УДК 519.81

ВЫДЕЛЕНИЕ ОДНОРОДНЫХ СОЦИАЛЬНЫХ ГРУПП В ОРГАНИЗАЦИОННЫХ СИСТЕМАХ

ОВЕЗГЕЛЬДЫЕВ А.О., ПИСКЛАКОВА В.П.
(Системы и процессы управления)

Рассмотрена задача классификации социальных групп по критерию близости матриц предпочтения факторов, определяющих поведение. Предложена универсальная функция оценки внутригруппового и межгруппового расстояния, основанная на функции обобщенного среднего Колмогорова.

Одним из условий эффективного управления организационными системами является корректный учет и управление поведением социальных групп, являющихся элементами системы. Постановка проблемы управления поведением группы предполагает, что известна усредненная матрица $A_{ср}$ предпочтений факторов, определяющих поведение группы.

При разработке процедур определения усредненного значения вектора предпочтений предполагается, что $A_{ср}$ вычисляется на однородном множестве векторов индивидуального предпочтения $A_{и}$. Такая однородность может быть обеспечена в следующих случаях.

1. Один ЛПР многократно принимает решения на однородном множестве ситуаций. Здесь под однородностью понимается близость по содержательной и целевой установке, по структуре характеризующих факторов и т.д.

2. Одна ситуация оценивается независимо членами группы ЛПР. В этом случае однородность предусматривает близость по целевым установкам, квалификации, информированности и т.д. экспертов.

3. Из множества всех идентифицированных векторов предпочтения $A_{и}$ выделены однородные, т.е. близкие по значению компонент, группы.

В первом случае возникает задача типизации ситуаций принятия решений, во втором и третьем - выделения типичных групп ЛПР. Примером последних двух задач является проблема структуризации множества покупателей и сегментации рынка товаров. Каждый покупатель совершая покупку выступает как ЛПР, осуществляющий выбор из множества однотипных (одинакового функционального назначения) товаров. Естественно покупатели различаются по возрасту, полу, месту проживания, душевому доходу, социальному статусу и т.д. Для принятия обоснованных маркетинговых решений необходимо выделить типичные группы покупателей. Это можно сделать двумя способами:

- сгруппировать покупателей по их объективным характеристикам и затем для каждой группы по множеству индивидуальных векторов предпочтений $A_{и}$ определить усредненный $A_{ср}$;
- выделить группы близких по значению векторов $A_{и}$ и затем решить задачу определения группы ЛПР, которая им соответствует.

Оба способа взаимосвязаны и дополняют друг друга. Особенно важны задачи группирования при анализе сложных социально-экономических систем, таких как муниципальные, региональные и т.д. Такие системы состоят из неоднородных социальных, производственных и т.д. групп и управление ими невозможно без четкой структуризации на типичные группы.

Независимо от содержательной постановки все описанные выше задачи являются задачами классификации множества объектов. Они формулируется следующим образом. Задано множество сравнимых между собой объектов, т.е. объектов, которые характеризуются одинаковым по количеству и смыслу набором характеристик (факторов). Это не исключает, что некоторые характеристики принимают нулевое значение. Необходимо на исходном множестве выделить близкие по значению характеристик группы объектов.

Каждый объект может быть интерпретирован как точка в многомерном пространстве характеристик (факторном пространстве). В такой интерпретации задача классификации состоит в решении одной из следующих задач:

- разбиение факторного пространства на непересекающиеся области;
- выделение групп (скоплений) точек, основанном на естественном расслоении исходного множества.

Для целей настоящего исследования интерес представляет вторая задача, при этом для определенности, но без потери общности, будем рассматривать классификацию векторов индивидуальных предпочтений $A_{и}$.

Концептуальным моментом проблемы классификации является формализация понятия «близость» объектов в пространстве характеристик и метрики, в которой она измеряется. С этой точки зрения можно выделить два основных подхода к задаче классификации.

Первый основан на предположении, что более или менее точно известны статистические характеристики генеральной совокупности и

классов, которые ее составляют. Это направление является наиболее глубоко разработанным и опирается на мощные статистические методы, например критерии максимального правдоподобия, байесовские оценки и т.д. [1, 2].

Второе направление связано с решением задачи классификации, когда априорные статистические параметры генеральной совокупности и классов неизвестны и их невозможно определить. В этом случае задача классификации решается на основе понятия расстояния между любой парой объектов или некоторой функции, характеризующей степень близости (сходства). Подобные методы получили название непараметрических, а все направление в целом - кластер-анализа или таксономии [3]. В рамках этого направления и будем рассматривать задачу разбиения множества индивидуальных оценок $A_и$ на однородные группы.

Исходная информация представляет собой множество векторов индивидуального предпочтения A_{ij}, $j = \overline{1,m}$, представленных в виде матрицы. Каждый вектор A_{ij} может быть интерпретирован как точка в n-мерном пространстве, лежащая на плоскости $\sum_{i=1}^{n} a_i = 1$, где a_i - компоненты матрицы индивидуального предпочтения $A_и$, n - ее размерность.

Основой выделения групп должна быть близость индивидуальных векторов предпочтения. Трудность решения данной задачи состоит в отсутствии априорной информации о количестве и характеристиках возможных групп, т.е. с формальной стороны это – задача классификации (кластеризации) без учителя. Ее суть заключается в том, чтобы заданное множество многомерных точек разбить на группы, представляющие собой «сгустки» точек. Основанием для группирования служат результаты анализа внутригрупповых и межгрупповых расстояний, определенных в принятой метрике [3]. Таким образом, необходимо найти рациональные, согласно принятым критериям, число и состав групп (классов) точек, т.е. решить задачу кластерного анализа.

Конкретизация критериев классификации связана с выбором метрики и уточнением вида критерия. Здесь нужно иметь в виду следующее. Формализация практически всех этапов кластеризации основана на эвристических соображениях, что адекватно выдвижению набора аксиом, выбор которых практически предопределяет решение, т.е. число и состав классов. Кроме того, решение существенно зависит от топологии точек классифицируемого множества. Поскольку классификация проводится при отсутствии информации о числе и характеристиках классов, нет уверенности в том, что принятая аксиоматика позволит выявить действительно объективные группировки точек. Таким образом, необходимо гарантировать выбор устойчивых классификационных решений, инвариантных до некоторой степени к критериям и алгоритмам кластеризации.

Одним из путей обеспечения такой устойчивости решения, наряду с глубокой аргументацией эвристики, положенной в основу формализации, является многократное решение задачи классификации при различных критериях, метриках, алгоритмах. Если несколько решений совпадает, то с большой долей уверенности можно предположить, что они отражают некоторые объективные закономерности. В противном случае у ЛПР имеется возможность проанализировать возможные варианты классификации и сознательно выбрать единственное решение. С этой точки зрения при выборе вида критерия близости необходимо стремиться к его универсальности в том смысле, чтобы он допускал при единой форме реализацию разных метрик в пространстве классификации и различных алгоритмов. В связи с этим заслуживает внимания критерий обобщенного среднего Колмогорова [3]:

$$\rho(A,B) = F^{-1}\left\{\frac{1}{n_1 n_2}\sum_{i\in A}\sum_{j\in B} F[\tau(x_i,x_j)]\right\}, \quad (1)$$

где $\rho(A,B)$ – расстояние между классами $A = \{x_i\}$, $i = \overline{1,n_1}$, $B = \{x_j\}$, $j = \overline{1,n_2}$;

F – некоторый оператор, определяющий конкретный вид критерия;

n_1, n_2 – число элементов в классах A, B;

$\tau(x_i,x_j)$ – расстояние между i-ми и j-ми точками.

Его вид обусловлен принятой в пространстве классификации метрикой.

Проведенные А. И. Орловым [4] исследования устойчивости классификационных решений показали, что выбор вида оператора F существенно зависит от шкалы, в которой проведены измерения. Для шкалы отношений, в которой измеряются векторы предпочтений индивидуумов, целесообразно применять логарифмические или степенные операторы. В случае использования степенного оператора F критерии (1) записывается следующим образом:

$$\rho(A,B) = \left[\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \tau^c(x_i, x_j) \right]^{1/c}. \quad (2)$$

Критерий вида (2) удобен и в отношении реализации различных метрик пространства. Наиболее часто используются евклидова метрика, τ_1 - супремум-, инфимум-нормы и расстояние Махаланобиса. Но все эти метрики, за исключением последней являются частными случаями τ_p -нормы вида

$$\tau(x_i, x_j) = \left[\sum_{k=1}^{K} |x_{ki} - x_{kj}|^p \right]^{1/p}. \quad (3)$$

Здесь k – размерность вектора x, $k = \overline{1,K}$. Действительно, при p=1 получаем τ_1-норму, при p=2 – евклидово расстояние τ_2, при $p \to \infty$ – супремум-норму τ_{sup}, а при $p \to \infty$ – инфимум-норму τ_{inf}.

С учетом выражения (3) критерий (2) в общем случае имеет вид

$$\rho(A,B) = \left\{ \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left[\sum_{k=1}^{K} |x_{ki} - x_{kj}|^p \right]^{p/c} \right\}^{1/c}. \quad (4)$$

Критерий (4) удобен при алгоритмизации, так как позволяет определять расстояние не только между кластерами, но и внутригрупповое расстояние от любой точки. Для последнего варианта критерий (4) примет вид

$$\rho(x_a, x_i) = \left[\frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} |x_{ka} - x_{ki}|^p \right)^{p/c} \right]^{1/c}, \quad (5)$$

$$i = \overline{1,n}; \quad a \in i.$$

Здесь n – число элементов в группе.

Перейдем к выбору алгоритма кластеризации. В общем случае задача классификации (кластеризации) без учителя представляет собой комбинаторную задачу, размерность которой при достаточно большом числе классифицируемых объектов такова, что ее практически невозможно реализовать на ЭВМ. Поэтому широко распространены различные эвристические алгоритмы классификации [3, 5, 6], что позволяет создать банк алгоритмов и реализовать идею многократного решения задачи классификации. Вместе с этим анализ опубликованных алгоритмов свидетельствует о серьезном недостатке, присущем всем им: отсутствует обоснование выбора числа групп. Так, в группе дендрограммных алгоритмов [5] число классов определяется критерием среза, в максиминном алгоритме [6] использован критерий расстояния между группами, в алгоритме ФОРЭЛЬ [3] – радиус сферы. Но во всех случаях нет рекомендаций по выбору значений критериев. Поэтому желателен подход, который хотя бы на эвристическом уровне обосновывал критерий выбора числа групп.

Такой критерий, как следует из содержательной постановки задачи классификации, должен одновременно учитывать два аспекта: плотность групп, т. е. внутригрупповое расстояние, и расстояние между группами. В основу синтеза критерия в данной работе положено следующее эвристическое предположение: оптимальна по числу классов и их составу такая классификация, которая минимизирует суммарное среднее расстояние на множестве классифицируемых точек, учитывающее внутри- и межгрупповые расстояния. Отметим противоречивую тенденцию: с ростом числа групп уменьшается внутригрупповое расстояние (в пределе оно нулевое при числе элементов в группе, равном единице), но возрастает суммарное межгрупповое расстояние. Это противоречие открывает путь к построению формального критерия, позволяющего найти компромисс между числом групп и их размером (составом).

Для формализации критерия примем следующие допущения. Межгрупповым является расстояние от центра группы до центра классифицируемого множества, нормированное по числу элементов в группе, а внутригрупповым – среднее расстояний от элементов группы до ее центра. Последнее определяется как суммарное расстояние от элементов до центра группы, нормированное по числу точек в ней. Таким образом, каждый i-й кластер характеризуется показателем

$$s_i = \tau_{1j} / n_i + \rho_{ij}, \; j = \overline{1,n_i}, \quad (6)$$

где j – номер элемента группы;

n_i – общее число элементов в группе.

Из определения ясно, что параметр τ_{1i} рассчитывается по формуле (3), а p_{ij} – по формуле (4). При этом в качестве центров множества и группы принимаются элементы, для которых показатель (6) минимален на всем множестве элементов и подмножестве, входящем в группу,

соответственно. С учетом равенства (6) критерий классификации имеет вид

$$S = \min_{N, A_i, x_i} \sum_{i=1}^{N} s_i, \quad i = \overline{1,N}. \qquad (7)$$

Здесь переменные оптимизации – количество групп N, их состав A_i и положение центра каждой группы $x_i \in A_i$.

Анализ критерия (7) показывает, что в случае, когда количество классов и места расположения их центров заданы, минимум достигается, если объекты включаются в класс, до центра которого расстояние наименьшее. Это означает, что при решении задачи классификации в критерии (7) можно перейти от трех оптимизируемых переменных к двум – числу классов N и расположению их центров x_i, а состав классов A_i, определяется однозначно по минимуму расстояния элементов до центров.

Рассмотрим зависимость значения критерия (7) от местоположения центров классов при фиксированном их числе. Если центры классов совпадают с классифицируемыми объектами, зависимость (7) – многоэкстремальная решетчатая функция. Область ее существования обусловлена числом сочетаний C_n^N, n – общее число элементов. Поэтому выявление глобального минимума (7) при фиксированном числе N связано с перебором всех сочетаний мест размещения центров и определением для них значений функций.

Чтобы обосновать алгоритм нахождения числа классов N, рассмотрим свойства огибающей глобальных минимумов критерия (7) как функции числа классов $S^*(N)$.

Утверждение 1. Функция $S^*(N)$ гладкая, выпуклая вниз, и ее значения при N = 1, N = n совпадают и являются максимальными.

Представим $S^*(N)$ с учетом выражения (6) в виде

$$S^*(N) = \sum_{i=1}^{N} \frac{\tau_{1i}}{n_i} + \sum_{i=1}^{N} \rho_{ij}, \quad N = 1,2,\ldots,n, \qquad (8)$$

где первое слагаемое – нормированное по числу элементов класса расстояние от центра i-го класса до центра классифицируемого множества элементов; второе слагаемое – сумма расстояний от элементов класса до его центра, нормированное по числу элементов. Значения функции (8) при N = 1 и N = n совпадают и равны сумме расстояний от элементов классифицируемого множества до его центра. Так, при N = 1 имеем: $\tau_{11} = 0$, $\rho_{ij} = \rho_{1i} = \tau_{1j}$, $j = \overline{1,n}$. При N = n (т.е. каждый элемент являете классом) $\rho_{ij} = 0$; $n_i = 1$, $i = \overline{1,n}$. Таким образом, в обоих случаях

$$S^*(N = 1) = S^*(N = n) = \sum_{i=1}^{n} \tau_{1i}. \qquad (9)$$

Это – максимальное из всех возможных значений функции $S^*(N)$. Выпуклость функции $S^*(N)$ вытекает из того, что первое слагаемое выражения (8) с ростом N монотонно возрастает от нуля при

N = 1 до максимального значения зависимости (9) при N = n, а второе – соответственно монотонно убывает от максимального значения функции (9) до нуля. Монотонность обеспечивается по определению кривой $S^*(N)$, как огибающей глобальных экстремумов. Таким образом, функция $S^*(N)$ одноэкстремальна и выпукла вниз, поскольку для любого $N \neq 1; N \neq n$ значение функции меньше максимального.

Утверждение 2. Минимум функции $S^*(N)$ достигается при числе классов, удовлетворяющем условиям

$$1 \le N \le n/2 \text{ для четных n;}$$
$$1 \le N \le (n-1)/2 \text{ для нечетных n.} \qquad (10)$$

Здесь N – число классов, содержащих более одного элемента. Это объясняется тем, что при N < n, в силу изложенного правила отнесения к классам, группы с одним элементом существовать не могут. Этот элемент обязательно относится к классу с ближайшим центром.

Из приведенного анализа вытекает, что исходную задачу оптимизации (7) можно представить в виде

$$S = \min_{N} \min_{C_n^N} \sum_{i=1}^{N} s_i. \qquad (11)$$

Не касаясь пока способа определения глобального минимума на множестве сочетаний C_n^N при фиксированном N, отметим, что характер зависимости $S^*(N)$ как огибающей локальных экстремумов открывает возможность построить алгоритм целенаправленного перебора, основанный на последовательном вычислении функции (8) при значениях N = 1, 2, . . ., до нахождения минимума.

Рассмотрим вычислительный аспект определения центра группы. При фиксированном

числе групп N зависимость критерия (7) от местоположения их центров является многоэкстремальной решетчатой функцией. Чтобы найти ее глобальный экстремум, нужно решать комбинаторную задачу, связанную с перебором C_n^N сочетаний и определением для каждого из них значения функции. Такая задача очень громоздка в вычислительном отношении и при больших n практически неразрешима. Поэтому необходимо разработать эвристические процедуры, которые хотя и не обеспечивают в общем случае достижения глобального экстремума, но дают решения, достаточно близкие к оптимальным, при гораздо меньших затратах времени. Здесь возможны разные подходы, в частности, основанные на различных модификациях методов последовательно-одиночного размещения, случайного поиска и т.д.

В данной работе предлагается использовать аналог метода покоординатного спуска на множестве сочетаний C_n^N.

Рассматриваемый алгоритм осуществляет поиск местоположения заданного числа центров классов, а следовательно, и. разбиение на классы, т. е. определение их состава, минимизирующего выбранный критерий качества. Задача решается путем последовательного перемещения одного центра (при фиксированном положении других) в пространстве возможных точек размещения до тех пор, пока не будет вычислен локальный по данной переменной экстремум. Затем процесс повторяется для следующей переменной (центра) и т.д. до нахождения глобального экстремума. Если не наложены ограничения на число шагов или радиус сферы возможного движения каждого центра, то алгоритм реализует полный перебор сочетаний C_n^N возможного размещения N центров на n допустимых точках. Но это требует больших затрат времени. Поэтому необходимо ввести эвристические процедуры, ограничивающие перебор. Возможны различные подходы к организации такой процедуры. В данной работе принято ограничение на число шагов перемещения каждого центра, не приводящих к улучшению функции цели. Кроме того, поскольку исследуемая поверхность многоэкстремальна, для поиска глобального экстремума используется традиционный подход, основанный на многократной реализации алгоритма спуска из различных начальных приближений (точек пространства).

Описываемый алгоритм спуска может быть использован самостоятельно для решения задачи классификации при заданном числе классов N и как подпрограмма алгоритма решения более общей задачи классификации, когда N – один из оптимизируемых параметров.

Литература: 1. *Рао С.Р.* Линейные статистические методы и их применение. - М.: Наука, 1968. - 240 с.2. *Овезгельдыев А.О.* Взаимосвязь задач оценивания и классификации в проблеме принятия решений // Труды 3 Междунар. конф. «Теория и техника передачи, приема и обработки информации». - Туапсе: Харьков, ХТУРЭ. - 1997. - с. 271.3. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с. 4. *Орлов А.И.* Устойчивость в социально-экономической моделях. – М.: Наука, 1979. – 296 с. 5. *Болч Б., Хуань К. Дж.* Многомерные статистические методы для экономики: Пер. с англ. - М.: Статистика, 1979. - 318с. 6. *Ту Дж., Гонсалес Р.* Принципы распознавания образов. - М.: Мир, 1978. - 412с. 7. *Петров Э.Г., Аннамухамедов О.Б., Овезгельдыев А.О.* и др. Синтез информационно–вычислительного обеспечения распределенных АСПИ. Часть 1. Методологические и инструментальные основы синтеза ИВС. – Ашхабад, Издательство АН ТССР «Ылым», 1988. – 198 с.

Сведения об авторах: **Овезгельдыев Атагельды Оразгельдыевич,** кандидат техн. наук, докторант кафедры системотехники, Харьковский государственный технический университет радиоэлектроники. Научные интересы - многофакторное оценивание и многокритериальная оптимизация. Служебный адрес: 310726, г. Харьков, пр. Ленина. 14, контактный телефон 40-93-06.

Писклакова Валентина Петровна, ст. научн. сотр., кандидат техн. наук, директор Центра информатизации органов управления, Харьковский государственный технический университет радиоэлектроники. Научные интересы - информатизация процессов управления. Служебный адрес: 310726, г. Харьков, пр. Ленина. 14, контактный телефон 30-24-29.

УДК 519.81

Выделение однородных социальных групп в организационных системах/А.О.Овезгельдыев, В.П. Писклакова//РИ. 1999. № 4. С. 00-00

Рассмотрена задача выделения однородных по близости предпочтений социальных групп в социально-экономических системах. Предложены критерии и алгоритмы определения числа и состава групп на множестве оценок, характеризующих индивидуальные предпочтения.

УДК 519.81

Виділення однорідних соціальних груп в організаційних системах/А.О.Овезгельдиєв, В.П. Пісклакова//РІ. 1999. № 4. С. 00-00

Розглянута задача виділення однорідних по близькості переваг соціальних груп у соціально-

економічних системах. Запропоновані критерії та алгоритми визначення числа та складу груп на множині оцінок, що характеризують індівідуальні переваги.

Рецензент: д.т.н.. проф. каф. фундаментальных дисциплин Харьковского института пожарной безопасности

В.М. Комяк