УДК 683.519

ЗАЙЧЕНКО Ю.П.,
ГОНЧАР М.А.

НЕЧЕТКИЕ МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ В ЭКОНОМИКЕ

В статье рассмотрена задача кластерного анализа в условиях неопределенности и описаны нечеткие методы кластерного анализа – *k*-средних и Густавссона-Кесселя. Проведены экспериментальные исследования предложенных алгоритмов на примере автоматической классификации стран СНГ по макроэкономическим показателям.

The problem of cluster analysis is considered and fuzzy clustering methods C-means and of Gustavssona-Kessel are described. The experimental investigations of the fuzzy clustering algorithm are described and it's application for automatic classification of CIS countries is presented.

Введение

Задачи кластер-анализа, или автоматической классификации получили широкое применение в экономике, социологии, медицине, геологии и других отраслях, всюду где имеются множества объектов произвольной природы, описываемых в виде векторов $x = \{x_1, x_2, \ldots, x_N\}$, которые необходимо автоматически разбить на группы однородных объектов по признакам «сходства-различия». В последние годы эти методы широко применяются в задачах интеллектуального анализа данных и Data mining. Традиционные методы кластер-анализа предполагают четкое разбиение исходного множества на подмножества, при котором каждая точка после разбиения попадает только в один кластер. Однако такое ограничение не всегда верно.

Зачастую необходимо произвести разбиение так, чтобы определить степень принадлежности каждого объекта к каждому множеству. В этом случае целесообразно использовать нечеткие методы кластер-анализа.

Цель настоящей статьи – провести исследование и анализ нечетких методов и алгоритмов кластер-анализа в экономике и финансовой сфере.

Рассмотрим постановку задачи нечеткого кластер-анализа.

1. Постановка задачи

Имеется N объектов $x = \{x_1, x_2, \ldots, x_N\}$, где $x_j = \lfloor x_{j1}, x_{j2}, \ldots, x_{jk} \rfloor$. Необходимо разбить их на k кластеров и определить места размещения центров кластеров c_i, $i = \overline{1,k}$.

Обозначим через u_{ij} – степень принадлежности точки x_j j-му кластеру.

На величины u_{ij} накладываются следующие ограничения [1]:

$$0 \le u_{ij} \le 1 \text{ и } \sum_{i=1}^{k} u_{ij} = 1 \text{ для всех } j. \qquad (1)$$

Требуется найти такое размещение центров кластеров c_i в пространстве объектов (x_1,…, x_n) и величины $\{u_{ij}\}$ при которых величина критерия (средневзвешенное отклонение точек x_j от центров кластеров) было минимально, т.е.:

$$\min_{C} E = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - c_i \right\|^2, \quad (2)$$

при условии (1), где $m > 1$, m – целое.

Для решения этой задачи разработаны нечеткий метод *k*-средних и алгоритм Густавссона-Кесселя.

2. Алгоритм нечёткой самоорганизации k-средних

Допустим, что в сети существует K нечетких нейронов с центрами в точках c_i (i = 1, 2, …, K). Начальные значения этих центров могут быть выбраны случайным образом из областей допустимых значений соответствующих компонентов векторов x_j $j = \overline{1, p}$, использованных для обучения. Пусть функция фаззификации задана в форме обобщенной функции Гаусса.

Подаваемый на вход сети вектор x_j будет принадлежать к различным группам, представляемым центрами c_i, в степени u_{ij}, причем $0 < u_{ij} < 1$, а суммарная степень принадлежности ко всем группам, очевидно, равна 1.

Поэтому

$$\sum_{i=1}^{K} u_{ij} = 1, \qquad (3) \qquad \text{для } j = \overline{1, p}.$$

Функцию погрешности, соответствующую такому представлению, можно определить как сумму частных погрешностей принадлежности к центрам c_i с учетом степени принадлежности u_{ij}. Следовательно,

$$E = \sum_{i=1}^{K} \sum_{j=1}^{p} u_{ij}^{m} \left\| c_i - x_j \right\|^2, \qquad (4)$$

где m – это весовой коэффициент, который принимает целочисленные значения из интервала [2,*N*].

Цель обучения с самоорганизацией состоит в таком подборе центров c_i, а также состава кластеров, чтобы для заданного множества обучающих векторов x_j – обеспечить достижение минимума функции (4) при одновременном соблюдении условий ограничения (3). Таким образом возникает задача минимизации нелинейной функции (4) с p ограничениями

типа (3). Решение этой задачи можно свести к минимизации функции Лагранжа, определенной в виде [1]:

$$LE = \sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij}^{m} \left\| c_i - x_j \right\|^2 + \sum_{j=1}^{p} \lambda_j \left(\sum_{i=1}^{K} u_{ij} - 1 \right), \quad (5)$$

Где λ_j (j=1,2, …, p) – это множители Лагранжа. В [1] доказано, что решение задачи (5) можно представить в виде

$$c_i = \frac{\sum_{j=1}^{p} u_{ij}^{m} x_j}{\sum_{j=1}^{p} u_{ij}^{m}}, \quad (6)$$

и

$$u_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}}, \quad (7)$$

где d_{ij} – это эвклидово расстояние между центром c_i, и вектором x_j, $d_{ij} = \left\| c_i - x_j \right\|$. Поскольку точные значения центров c_i, в начале процесса не известны, алгоритм решения данной задачи должен быть итерационным. Он может быть сформулирован в следующем виде:

Выполнить случайную инициализацию переменных u_{ij}, выбирая их значения из интервала [0,1] таким образом, чтобы соблюдалось условие (3).

Определить k центров c_i, в соответствии с (6).

Рассчитать значение функции погрешности согласно выражению (4). Если ее значение окажется ниже установленного порога либо, если уменьшение этой погрешности относительно предыдущей итерации пренебрежимо мало, то завершить вычисления. Последние значения центров составляют искомое решение. В противном случае перейти к п. 4.

Рассчитать новые значения u_{ij} по формуле (7) и перейти к п. 2.

Такую процедуру нечеткой самоорганизации будем называть алгоритмом C-means.

Многократное повторение итерационной процедуры ведет к достижению минимума функции E, которая необязательно будет глобальным минимумом. Качество находимых центров, оцениваемое значением функции погрешности E существенным образом зависит от предварительного

подбора как значений u_{ij}, так и центров c_i. Наилучшим может быть признано такое размещение центров, при котором они располагаются в областях, содержащих наибольшее количество предъявленных векторов x_j. При таком подборе центров они будут представлять векторы данных x_j с наименьшей суммарной погрешностью.

Поэтому начало итерационной процедуры расчета оптимальных значений центров должно предваряться процедурой их инициализации. К наиболее известным алгоритмам инициализации относятся алгоритмы пикового группирования и разностного группирования данных.

3. Алгоритм нечеткого кластер-анализа Густавссона-Кесселя

В классическом алгоритме *k*-средних (c-means) элементы выбираются с помощью обычного евклидового расстояния между вектором *x* и центром кластера *c*:

$$d(x,c) = \|x-c\| = \sqrt{(x-c)^T (x-c)}\,.$$

При таком задании расстояния между двумя векторами множество точек, равноудаленных от центра, принимает вид сферы с одинаковым масштабом по всем осям. Но если данные создают группы, форма которых отличается от сферической или если шкалы отдельных координат вектора сильно отличаются, в этом случае метрика становится неадекватной. В этом случае качество кластеризации можно значительно повысить за счет улучшенной версии алгоритма самоорганизации, который называется алгоритмом Густавссона-Кесселя.

Основные изменения относительно базового алгоритма *k*-средних в ведении в формулу расчета метрики масштабирующей матрицы *A*. При таком масштабировании расстояние между центром *c* и векторами *x* определяется формулой:

$$d(x,c) = \|x-c\| = \sqrt{(x-c)^T A(x-c)}$$

В качестве масштабирующей обычно используется положительно-определенная матрица, то есть матрица, у которой все собственные числа действительные и положительные.

Аналогично базовому алгоритму *k*-средних цель обучения с использованием алгоритма Густавсона-Кесселя в таком размещении центров, при котором минимизируется критерий:

$$E = \sum_i \sum_j u_{ij}^m d^2(x_j, c_i)\,. \qquad (8)$$

Описание алгоритма Густавссона-Кесселя:

Провести начальное размещение центров в пространстве данных. Создать элементарную форму масштабирующей матрицы A.

Сформировать матрицу коэффициентов принадлежностей всех векторов *x* к центрам *c* по формуле:

$$u_{ij} = \frac{1}{\sum_{k=1}^{K}\left(\frac{d^2(x_j, c_i)}{d^2(x_j, c_k)}\right)^{\frac{1}{m-1}}}. \qquad (9)$$

Рассчитать новое размещение центров *c* в соответствии с формулой:

$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}. \qquad (10)$$

Сгенерировать для каждого вектора матрицу ковариаций:

$$S_i = \sum_{j=1}^{N} u_{ij}^m (x_j - c_i)(x_j - c_i)^T. \qquad (11)$$

Рассчитать новую масштабную матрицу для каждого *i*-го центра по формуле:

$$A_i = \sqrt[n]{\det(S_i)} \cdot S_i^{-1}. \qquad (12)$$

Если последние изменения положений центров и матрицы ковариации достаточно малые по отношению к предыдущим значениям и не превышают заданной величины, то – завершить итерационный процесс.

Данные методы обладают следующим недостатком – число кластеров *k* должно быть известным. Для устранения первого недостатка и решения задачи кластерного анализа при незаданном числе кластеров разработан подход, основанный на вычислении величины: $\Delta E(k)=E(k-1)-E(k)$,
где $E(k-1)$ и $E(k)$ – оптимальное значение критерия для числа кластеров $(k-1)$ и (k).

Как только величина $\frac{\Delta E(k)}{E_0} \le \Theta$, где $\Theta = (0{,}2 \div 0{,}3)$ – величина заданного порога, процесс увеличения числа кластеров прекращается.

4. Применение нечетких методов *k*-средних и Густавссона-Кесселя в задачах автоматической классификации

Кластеризация регионов Украины по данным Госкомстата за январь-март 2005 года.

Исходные данные:

Табл.1

	Темпы роста промышленного производства	Темпы роста сельского хозяйства	Темпы роста строительства	Темпы роста розничной торговли
Автономная Республика Крым	**118,6**	**110,5**	**81.2**	**127,5**
Винницкая	**108,5**	**101,9**	**100,1**	**116,8**
Волынская	**117,2**	**105,7**	**83.1**	**117,9**
Днепропетровская	**112,3**	**117,1**	**126,9**	**115,4**
Донецкая	**94.4**	**113,3**	**115,9**	**120,2**
Житомирская	**114,8**	**109,5**	**86.7**	**113,6**
Закарпатская	**113,2**	**98.9**	**72.9**	**118,0**
Запорожская	**112,8**	**114,1**	**106,2**	**132,2**
Ивано-Франковкая	**106,2**	**96.4**	**91.5**	**120,7**
Киевская	**115,9**	**110,4**	**106,8**	**104,5**
Кировоградская	**115,1**	**82.3**	**118,9**	**120,3**
Луганская	**100,0**	**98.2**	**111,5**	**142,9**
Львовская	**101,3**	**100,1**	**102,7**	**114,4**
Николаевская	**110,9**	**101,6**	**106,5**	**120,1**
Одесская	**103,9**	**102,9**	**153,6**	**116,1**
Полтавская	**101,4**	**103,4**	**115,5**	**106,5**
Ровенская	**116,2**	**102,3**	**102,6**	**118,2**
Сумская	**117,4**	**100,7**	**90.5**	**115,1**
Тернопольская	**85.3**	**98.1**	**93.8**	**120,3**
Харьковская	**113,4**	**101,3**	**82.1**	**122,6**
Херсонская	**138,0**	**118,3**	**157,7**	**115,1**
Хмельницка	**133,2**	**94.5**	**78.3**	**112,6**
Черкасская	**112,4**	**112,5**	**94.5**	**110,5**
Черновецкая	**94**	**98.4**	**114,8**	**115,2**
Чернигивская	**109,6**	**103,3**	**115,7**	**108,6**
г.Киев	**117,2**	**100**	**310**	**119,7**
г.Севастополь	**107,3**	**100**	**107,7**	**128,7**

Табл. 2. Результаты кластеризации

Автономная Республика Крым	**0.980**	**0.018**	**0.001**	**0.000**
Винницкая	**0.000**	**0.996**	**0.002**	**0.001**

Волынская	**0.000**	**0.318**	**0.678**	**0.004**
Днепропетровская	**0.000**	**0.009**	**0.001**	**0.990**
Донецкая	**0.000**	**0.037**	**0.942**	**0.020**
Житомирская	**0.000**	**0.050**	**0.946**	**0.003**
Закарпатская	**0.576**	**0.419**	**0.003**	**0.002**
Запорожская	**0.988**	**0.010**	**0.000**	**0.001**
Ивано-Франковкая	**0.795**	**0.199**	**0.003**	**0.003**
Киевская	**0.000**	**0.142**	**0.812**	**0.046**
Кировоградская	**0.000**	**0.005**	**0.005**	**0.990**
Луганская	**0.000**	**0.991**	**0.007**	**0.002**
Львовская	**0.000**	**0.991**	**0.001**	**0.008**
Николаевская	**0.000**	**0.986**	**0.013**	**0.001**
Одесская	**0.000**	**0.814**	**0.108**	**0.078**
Полтавская	**0.000**	**0.985**	**0.002**	**0.012**
Ровенская	**0.000**	**0.988**	**0.009**	**0.003**
Сумская	**0.000**	**0.937**	**0.057**	**0.006**
Тернопольская	**0.000**	**0.992**	**0.001**	**0.007**
Харьковская	**0.000**	**0.916**	**0.082**	**0.002**
Херсонская	**0.000**	**0.016**	**0.000**	**0.984**
Хмельницка	**0.000**	**0.065**	**0.929**	**0.006**
Черкасская	**0.000**	**0.019**	**0.976**	**0.005**
Черновецкая	**0.000**	**0.617**	**0.002**	**0.381**
Чернигивская	**0.000**	**0.975**	**0.011**	**0.014**
г.Киев	**0.000**	**0.103**	**0.002**	**0.895**
г.Севастополь	**0.000**	**0.993**	**0.006**	**0.001**

Центры кластеров:

Первый	**113,38**	**107,19**	**90,46**	**126,35**
Второй	**105,90**	**100,98**	**104,07**	**119,21**
Третий	**114,49**	**107,71**	**94,33**	**113,18**
Четвертый	**119,57**	**104,36**	**124,62**	**117,42**

Из расположения центров кластеров видно:

В первом и третьем кластере находятся области с высокими темпами роста производства и сельского хозяйства и низкими темпами роста строительства. Различие между кластерами в темпах роста розничной торговли (для АР Крыма, Закарпатской, Запорожской и Ивано-Франковской областей (первый кластер) они высокие, для Волынской, Донецкой, Житомирской, Киевской, Хмельницкой, Черкасской (третий кластер) – ниже).

Во втором кластере находятся области с низкими темпами роста (Винницкая, Луганская, Львовская, Николаевская, Одесская, Полтавская, Ровенская, Сумская, Тернопольская, Харьковская, Сумская, Черновицкая, Черниговская, г. Севастополь).

Днепропетровская, Кировоградская, Херсонская области и г. Киев находятся в четвертом кластере (высокие темпы роста).

Значение минимизируемого критерия ошибки: E=4287.

Табл. 3. Расстояние между центрами кластеров

первый – второй	18,18588
второй – третий	15,82117
Третий – четвертый	31,18478
первый – третий	13,78145
второй – четвертый	24,97599
первый – четвертый	35,95797

Как видим, наименьшее расстояние между первым и третьим кластерами: собственно различия только в значениях четвертого параметра. Наибольшее – между первым и четвертым и между третьим и четвертым кластерами.

Заключение

1. В статье сформулирована задача нечеткой кластеризации экономических объектов по их характеристикам.
2. Описаны алгоритмы нечеткой классификации *k*-средних и Густавсона-Кесселя и описан подход к их использованию в случаях, когда число кластеров *k* заранее не задано.
3. Проведены экспериментальные исследования разработанных алгоритмов кластер-анализа на примере автоматической классификации стран СНГ.

Список использованной литературы

1. Зайченко Ю.П. Основи проектування інтелектуальних систем. Навчальний посібник. – К.: Видавничий Дім «Слово», 2004. – 352с.
2. Никифорова Н.С. Кластерный анализ в задачах социально-экономического прогнозирования.