Министерство образования и науки Российской Федерации

Самарский государственный аэрокосмический университет
имени академика С.П. Королева

А В Т О М А Т И Ч Е С К А Я К Л А С С И Ф И К А Ц И Я

Методические указания к лабораторной работе № 4
по курсу «МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ»

САМАРА
2005

Составители: к.т.н. В.В.Мясников

аспирант А.Ю. Баврина

УДК 681.3

Автоматическая классификация

Методические указания к лабораторной работе № 4
Самарский государственный аэрокосмический университет
имени академика С.П.Королева
Составители: А.Ю. Баврина, В.В.Мясников
Самара, 2005. 22 с.

В лабораторной работе № 4 по курсу «Методы распознавания образов» изучаются методы автоматической классификации. Рассматривается постановка алгоритма автоматической классификации, меры сходства, критерии кластеризации и два известных алгоритма: минимаксный и К внутрегрупповых средних.

Методические указания предназначены для студентов специальности 01.02.00 “Прикладная математика и информатика”, обучающихся по специализации «Математическое обеспечение обработки изображений».

Печатается по решению редакционно-издательского совета
Самарского государственного аэрокосмического университета
имени академика С.П.Королева

Рецензент: д.т.н., профессор В.А.Фурсов

Данные методические указания разработаны при поддержке Министерства образования РФ, Администрации Самарской области и Американского фонда гражданских исследований и развития (CRDF Project SA-014-02) в рамках российско-американской программы "Фундаментальные исследования и высшее образование" (BRHE).

Цель работы - изучение теоретических основ и экспериментальное исследование методов автоматической классификации для распознавания образов.

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛАБОРАТОРНОЙ РАБОТЫ

1.1. Постановка задачи автоматической классификации

Пусть классификации подлежат N объектов, каждый из которых характеризуется n-мерным вектором признаков $\bar{x}$, то есть дано множество векторов $\{\bar{x}_i\}_{i=0}^{N-1}$. Эти вектора рассматриваются как фиксированные. Каждый объект должен быть отнесен к одному из L классов: $\Omega_0, \Omega_1, \ldots \Omega_{L-1}$, где число классов L может быть известно заранее или может быть не известно. Таким образом основной вопрос задачи автоматической классификации (АК), так же как и в других задачах классификации, это вопрос об определении класса[1].

Один из возможных подходов к определению класса состоит в его понимании как *кластера (таксона)*, то есть компактной в некотором смысле области в признаковом пространстве. В этом случае задача АК является задачей *кластер-анализа* и представляет собой задачу идентификации групп схожих образов в анализируемом множестве данных (или задачу выделения кластеров).

Процесс выделения кластеров является искусством весьма "эмпирическим", так как работа конкретного алгоритма зависит не только от характера анализируемых данных, но в значительной степени и от выбранной меры подобия образов и метода, используемого для идентификации кластеров, и даже от последовательности просмотра образов.

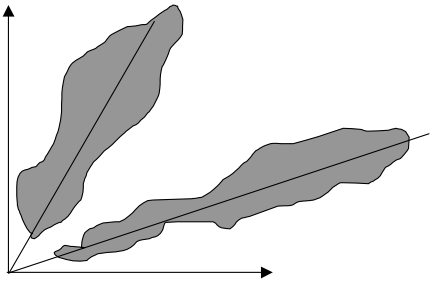
1.2. Меры сходства

Для того чтобы определить на множестве данных кластеры, необходимо *в первую очередь ввести меру сходства (подобия) образов* (векторов признаков), которая может быть положена в основу правила отнесения образов характеризуемой некоторым центром кластера. В качестве мер сходства широко используются следующие:

---

[1] Речь идет не о задании самого класса, а о задании областей признакового пространства которые "соответствуют" классам.

- *евклидово расстояния* $d(\bar{x},\bar{z}) = \|\bar{x} - \bar{z}\|$ (меньше d - больше сходство);
- *расстояние Махаланобиса* $d(\bar{x},\bar{z}) = (\bar{x} - \bar{z})^T B^{-1}(\bar{x} - \bar{z})$ (меньше d - больше сходство). Используется, когда известны статистические характеристики кластеров (матрицы разброса);
- *косинус угла между векторами* $d(\bar{x},\bar{z}) = \dfrac{\bar{x}^T \bar{z}}{\|\bar{x}\| \cdot \|\bar{z}\|}$. Данную меру удобно использовать, когда кластеры имеют тенденцию располагаться вдоль главных осей, как например изображено на рисунке:



- *мера Такимото* $d(\bar{x},\bar{z}) = \dfrac{\bar{x}^T \bar{z}}{\bar{x}^T \bar{x} + \bar{z}^T \bar{z} - \bar{x}^T \bar{z}}$.

В дальнейшем ограничимся евклидовой мерой подобия.

1.3. Критерии кластеризации

После выбора меры сходства необходимо определить *критерий кластеризации*. Критерий кластеризации может либо воспроизводить некие эвристические соображения, либо основываться на минимизации (или максимизации) какого-нибудь показателя качества.

При *эвристическом подходе* решающую роль играют интуиция и опыт. Он предусматривает задание набора правил, которые обеспечивают использование выбранной меры сходства для отнесения образов к одному из кластеров.

Подход к кластеризации, предусматривающий *использование показателя качества*, связан с разработкой процедур, которые обеспечат минимизацию или максимизацию выбранного показателя качества. Одним из наиболее популярных показателей является сумма квадратов ошибки:

$$\Im = \sum_{l=0}^{L-1} \sum_{\bar{x} \in S_l} \left\| \bar{x} - \bar{M}_j \right\|^2 \qquad (1)$$

где L — число кластеров, S_l — множество образов (векторов признаков), относящихся к l-му кластеру, а $\overline{M}_l = \frac{1}{N_l}\sum_{\bar{x}\in S_l}\bar{x}$ - вектор выборочных средних значений для множества образов S_l, величина N_l характеризует количество образов, входящих во множество S_l. Как видно, показатель качества (1) определяет общую сумму квадратов отклонений характеристик всех образов, входящих в некоторый кластер, от соответствующих средних значений по этому кластеру. Алгоритм, основанный на этом показателе качества, рассматривается в ниже. Естественно, существует масса показателей качества помимо рассмотренного. Нередко применяются алгоритмы отыскания кластеров, основанные на совместном использовании эвристического подхода и показателя качества. Подобной комбинацией является алгоритм ИСОМАД (ISODATA) [2].

Ниже рассмотрены три наиболее известные алгоритма [2], которые являются примерами как эвристического подхода, так и подхода, использующего показатель качества.

1.4. Простой алгоритм выделения кластеров

Пусть задано множество N образов $\{\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_{N-1}\}$. Пусть также центр первого кластера $\overline{M}_0$ совпадает с любым из заданных образов и определена произвольная неотрицательная пороговая величина T; для удобства можно считать, что $\overline{M}_0 = \bar{x}_0$. После этого вычисляется расстояние $d(\overline{M}_0, \bar{x}_1)$ между существующим центром кластера $\overline{M}_0$ и образом $\bar{x}_1$. Если это расстояние больше значения пороговой величины T $(d(\overline{M}_0, \bar{x}_1) > T)$, то учреждается новый кластер с центром $\overline{M}_1 = \bar{x}_1$. В противном случае образ $\bar{x}_1$ включается в кластер, центром которого является $\overline{M}_0$. Пусть условие $d(\overline{M}_0, \bar{x}_1) > T$ выполнено, и $\overline{M}_1$ - центр нового кластера.

На следующем шаге вычисляются расстояния $d(\overline{M}_0, \bar{x}_2)$ и $d(\overline{M}_1, \bar{x}_2)$ до образа $\bar{x}_2$ от центров кластеров $\overline{M}_0$ и $\overline{M}_1$. Если оба расстояния оказываются больше порога T $(d(\overline{M}_0, \bar{x}_2) > T$ и $d(\overline{M}_1, \bar{x}_2) > T)$, то учреждается новый кластер с центром $\overline{M}_2 = \bar{x}_2$. В противном случае образ $\bar{x}_2$ зачисляется в тот кластер l $(l = \overline{0,1})$, чей центр $\overline{M}_l$ к нему ближе.

Подобным же образом расстояния от каждого нового образа $\bar{x}_j \quad \left(j = \overline{1, N-1}\right)$ до каждого известного центра кластера $\overline{M}_l \left(l = 0, L-1\right)$ вычисляются и сравниваются с пороговой величиной. Если все эти расстояния превосходят значение порога T $\left(\forall\, l = \overline{0, L-1} \quad d\left(\overline{M}_l, \bar{x}_j\right) > T\right)$, то учреждается новый кластер с центром $\overline{M}_L = \bar{x}_j$ (и число кластеров увеличивается на единицу). В противном случае образ зачисляется в кластер с самым близким к нему центром.

Результаты описанной процедуры определяются выбором первого центра кластера, порядком просмотра образов, значением пороговой величины T и, конечно, геометрическими характеристиками данных.

1.5. Алгоритм максиминного расстояния

Алгоритм, основанный на принципе *максиминного* (максимально-минимального) *расстояния*, представляет собой еще одну простую эвристическую процедуру, использующую евклидово расстояние. Этот алгоритм в принципе аналогичен схеме из п.1.4, за исключением того обстоятельства, что в первую очередь он выявляет наиболее удаленные кластеры.

Алгоритм состоит из нескольких шагов.

Шаг 1. Произвольным образом выбирается цент первого кластера $\overline{M}_0$. Удобно выбирать в качестве центра кластера $\overline{M}_0$ тот вектор признаков $\bar{x} \in \left\{\bar{x}_j\right\}_{j=0}^{N-1}$, который обладает некоторыми «экстремальными» свойствами. Например, удобным является тот вектор, удаление которого от среднего всех векторов из выборки максимально:

$$\overline{M}_0 = arg \max_{\bar{x} \in \left\{\bar{x}_j\right\}_{j=0}^{N-1}} d\left(\bar{x}, \frac{1}{N}\sum_{i=0}^{N-1} \bar{x}_i\right).$$

Шаг 2. Выбирается цент второго кластера $\overline{M}_1$. В качестве центра используется тот вектор $\bar{x} \in \left\{\bar{x}_j\right\}_{j=0}^{N-1}$, который наиболее удален от первого центра кластера:

$$\overline{M}_1 = arg \max_{\bar{x} \in \left\{\bar{x}_j\right\}_{j=0}^{N-1}} d\left(\overline{M}_0, \bar{x}\right).$$

Шаг L $(L>2)$. Выбирается цент кластера $\overline{M}_{L-1}$. Для этого вычисляются все расстояния между оставшимися образами (векторами признаков) $\bar{x}\in\{\bar{x}_j\}_{j=0}^{N-1}\setminus\{\overline{M}_l\}_{l=0}^{L-2}$ и уже существующими центрами кластеров $\overline{M}_0,\overline{M}_1,\dots\overline{M}_{L-2}$. Для каждого оставшегося образа $\bar{x}$ находится тот центр кластера l, расстояние до которого минимально: $l=arg\ \underset{l=0,L-2}{min}\, d(\overline{M}_l,\bar{x})$ (вектора распределяются по кластерам по критерию близости к их центру). В качестве претендента на новый центр кластера $\overline{M}_{L-1}$ берется тот вектор признаков $\bar{x}$, у которого это минимальное расстояние (расстояние до центра «своего» кластера) максимально:

$$\widetilde{\overline{M}}_{L-1}=arg\ \underset{\bar{x}\in\{\bar{x}_j\}_{j=0}^{N-1}}{max}\ \underset{l=0,L-2}{min}\, d(\overline{M}_l,\bar{x}).$$

Полученное для выбранного вектора расстояние $d_{min}=\underset{l=0,L-2}{min}\, d(\overline{M}_l,\widetilde{\overline{M}}_{L-1})$ от него до ближайшего кластера сравнивается с некоторым «типичным» расстоянием $d_{typical}$ между кластерами. Если полученное расстояние больше «типичного», то этот вектор становится новым центром кластера:

$$d_{min}>d_{typical}\Rightarrow\overline{M}_{L-1}=\widetilde{\overline{M}}_{L-1}.$$

В противном случае – процесс выделения кластеров останавливается.

Выбор типичного расстояния может быть осуществлен различными способами. Один из наиболее типичных способов, это задание типичного расстояния равного некоторой части η от среднего расстояния между уже существующими кластерами:

$$d_{typical}=\eta\left(\frac{2}{(L-1)(L-2)}\right)\sum_{l=0}^{L-2}\sum_{j=l+1}^{L-2}d(\overline{M}_l,\overline{M}_j).$$

Величина η выбирается из условия $0<\eta<1$. Типичное значение: $\eta=\frac{1}{2}$.

1.6. Алгоритм *K* внутригрупповых средних

Алгоритмы, рассмотренные в п.1.4 и п.1.5 являются, в сущности, эвристическими процедурами. Алгоритм, представленный ниже, минимизирует показатель качества, заданный как сумма квадратов расстояний всех точек, входящих

в кластерную область, до центра кластера. Эта процедура, которую часто называют алгоритмом, основанным на вычислении K внутригрупповых средних, состоит из следующих шагов.

Шаг 1. Выбираются K исходных центров кластеров на первой итерации $(r=1)$: $\overline{M}_0(r), \overline{M}_1(r), \ldots \overline{M}_{K-1}(r)$. Этот выбор производится произвольно, и обычно в качестве исходных центров используются первые K образов из заданного множества $\{\overline{x}_j\}_{j=0}^{N-1}$ $(N > K)$:

$$\overline{M}_0(1) = \overline{x}_0, \quad \overline{M}_1(1) = \overline{x}_1, \quad \ldots \overline{M}_{K-1}(1) = \overline{x}_{K-1} .$$

Номер итерации увеличивается: := +1.

Шаг 2. На r-ой итерации $(r = 2,3...)$ исходное множество образов $\{\overline{x}_j\}_{j=0}^{N-1}$ распределяется по K кластерам по правилу близости. То есть некоторый образ $\overline{x} \in \{\overline{x}_j\}_{j=0}^{N-1}$ относят в кластер $S_k(r)$ $(k = \overline{0, K-1})$ с центром $\overline{M}_k(r-1)$, рассчитанным на предыдущем шаге, если этот центр - ближайший:

$$\overline{x} \in S_k(r), \quad k = arg \min_{j=\overline{0,K-1}} d\left(\overline{x}, \overline{M}_j(r-1)\right) .$$

Таким образом $S_k(r)$ - множество образов, входящих в кластер с номером k на r-ой итерации алгоритма. В случае равенства расстояний от некоторого образа до нескольких центров кластеров решение об отнесении этого образа к одному из них принимается произвольным образом.

Шаг 3. На основе результатов шага 2 определяются новые центры кластеров $\overline{M}_0(r), \overline{M}_1(r), \ldots \overline{M}_{K-1}(r)$ на r-ой итерации алгоритма. Они определяются из условия, чтобы сумма квадратов расстояний между всеми образами, принадлежащими кластеру $S_k(r)$, и новым центром кластера $\overline{M}_k(r)$ должна быть минимальной. Другими словами, новые центры кластеров выбираются таким образом, чтобы минимизировать частные показатели

$$J_k(r) = \sum_{\overline{x} \in S_k(r)} d\left(\overline{x}, \overline{M}_k(r)\right)^2 , \; k = 0,1, \ldots, K-1$$

и, следовательно, интегральный показатель качества кластеризации на r-ом шаге

$$J(r)=\sum_{k=0}^{K-1}J_k(r)=\sum_{k=0}^{K-1}\sum_{\bar{x}\in S_k(r)}d\left(\bar{x},\overline{M}_k(r)\right)^2 .$$

Для случая евклидового расстояния новый центр кластера $\overline{M}_k(r)$, обеспечивающий минимизацию соответствующего частного показателя $J_k(r)$, является, в сущности, выборочным средним, определенным по множеству образов в кластере $S_k(r)$:

$$\overline{M}_k(r)=\frac{1}{N_k}\sum_{x\in S_k(r)}\bar{x}\ ,\ k=0,1,\ldots,K-1\ ,$$

где $N_k=|S_k(r)|$ - число образов, входящих в кластер $S_k(r)$ на *r*-ом шаге. Как видно, что название алгоритма «*K* внутригрупповых средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

Шаг 4. Равенство центров кластеров на соседних шагах $\overline{M}_k(r)=\overline{M}_k(r-1)\ \ \left(k=\overline{0,K-1}\right)$ является условием сходимости алгоритма, и при его достижении выполнение алгоритма заканчивается. В противном случае алгоритм повторяется с шага 2 с новым номером итерации $:=\ +1$.

Качество работы алгоритма *K* внутригрупповых средних зависит от числа выбираемых центров кластеров *K*, от выбора центров кластеров на первой итерации $\overline{M}_0(1),\overline{M}_1(1),\ldots\overline{M}_{K-1}(1)$ и, естественно, от геометрических особенностей данных. От последовательности просмотра данных результаты не зависят.

Хотя для этого алгоритма общее доказательство сходимости не известно, получения приемлемых результатов можно ожидать в тех случаях, когда данные образуют характерные гроздья, отстоящие друг от друга достаточно далеко. В большинстве случаев практическое применение этого алгоритма потребует проведения экспериментов, связанных с выбором различных значений параметра *K* и расположения первоначальных центров кластеров.

Пример. В качестве простой числительной иллюстрации алгоритма *K* внутригрупповых средних рассмотрим образы, представленные на рисунке 1. Процедура кластеризации протекает следующим образом:

Шаг 1. Задается $K = 2,\ \ = 1$ и выбирается $\overline{M}_0(1) = \bar{x}_0 = (0, 0)'$, $\overline{M}_1(1) = \bar{x}_1 = (1, 0)'$. Увеличивается номер итерации: $= 2$.

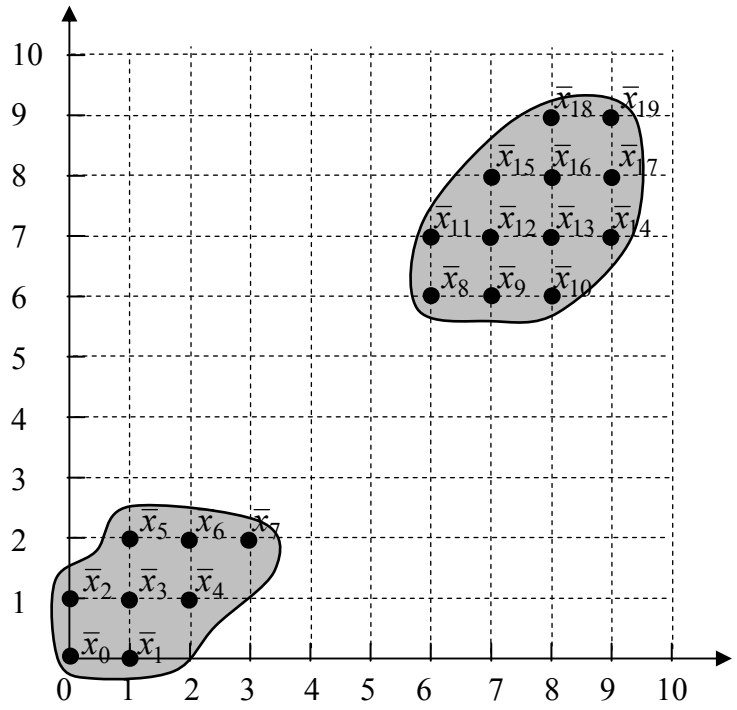


Рисунок 1. *Выборка образов, иллюстрирующая работу алгоритма К внутригрупповых средних*

Шаг 2. Так как $\|\bar{x}_0 - \overline{M}_0(1)\| < \|\bar{x}_0 - \overline{M}_1(1)\|$ и $\|\bar{x}_2 - \overline{M}_0(1)\| < \|\bar{x}_2 - \overline{M}_1(1)\|$, то $S_0(2) = \{\bar{x}_0, \bar{x}_2\}$. Аналогично устанавливается, что остальные образы расположены ближе к центру кластера $\overline{M}_1(1)$, и поэтому $S_1(1) = \{x_1, x_3, x_4, \ldots, x_{19}\}$.

Шаг 3. Коррекция центров кластеров:

$$\overline{M}_0(2) = \frac{1}{N_0} \sum_{\bar{x} \in S_0(2)} \bar{x} = \frac{1}{2}(\bar{x}_0 + \bar{x}_2) = \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix},$$

$$\overline{M}_1(2) = \frac{1}{N_1} \sum_{\bar{x} \in S_1(2)} \bar{x} = \frac{1}{18}(\bar{x}_1 + \bar{x}_3 + \ldots + \bar{x}_{19}) = \begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}.$$

Шаг 4. Так как $\overline{M}_k(2) \neq \overline{M}_k(1)\ \ (k = 1, 2)$, то

- увеличивается номер итерации $= 3$ и
- производится возврат к шагу 2.

Шаг 2. Выбор новых центров кластеров приводит к неравенствам:

- $\left\|\bar{x}_j - \overline{M}_0(2)\right\| < \left\|\bar{x}_j - \overline{M}_1(2)\right\|$ для $j = 0, 1, \ldots, 7$ и
- $\left\|\bar{x}_j - \overline{M}_0(2)\right\| > \left\|\bar{x}_j - \overline{M}_1(2)\right\|$ для $j = 8, 9, \ldots, 19$.

Следовательно, $S_0(3) = \{\bar{x}_0, \bar{x}_1, ..., \bar{x}_7\}$ и $S_1(3) = \{\bar{x}_8, \bar{x}_9, ..., \bar{x}_{19}\}$.

Шаг 3. Коррекция центров кластеров:

$$\overline{M}_0(3) = \frac{1}{N_0} \sum_{\bar{x} \in S_0(3)} \bar{x} = \frac{1}{8}\left(\bar{x}_0 + \bar{x}_1 + ... + \bar{x}_1\right) = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix},$$

$$\overline{M}_1(3) = \frac{1}{N_1} \sum_{\bar{x} \in S_1(3)} \bar{x} = \frac{1}{12}\left(\bar{x}_9 + \bar{x}_{10} + ... + \bar{x}_{19}\right) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}.$$

Шаг 4. Так как $\overline{M}_k(3) \neq \overline{M}_k(2)$ $(k = 1, 2)$, то , то

- увеличивается номер итерации $= 4$ и
- производится возврат к шагу 2.

Шаг 2. Получаем те же результаты, что и на предыдущей итерации, то есть $S_0(4) = S_0(3) = \{\bar{x}_0, \bar{x}_1, ..., \bar{x}_7\}$ и $S_1(4) = S_1(3) = \{\bar{x}_8, \bar{x}_9, ..., \bar{x}_{19}\}$.

Шаг 3. Также получаем идентичные результаты.

Шаг 4. Так как $\overline{M}_k(4) = \overline{M}_k(3)$ $(k = 1, 2)$, то алгоритм сошелся и в результате получены следующие центры кластеров

$$\overline{M}_0(3) = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}, \quad \overline{M}_1(3) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}.$$

Эти результаты согласуются с человеческим интуитивным представлением о группировке искомых образов в пространстве признаков.

2. РЕАЛИЗАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В СРЕДЕ MATHCAD

2.1. Генерация исходных данных

Для генерации данных, которые подлежат кластеризации, используйте алгоритм генерации случайного двумерного вектора, описанный в методических указаниях для первой лабораторной работы [4]. Ниже представлен код MathCAD программы для генерации 200 значений-реализаций случайного вектора признаков.

Текст программы в MathCad	**Комментарии**
$M := \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad B := \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$	Задание параметров нормального закона распределения
$A_{0,0} := \sqrt{R_{0,0}} \qquad A_{0,1} := 0$ $A_{1,0} := \frac{R_{0,1}}{\sqrt{R_{0,0}}} \qquad A_{1,1} := \sqrt{R_{1,1} - \frac{(R_{0,1})^2}{R_{0,0}}}$	Определение параметров линейного преобразования
$A = \begin{bmatrix} 2.236 & 0 \\ 0.894 & 0.447 \end{bmatrix}$	Отображение полученного результата для матрицы линейного преобразования
$n := 2 \qquad l := 0..n-1 \qquad k := 0..n-1$ $N := 200 \qquad i := 0..N-1$ $j := 0..11$	Вспомогательные переменные, отвечающие за двукомпонентность вектора (n, l, k), число выборочных значений (N, i) и за процесс генерации стандартной нормально распределенной случайной величины (j).
$y_{l,i} := \sum_{j} (rnd(1) - 0.5)$	Генерация N реализаций случайного вектора, компоненты которого – суть независимые и нормально распределенные $N(0,1)$ случайные величины.
$X_{k,i} := \sum_{l} A_{k,l} \cdot y_{l,i} + M_k$	Генерация N реализаций случайного вектора с требуемым нормальным законом распределения $N(\overline{M}, B)$.

2.2. Основные функции кластеризации

Большинство алгоритмов кластеризации состоит из трех блоков: блока задания начальных центров кластеров, блока отнесения вектора к некоторому кластеру/классу и блока пересчета центров кластеров по данным векторам признаков и их номерам классов. Для этой цели удобно использовать следующий код в системе MathCad.

$L := 2 \qquad l := 0..L-1 \qquad i := 0..N-1$

$M_{0,l} := x_{0,l} \quad M_{1,l} := x_{1,l}$

Инициирование счетчиков и первоначальных центров кластеров

$d_{i,l} := \sqrt{\left(M_{0,l} - x_{0,i}\right)^2 + \left(M_{1,l} - x_{1,i}\right)^2}$

$ClassNum_i := min\_ind(GetVector(d, i, L), L)$

Получение массива, содержащего номера классов для каждого вектора признаков

$k_l := 0 \qquad k_{(ClassNum_i)} := k_{(ClassNum_i)} + 1$

$M_{0,l} := 0 \qquad M_{0,(ClassNum_i)} := M_{0,(ClassNum_i)} + x_{0,i}$

$M_{1,l} := 0 \qquad M_{1,(ClassNum_i)} := M_{1,(ClassNum_i)} + x_{1,i}$

$M_{0,l} := \frac{M_{0,l}}{k_l} \qquad M_{1,l} := \frac{M_{1,l}}{k_l}$

Пересчет центров кластеров

Описание дополнительных функций и способ визуализации результатов кластеризации приведены в п.2.4.

2.3. Вспомогательные функции, используемые для кластеризации

Среда математического программирования MathCAD предназначена для математического программирования. Поэтому реализацию подпрограммы в ней удобно выполнить в виде функции, объявление которой должно предшествовать ее первому вызову. Для кластеризации удобно использовать следующие функции.

2.3.1. Функция поиска и отбора

```
AnyIndexOfValue(array, value, Len) :=
    ind ← 0
    i ← 1
    while i < Len
        ind ← i  if array_i = value
        i ← i + 1
    return ind
```

Функция нахождения некоторого (последнего) индекса элемента *value* в массиве *array* длины *Len*.

```
min_ind(x, Len) :=
    ind ← 0
    min ← x_0
    i ← 1
    while i < Len
        if x_i < min
            min ← x_i
            ind ← i
        i ← i + 1
    return ind
```

Функция выдачи индекса минимального элемента в массиве *x* размером *Len*.

```
Max_Ind(x, Len) :=
    ind ← 0
    max ← x_0
    i ← 1
    while i < Len
        if x_i > max
            max ← x_i
            ind ← i
        i ← i + 1
    return ind
```

Функция выдачи индекса максимального элемента в массиве *x* размером *Len*.

```
GetVector(x, i, Len) :=
    j ← 0
    while j < Len
        y_j ← x_{i,j}
        j ← j + 1
    return y
```

Функция выдачи вектора, сформированного из элементов *i*-ой строки матрицы *x* размером *Len* элементов по горизонтали

2.3.2. Метод отображения результатов кластеризации

Результатом кластеризации является массив *ClassNum*, элементы которого содержат номер класса/кластера, к которому отнесен соответствующий вектор из набора векторов *x*. Для отображения векторов, отнесенных к различным классам, следует распределить весь набор векторов на два набора с векторами, принадлежащими только какому-либо одному классу. Для формирования списка векторов, отнесенных к некоторому классу, например к классу «0» следует использовать следующий код MathCad:

$$ind1 := AnyIndexOfValue(ClassNum, 1, N)$$

$$x1_{s,i} := if\left(ClassNum_i = 1, x_{s,i}, x_{s,ind1}\right)$$

При таком формировании списка вектора, отнесенные к необходимому классу, переносятся в новый вектор, а на место остальных векторов записывается некоторый наперед заданный вектор нужного класса.

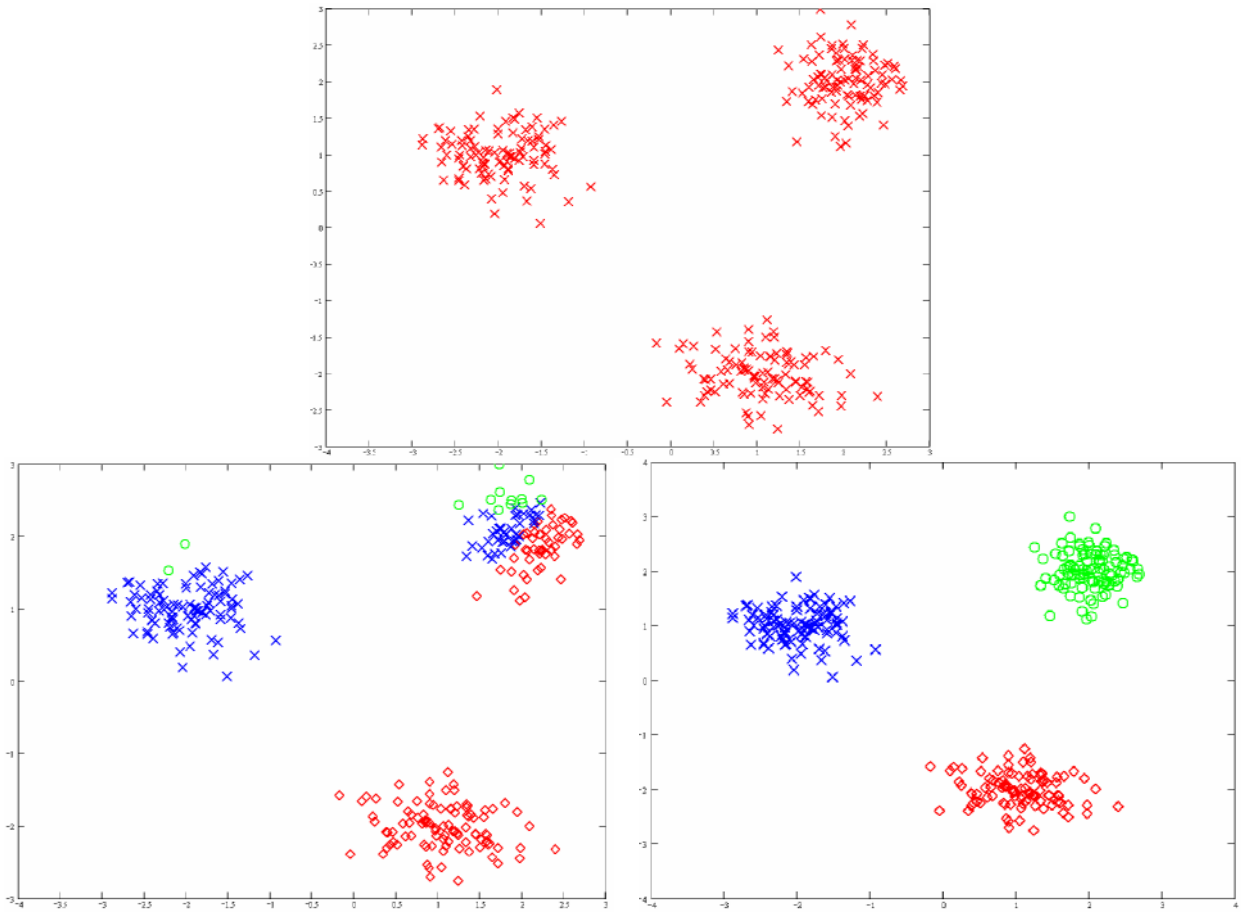


Рисунок 2. *Пример кластеризации с использованием алгоритма К внутригрупповых средних (2 шага алгоритма)*

3. ЛИТЕРАТУРА

1. Дуда Р., Харт П. Распознавание образов и анализ сцен: Пер. с англ. - М.: Мир, 1976. - 512 с.
2. Ту Дж., Гонсалес Р. Принципы распознавания образов: Пер. с англ. - М.: Мир, 1978. - 412с.
3. Фукунага К. Введение в статистическую теорию распознавания образов: Пер. с англ. - М.: Наука, 1979. - 368с.
4. Коломиец Э.И., Мясников В.В. Моделирование экспериментальных данных для решения задач распознавания образов: Методические указания к лабораторной работе № 1 по курсу «Методы распознавания образов».

4. ПОРЯДОК ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ РАБОТЫ

4.1. Исходные данные

- математические ожидания для пяти случайных векторов признаков, задаются по номеру варианта учащегося в п.5 настоящего руководства;
- исполняемые в системе MathCad файлы, необходимые при выполнении лабораторной работы (предоставляются преподавателем).

4.2. Общий план выполнения работы

1. Смоделировать и изобразить графически обучающие выборки объема N=50 для пяти нормально распределенных двумерных случайных векторов с заданными математическими ожиданиями и самостоятельно подобранными корреляционными матрицами, которые обеспечивают линейную разделимость классов.
2. Объединить пять выборок в одну. Общее количество векторов в объединенной выборке должно быть 250. Полученная объединенная выборка используется для выполнения пунктов 3 и 4 настоящего плана.
3. Разработать программу кластеризации данных с использованием минимаксного алгоритма. В качестве типичного расстояния взять половину среднего расстояния между существующими кластерами. Построить отображение результатов кластеризации для числа кластеров, начиная с двух. Построить график зависимости максимального (из минимальных) и типичного расстояний от числа кластеров.
4. Разработать программу кластеризации данных с использованием алгоритма K внутригрупповых средних для числа кластеров равного 3 и 5. Для ситуации 5 кластеров подобрать начальные условия так, чтобы получить два результата: а) чтобы кластеризация максимально **соответствовала** первоначальному разбиению на классы («правильная» кластеризация); б) чтобы кластеризация максимально **не соответствовала** первоначальному разбиению на классы («неправильная» кластеризация). Для всех случаев построить графики зависимости числа векторов признаков, сменивших номер кластера, от номера итерации алгоритма.

4.3. Содержание отчета

Отчет по работе должен содержать:

1. Исходные данные генерируемых векторов признаков – средние и ковариационные матрицы.
2. Для минимаксного алгоритма кластеризации включить в отчет: число классов, полученное в результате работы алгоритма; график зависимости максимального (из минимальных) и типичного расстояний от числа кластеров.
3. Для алгоритма K внутригрупповых средних: нарисовать график зависимости числа векторов признаков, сменивших номер кластера, от номера итерации алгоритма (для 3 и 5 кластеров, в последнем случае - для ситуаций «правильной» и «неправильной» кластеризации).

5. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Меры сходства, используемые при автоматической классификации.
2. Критерии, используемые при автоматической классификации.
3. Простой алгоритм выделения кластеров. Достоинства и недостатки.
4. Минимаксный алгоритм. Достоинства и недостатки.
5. Алгоритм К внутригрупповых средних. Достоинства и недостатки.
6. Алгоритм ИЗОМАД.

6. ВАРИАНТЫ ЗАДАНИЙ

Вариант	Математические ожидания пяти наборов нормально распределенных случайных векторов
1.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$
2.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$
3.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$
4.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$
5.	$\overline{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}.$
6.	$\overline{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}.$
7.	$\overline{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
8.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}.$
9.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
10.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$
11.	$\overline{M}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$
12.	$\overline{M}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$
13.	$\overline{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
14.	$\overline{M}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$
15.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$
16.	$\overline{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$
17.	$\overline{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$

18. $\overline{M}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$

19. $\overline{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$

20. $\overline{M}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$

21. $\overline{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$

22. $\overline{M}_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$

23. $\overline{M}_1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$

24. $\overline{M}_1 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$

25. $\overline{M}_1 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \quad \overline{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \overline{M}_3 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad \overline{M}_4 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \overline{M}_5 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$

СОДЕРЖАНИЕ

Учебное издание

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ

Методические указания к лабораторной работе № 4

по курсу «Методы распознавания образов»

Составители: Мясников Владислав Валерьевич

Баврина Алина Юрьевна

Самарский государственный аэрокосмический университет

имени академика С.П.Королева

443086, Самара, Московское шоссе, 34

Отпечатано на кафедре геоинформатики СГАУ

Тираж 20 экз.