Лабораторная работа № 1

КЛАСТЕРНЫЙ АНАЛИЗ В РАСПОЗНАВАНИИ ОБРАЗОВ

Цель работы: Ознакомиться с методами кластеризации образов и расстояниями в пространстве признаков, разработать программу, выполняющую кластеризацию заданного множества образов.

Теоретические пояснения

Любой физический объект или процесс обладает набором некоторых параметров (характеристик, свойств), которые, собственно, и позволяют отличать один объект от другого. Объекты, имеющие похожие параметры, можно объединить в группу (класс). Отнесение некоторого объекта к одной из известных групп называется распознаванием или классификацией.

Измеряемые или вычисляемые свойства объектов, позволяющие отличить классы друг от друга, называются признаками. Совокупность конкретных значений признаков, относящихся к одному объекту, называется образом объекта. Тогда класс можно определить как множество образов, обладающих рядом близких значений признаков. Эти образы называются элементами класса.

В пределе каждый класс может состоять только из одного элемента, как, например, при распознавании человека. В таком случае принято говорить об опознавании или идентификации образа. С другой стороны, все множество образов может быть разделено всего на два класса, например «свой», «чужой».

При разработке системы распознавания некоторого множества образов, разделенного на классы, используется одно множество признаков, отличающихся значениями для разных классов. Если число признаков, используемых при классификации, равно n, то образ можно представить в виде некоторого упорядоченного набора значений признаков или вектора признаков вида $\mathbf{x} = (x_1, ..., x_n)$, где x_i – значение i-го признака данного образа. В многомерном пространстве, осями координат которого являются отдельные признаки, каждый образ определяется точкой, причем расстояние от этой точки до начала координат и расстояние между

точками определяется выбранной метрикой пространства.

Каждый класс занимает некоторую область в n-мерном пространстве признаков. Форма этой области определяется степенью отличия элементов класса и может задаваться границами, например, плоскостями, или характеристиками признаков класса, например, векторами средних значений и среднеквадратичных отклонений признаков.

Множество образов, используемое при разработке системы распознавания, называется обучаемым. Элементы этого множества относятся к разным классам, причем иногда заранее неизвестно к какому классу относится каждый образ. Порой неизвестно и число классов, на которые можно разделить множество имеющихся образов. Процедуру разбиения множества образов на классы называют кластеризацией, а совокупность отнесенных в результате кластеризации к отдельному классу образов – кластером. Результаты кластеризации одного и того же множества образов зависят от выбранной метрики пространства признаков и применяемого метода кластеризации.

Для определения расстояния между точками в пространстве признаков необходимо выбрать метрику, т.е. определить процедуру измерения расстояния d_{lp} между точками l и p в этом пространстве так, чтобы выполнялись следующие аксиомы:

- симметричность расстояния ($d_{lp} = d_{pl}$);
- правило треугольника ($d_{lh} + d_{hp} > d_{lp}$);
- положительность расстояния ($d_{lp} >= 0$, причем $d_{lp} = 0$ только если $l = p$)

Наиболее часто в пространстве признаков используется Евклидово расстояние: $d_{lp} = \sqrt{\sum_{i=1}^{n}(x_{il} - x_{ip})^2}$, (1)

где x_{il}, x_{ip} – i-ые координаты точек l и p соответственно. Так как при распознавании важно не абсолютное, а относительное расстояние между точками в пространстве признаков, то квадратный корень в (1), как правило, не вычисляется. В векторной форме Евклидово расстояние между двумя образами в пространстве признаков можно представить как: $d_{lp} = (\mathbf{x}_l - \mathbf{x}_p) \times (\mathbf{x}_l - \mathbf{x}_p)^{\mathrm{T}}$, (2)

где × – операция умножения векторов, $\mathbf{x}_l$, $\mathbf{x}_p$ – вектора-строки признаков для образов l и p соответственно.

Евклидово расстояние является частным случаем расстояния Минковского, которое вычисляется как: $d_{lp} = \sum_{i=1}^{n} (x_{il} - x_{ip})^{\lambda}$, (3)

где λ – целое положительное число.

Манхеттенское расстояние представляет собой сумму абсолютных значений разности двух образов по каждому признаку, т.е. $d_{lp} = \sum_{i=1}^{n} \left| x_{il} - x_{ip} \right|$. (4)

За расстояние доминирования между двумя образами принимается максимальная разность значений признаков, т.е. $d_{lp} = \max_{i=1,n} \left(\left| x_{il} - x_{ip} \right| \right)$. (5)

Если диапазоны возможных значений признаков существенно отличаются, например, число углов замкнутого контура и его длина в миллиметрах, то признаки нормируются. $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, где x' – нормированное значение признака x, $x_{\max}$, $x_{\min}$ – возможные максимальное и минимальное значения этого признака. В результате все признаки получают один диапазон изменения – от нуля до единицы.

Учесть различие диапазонов изменений признаков позволяет и расстояние Камберра: $d_{lp} = \sum_{i=1}^{n} \frac{\left| x_{il} - x_{ip} \right|}{\left| x_{il} + x_{ip} \right|}$, (6)

которое не требует предварительного нормирования признаков. Из (6) следует, что чем меньше абсолютное значение признака, тем большее влияние он оказывает на результат вычисления расстояния.

Признаки могут иметь разную ценность при распознавании, например, отличие в числе вершин на единицу более существенно, чем такое же по величине отличие в длине контура. Чтобы учесть этот момент применяют весовые коэффициенты признаков, что приводит к изменению формул (1) – (6), например, формула для вычисления Евклидова расстояния будет иметь вид $d_{lp} = \sqrt{\sum_{i=1}^{n} \eta_i (x_{il} - x_{ip})^2}$,

где η_i – весовой коэффициент i-го признака.

Косинусное расстояние представляет собой угол α_{lp} между векторами $\mathbf{x}_l$ и $\mathbf{x}_p$. В векторной форме оно вычисляется как:

$$\alpha_{lp} = \arccos\left(\frac{\mathbf{x}_l \times \mathbf{x}_p^T}{(\mathbf{x}_l \times \mathbf{x}_l^T)^{1/2} (\mathbf{x}_p \times \mathbf{x}_p^T)^{1/2}} \right), \quad (7)$$

и дает хорошие результаты при распознавании классов, образы которых вытянуты вдоль радиус-вектора в пространстве признаков. Так как расстояние (7) основано на скалярном произведении векторов $\mathbf{x}_l \times \mathbf{x}_p^T = (\mathbf{x}_l \times \mathbf{x}_l^T)^{1/2} (\mathbf{x}_p \times \mathbf{x}_p^T)^{1/2} \cos(\alpha)$, то в качестве альтернативы косинусному расстоянию можно использовать обратное значение скалярного произведения векторов.

При разработке системы распознавания кроме расстояния между точками в пространстве признаков необходимо задать и способ определения расстояния между точкой и множеством точек, например, классом. Как правило, при задании этого способа используются статистические характеристики признаков класса: вектор средних значений **m** и ковариационная матрица **Cov**.

Эти статистические характеристики признаков класса вычисляются следующим образом. Пусть множество n-мерных векторов, описывающих в пространстве признаков образы, относящиеся к одному классу, задано в виде матрицы: $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & \dots & x_{nm} \end{pmatrix}$, где m – число образов, составляющих данный класс.

$$\text{Тогда } \mathbf{m} = (\mu_1, \mu_2, \dots, \mu_n) : \mu_i = \frac{1}{m} \sum_{k=1}^{m} x_{ik}; \quad (8)$$

$$\mathbf{Cov} = \begin{pmatrix} D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \dots & \dots & \dots & \dots \\ D_{n1} & D_{n2} & \dots & D_{nn} \end{pmatrix} : \begin{cases} D_{ii} = \frac{1}{m} \sum_{k=1}^{m} (x_{ik} - \mu_i)^2; \\ D_{ij} = \frac{1}{m} \sum_{k=1}^{m} (x_{ik} - \mu_i)(x_{jk} - \mu_j), \end{cases} \quad (9)$$

где x_{ik} – значение i-го признака k-го образа (k=1,…,m), μ_i – среднее значение (оценка математического ожидания) i-ой компоненты вектора признаков, D_{ii} – дисперсия i-го признака, D_{ij} – коэффициент ковариации i-го и j-го признаков.

Из (9) видно, что ковариационная матрица симметрична относительно главной диагонали, т.е. необходимо вычислять только половину ее элементов. Ковариация характеризует степень линейной зависимости случайных величин. Если ковариация равна нулю, то величины считаются некоррелированными.

Необходимо отметить важный факт, касающийся ковариационной матрицы. Если число образов, относящихся к некоторому классу, меньше или равно числу признаков, то ковариационная матрица, вычисляемая по этому множеству образов, будет вырожденной при любых значениях признаков каждого образа. Справедливость данного утверждения легко показать, если представить ковариационную матрицу как результат матричного умножения $\mathbf{Cov} = \mathbf{X}_c \times \mathbf{X}_c^T$, где $\mathbf{X}_c$ – центрированная, т.е. смещенная на вектор **m,** матрица **X.** Как известно из линейной алгебры определитель матрицы, полученной таким образом, равен нулю, если $m \leq n$, следовательно, обратная ковариационная матрица существует только при выполнении условия $m>n$.

Наиболее простой способ определения расстояния между точкой и классом в пространстве признаков состоит в использовании формул (1) – (7) с заменой вектора признаков одной из точек на вектор средних значений признаков класса. Такой подход не учитывает степень компактности класса, которая определяется дисперсиями его признаков. Следующие два расстояния позволяют устранить этот недостаток.

Евклидово расстояние с учетом дисперсий между точкой и классом в пространстве признаков: $d = (\mathbf{x} - \mathbf{m})\, \mathbf{D}^{-1}\, (\mathbf{x} - \mathbf{m})^T$, (10)

где, **x** – вектор признаков точки, **m** – вектор средних значений признаков класса, **D** – диагональная матрица (диагональные элементы – дисперсии признаков).

Более полно учесть статистические особенности признаков позволяет расстояние Махаланобиса: $d = (\mathbf{x}-\mathbf{m})\, \mathbf{Cov}^{-1}\, (\mathbf{x}-\mathbf{m})^T$, (11)

где $\mathbf{Cov}^{-1}$ – обратная ковариационная матрица признаков.

Расстояния (10) и (11) дают хорошие результаты для классов, имеющих эллипсоидную форму в пространстве признаков. Для классов, имеющих более сложную форму, в качестве расстояния между точкой и классом часто использу-

ется расстояние от точки до ближайшего представителя класса, что существенно увеличивает время вычислений, особенно при большом количестве точек, принадлежащих классу.

Третьим видом расстояний, которое иногда приходится задавать при разработке системы распознавания, является расстояние между классами. Наиболее простой способ – использование расстояний (1) – (7) с заменой векторов признаков точек на вектора средних значений. Такой способ дает хорошие результаты для компактных классов, имеющих эллипсоидную форму. Для классов сложной формы можно использовать приведенные ниже расстояния.

Расстояние ближнего соседа – расстояние между ближайшими точками, принадлежащими разным классам w_1 и w_2, т.е.

$$d(w_1,w_2) = \min(d_{lp}),\ (l = 1,m_1;\ p = 1,m_2), \qquad (12)$$

где m_1, m_2 – число точек классов w_1 и w_2 соответственно.

Расстояние дальнего соседа – расстояние между самыми дальними точками, принадлежащими разным классам w_1 и w_2, т.е.

$$d(w_1,w_2) = \max(d_{lp}),\ (l = 1,m_1;\ p = 1,m_2), \qquad (13)$$

К-расстояние или расстояние по Колмогорову между классами определяется как:

$$d(w_1,w_2) = \left(\frac{1}{m_1 m_2} \sum_{l=1}^{m_1} \sum_{p=1}^{m_2} d_{lp}^{\lambda} \right)^{1/\lambda}, \qquad (14)$$

где λ – целое число в диапазоне $\pm \infty$. При $\lambda \to +\infty$ К-расстояние вырождается в расстояние дальнего соседа, а при $\lambda \to -\infty$ – в расстояние ближнего соседа.

Отметим, что в (12) – (14) способ определения расстояние между точками d_{lp} выбирается разработчиком системы распознавания, например, из (1) – (7).

Основной целью кластеризации является разбиение множества образов на подмножества близких между собой образов. Определение степени близости зависит от метода кластеризации и способов вычисления различных видов расстояний. Ниже рассмотрено несколько наиболее простых методов кластеризации.

Пороговый алгоритм кластеризации. Пусть в пространстве признаков задано множество образов $M=\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_m\}$, где $\mathbf{x}_k$ – вектор признаков k-го обра-

за, m – мощность множества. Будем считать, что центр первого кластера w_1 совпадает с любым из образов множества, например с $\mathbf{x}_1$, т.е. $w_1 = \{\mathbf{x}_1\}$. Далее вычисляется расстояние d_{21} между образом $\mathbf{x}_2$ и центром кластера w_1. Напомним, что способ вычисления расстояния между точкой и кластером выбирается разработчиком. Если значение расстояния больше заранее заданной пороговой величины t, то образ $\mathbf{x}_2$ принимается за центр нового кластера w_2, т.е. $w_2 = \{\mathbf{x}_2\}$. В противном случае образ $\mathbf{x}_2$ включается в кластер w_1, т.е. $w_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$. Для следующего образа оцениваются расстояния от него до имеющихся кластеров. Если **все** расстояния больше порога, то образ принимается за новый кластер. Если часть расстояний меньше порога, то образ относится к ближайшему кластеру. Процедура продолжается пока не будут исчерпаны все образы множества M.

Следует отметить, что при добавлении образа в кластер характеристики кластера пересчитываются, а результат кластеризации зависит от выбора порога и порядка просмотра образов из множества M.

Для кластеризации множества образов сложной формы, например, вытянутой или кольцеобразной, можно применить метод цепной кластеризации, являющийся разновидностью порогового алгоритма кластеризации. В данном методе за расстояние между точкой и кластером в пространстве признаков принимается расстояние от точки до ближайшего представителя кластера.

Метод заключается в следующем. В начале кластеризации произвольный образ считается принадлежащим к первому кластеру. Далее к данному кластеру присоединяются все образы, принадлежность которых к какому-либо кластеру еще не установлена, и расстояние от которых до исходного образа меньше заранее заданного порога t. Затем для каждого из присоединенных образов данная процедура повторяется. После того как к первому кластеру больше нельзя отнести ни одного образа, среди оставшихся не присоединенными образов в качестве исходного образа для второго кластера берется произвольный образ. Процедура повторяется до тех пор, пока не будут исчерпаны все образы.

Метод кластеризации слиянием тоже достаточно прост. В начале кластеризации каждый образ считается отдельным кластером, далее вычисляются расстоя-

ния между всеми кластерами, т.е. формируется квадратная, диагонально-симметричная таблица расстояний, строки и столбцы которой – имеющиеся кластеры. На каждом шаге сливаются два самых близких кластера, после чего размер таблицы уменьшается, а расстояния между кластерами пересчитываются. Процесс прекращается при достижении заданного числа кластеров или когда расстояние между ближайшими кластерами станет больше заданного порога.

Данный метод требует многократных вычислений изменяющихся на каждом шаге расстояний, что может стать достаточно трудоемкой задачей при большом количестве образов.

Метод кластеризации по k средним требует задания числа кластеров – k. На первом шаге в пространстве признаков произвольно выбирается положение k центров кластеров, не обязательно совпадающих с какими-либо образами. Далее на каждом шаге, во-первых, каждый образ относится к тому кластеру, расстояние до центра которого для него минимально, а во-вторых, после распределения всех образов по кластерам производится перерасчет положения центров кластеров. Процесс продолжается до тех пор, пока состав кластеров не стабилизируется.

Цель метода – минимизировать суммарное расстояние от центров кластеров до отнесенных к ним образов по всем кластерам. Возможно схождение процесса к локальному минимуму, а также отсутствие образов в некоторых кластерах по окончании процесса кластеризации.

Порядок выполнения работы

1. Разработать программу, выполняющую кластеризацию заданного множества образов с возможностью нормирования признаков и введения весов (способы задания расстояний, метод кластеризации и состав множества образов выбирается в соответствии с номером варианта).
2. Провести эксперименты по кластеризации, изменяя порог и/или число формируемых кластеров, а также используя нормирование признаков и весовые коэффициенты.
3. Проанализировать полученные результаты и подготовить отчет.

Содержание отчета

1. Цель работы и задание.
2. Используемые в программе расстояния и метод кластеризации.
3. Описание интерфейса и текст разработанной программы.
4. Результаты экспериментов.
5. Анализ результатов и выводы.

Контрольные вопросы

1. Чем характеризуется образ в пространстве признаков?
2. Что такое расстояние в метрическом пространстве?
3. Какие виды расстояний необходимо задавать для кластеризации?
4. Как вычисляются статистические характеристики кластера?
5. Какие методы кластеризации требуют задания порога?

Варианты заданий

В вариантах заданий, приведенных в таблице 1, используется следующая нумерация расстояний и методов кластеризации.

Расстояние между образами: 1 – Евклидово; 2 – Манхеттенское; 3 – доминирования; 4 – Камберра; 5 – косинусное.

Расстояние между образом и кластером: 1 – до центра кластера (в скобках – вариант расстояния между точками); 2 – до ближнего образа кластера (в скобках – вариант расстояния между точками); 3 – Евклидово с учетом дисперсии; 4 – Махаланобиса.

Расстояние между кластерами (в скобках – вариант расстояния между точками): 1 – между центрами; 2 – ближнего соседа; 3 – дальнего соседа; 4 – К-расстояние.

Метод кластеризации: 1 – пороговый; 2 – цепной; 3 – слиянием; 4 – по k средним.

Таблица 1 – Варианты заданий

№ вар.	Расстояние между …			Метод кластеризации	№ множества образов
	образами	кластером и образом	кластерами		
1	1, 3	1(1), 1(3)	-	1	1
2	2, 4	2(2), 2(4)	-	2	2
3	5	-	1(5)	3	3
4	1	1(1), 3(1)	-	4	4
5	2, 4	1(2), 1(4)	-	1	5
6	1, 3	2(1), 2(3)	-	2	6
7	2, 4	-	2(2), 2(4)	3	7
8	2	1(2), 4(2)	-	4	8
9	1	3(1)	-	1	9
10	5	2(5)	-	2	10
11	1, 3	-	3(1), 3(3)	3	1
12	1	1(1), 4(1)	-	4	2
13	1	4(1)	-	1	3
14	1, 3	-	1(1), 1(3)	3	4
15	5	1(5)	-	4	5
16	5	1(5)	-	1	6
17	1, 3	-	2(1), 2(3)	3	7
18	1	-	4(1)	3	8
19	2, 4	-	1(2), 1(4)	3	9
20	5	-	4(5)	3	10

Множество образов для кластеризации выбирается из таблицы 2 в соответствии с номером варианта задания.

Таблица 2 – Множества образов для кластеризации

№ мн.	Состав множества образов для кластеризации
1	(10,20,1), (0,1,7), (5,7,4), (7,1,0) (10,0,19), (0,12,7), (-5,-4,5), (20,10,15), (0,15,-16), (-1,9,-30), (6,18,4), (-5,-3,2), (18,0,17), (16,19,20)
2	(11,2,15), (5,18,4), (-10,-3,3), (18,15,14), (2,19,-13), (-3,5,-33), (13,5,12), (6,11,1), (-7,-2,2), (14,13,11), (4,20,-11), (-5,8,-21), (9,16,4), (-10,-2,3)
3	(15,6,15), (1,19,0), (-8,-1,4), (13,19,15), (15,17,-14), (-3,9,-35), (12,4,16), (8,14,9), (-6,0,5), (11,17,10), (12,17,-10), (-1,10,-25), (18,17,-11), (-4,9,-31)
4	(16,1,11), (9,13,8), (-9,-2,1), (17,12,14), (8,16,-19), (-2,7,-29), (19,7,10), (4,12,5), (-10,-4,3), (16,15,16), (6,16,-17), (-4,6,-32), (12,7,10), (4,20,7)
5	(20,3,19), (7,18,4), (-5,-5,2), (15,19,20), (11,19,-20), (-3,8,-30), (17,5,13), (6,15,3), (-8,-3,4), (11,13,18), (18,17,-15), (-4,7,-34), (0,5,5), (9,4,5)
6	(17,0,11), (5,13,0), (-5,-4,0), (14,16,18), (5,15,-11), (-3,10,-35), (16,2,15), (6,15,3), (-9,-2,5), (19,17,11), (6,13,-14), (-4,5,-25), (-6,0,1), (20,10,20)
7	(11,5,19), (9,10,4), (-7,0,3), (13,14,15), (12,19,-17), (-5,9,-30), (20,7,20), (0,19,9), (-6,-1,2), (13,11,18), (0,15,-20), (-2,6,-24), (15,12,20), (-2,10,-32)
8	(19,4,13), (8,14,10), (-6,-5,1), (20,20,20), (7,16,-17), (-1,7,-26), (15,1,10), (0,11,8), (-8,-1,5), (10,10,10), (12,15,-10), (-4,5,-27), (11,4,15), (3,15,6)
9	(13,6,14), (1,12,6), (-9,-3,0), (19,11,10), (4,5), (-5,8,-28), (12,5,12), (3,17,7), (-10,0,4), (15,16,17), (7,20,-14), (0,0,-34), (-7,-1,4), (13,17,11)
10	(14,3,16), (2,16,1), (-9,-4,3), (18,11,11), (11,15,-12), (-1,7,-32), (10,2,18), (7,19,2), (-8,-5,6), (15,14,18), (16,16,-20), (-3,6,-33), (19,19,-19), (-1,10,-25)