, относительно которых распознавание является устойчивым [4].

3.3. Проверка (формирование) сильной $\mathcal{A}$-разделимости

Образуем матрицу (массив) $M_{raz}(n)$ размера $m^2 \times n$ (соответствующую множеству $\overline{M} = \{\overline{a} - \overline{b} | \overline{a} \in M_1, \overline{b} \in M_2\}$.

После умножения матрицы M на транспонированную матрицу M_{raz}^T аналогично шагу 3.2 можем вычислить величины $M_{imssv}(I)$, $M_{imssv}(J)$. Если $\min(M_{imssv}(I), M_{imssv}(J)) > 0$, то матрицы M_1 и M_2 сильно $\mathcal{A}$-разделимы [7]. В противном случае удаляем из матрицы M строки с номерами I и J (возможно, несколько раз) до тех пор, пока не придем к выполнению условия сильной $\mathcal{A}$-разделимости (запоминание номеров удаляемых из M строк происходит в массиве M_{ASKA}).

3.4. Сокращение множества признаков I (построение 1-тупиковой матрицы)

Введем следующие понятия.

Матрицу назовем тупиковой, если удаление любой группы столбцов не является допустимым, то есть приводит к снижению порога распознавания.

Матрицу назовем l-тупиковой, если удаление любой группы столбцов в количестве не более l не является допустимым, то есть приводит к снижению порога распознавания.

Замечание. Очевидно, что из l-тупиковости матрицы не следует ее тупиковость.

На этапах 3.4 и 3.5 множество признаков обрабатывается (сокращается) для полученной матрицы M.

Для матрицы M, столбцы которой разбиты на группы, пусть $q_{(i)}$ обозначает общее значение компонент вектора $\bar{q}(M)$ на столбцах i-ой группы.

Используем следующие обозначения:

D — порог распознавания для M;

$\Sigma(I)$ — скалярное произведение I-ой строки матрицы M на вектор $\bar{q}(M)$.

$SB(I)$ — часть $\Sigma(I)$ на номерах столбцов из I-ой группы;

$\Sigma_M(I) = \Sigma(I) - SB(I)$.

Подмножество номеров строк $S \subseteq \{1, 2, \dots, m\}$ будем называть граничным множеством для матрицы M, если при $i \in S$ имеет место $d \sum(\tilde{\alpha}_i) = B$, то есть $\sum(\tilde{\alpha}_i) = \frac{B}{d}$, где B — порог распознавания, $d = \sum\limits_{\tilde{\alpha} \in M} \tilde{\alpha}$.

С учетом этих обозначений строка с номером I принадлежит граничному множеству S тогда и только тогда, когда

$$D - \Sigma_M(I) \leqslant SB(I) < D - \Sigma_M(I) + q_{(i)}.$$

Пусть I_{GRAN} — массив, состоящий из номеров строк граничного множества S.

Среди столбцов i-ой группы отыскиваются такие, у которых в пересечении со строками из граничного множества S стоит -1. Пусть L — номер первого такого столбца из i-ой группы (если таких столбцов нет, то осуществляется переход к рассмотрению $(i+1)$-ой группы).

Удаляем из матрицы M столбец с номером L. Можно показать, что порог распознавания для сокращенной матрицы не меньше, чем исходный порог D.

Полагаем $M_{asst}(L) = 1$ (в массиве M_{asst} запоминаются номера удаленных столбцов) и заменяем $SB(I)$ на $SB(I) - \operatorname{sign}(M(I, L))q_{(i)}$.

Описанный процесс удаления столбцов матрицы M (без уменьшения порога распознавания) в результате циклического просмотра всех групп столбцов осуществляется до тех пор, пока еще возможно удаление какого-либо столбца хотя бы из одной группы.

В итоге будем иметь 1-тупиковую подматрицу матрицы M.

3.5. Сокращение множества признаков II (построение 2-тупиковой матрицы)

Для простоты осуществляется построение 2-тупиковой на группах столбцов матрицы исходной матрицы M, то есть возможность удаления пары столбцов проверяется только на имеющихся группах столбцов. Такая проверка осуществляется вполне аналогично описанному в пункте 3.4 с тремя изменениями:

А) $D - \Sigma_M(I) \leqslant SB(I) < D - \Sigma_M(I) + 2q_{(i)}$;

Б) Среди пар столбцов i-ой группы отыскиваются такие, у которых в пересечении со строками из граничного множества стоят либо противоположные по знаку числа, либо оба отрицательные.

В) $SB(I) = SB(I) + \sigma q_{(i)}$, где σ есть 2 либо -2, в зависимости от того, какой случай имеет место в предыдущем пункте.

Таким образом, в результате применения шага 3.5 получаем 2-тупиковую подматрицу матрицы M.

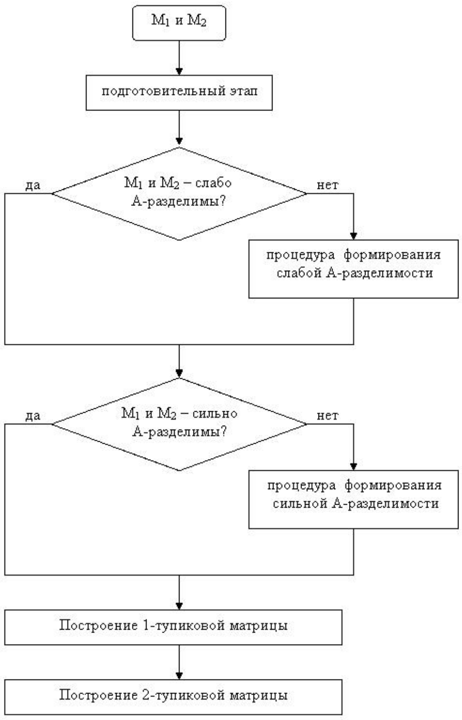
В заключение отметим следующую возможную схему обработки информации, представленной эталонным множествами M_1 и M_2.

При движении по данной блок-схеме происходит сжатие (как по строкам, так и по столбцам) исходной информации об эталонных множествах, при этом порог распознавания может только увеличиваться, что позволяет ожидать повышения качества распознавания и таким образом может рассматриваться как самоусовершенствование представленного алгоритма распознавания.

4. Практические приложения и направления дальнейших исследований

Эта работа реализована в нескольких областях в геологии и в социально-экономических областях [3], [4] и [5]. Анализ результатов распознавания показал, что процедуру распознавания самоорганизация алгоритма $\mathcal{A}$, заложенная в него, приводит к повышению качества распознавания.

Интересным обобщением данной задачи является рассмотрение обучающих множеств, в которых присутствуют ошибки. Интерес



представляет модификации предложенных алгоритмов для решения практических задач, а также теоретические оценки качества распознавания для таких моделей.

Список литературы

[1] Константинов Р.М., Королева З.Е., Кудрявцев В.Б. О комбинаторно-логическом подходе к задачам прогноза рудоносности // Проблемы кибернетики. Вып. 31. М.: Наука, 1976. С. 5–33.

[2] Журавлев Ю.И. Алгебраический подход к задачам распознавания // Проблемы кибернетики. Вып. 33. М.: Наука, 1978. С. 5–68.

[3] Переяславский В.И. Об одном линейном методе распознавания образов // Комбинаторно-алгебраические методы в прикладной математике. Горький: Изд-во ГГУ, 1982. С. 89–121.

[4] Сиротинская С.В. Метод вариационных рядов и его применение к исследованию некоторых геологических особенностей оловянно-вольфрамовых месторождений // Логико-информационные решения геологических задач. М.: Наука, 1975. С. 5–82.

[5] Кудрявцев Вит.Б., Чижова И.А. Дифференцированная оценка рекреационных территорий // Математические методы в биологии. М.: Изд-во МГУ, 1985.

[6] Шайеб А. Об одном алгоритме распознавания типа голосования // Дискретная математика и ее приложения. М.: Изд-во МГУ, 1985.

[7] Шайеб А. Болотов А.А. О линейных метрических алгоритмах распознавания // Тезисы VI всесоюзной конференции по математической кибернетике. Саратов, 1983.

[8] Шайеб А. К задаче сокращения признакового пространства в алгоритмах распознавания // Дискретная математика и ее приложения. М.: Изд-во МГУ, 1985.

[9] Шайеб А. Болотов А.А. О метрических алгоритмах классификации // ДАН. 1987. Т. 292. № 3.

[10] Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. С. 416.