УДК 519.237.8 : 519.87 : 004.9

И.Г. Яр-Мухамедов, aldar@email.su
Институт машиноведения и автоматики НАН КР

ЛИНЕЙНАЯ МОДЕЛЬ КЛАСТЕРА

В статье рассмотрены специфические черты кластерного анализа и показаны некоторые возможности решения его проблем с использованием аппарата целочисленного линейного программирования. Представлен один из путей формализации модели кластерного анализа, сведения ее к системе линейных ограничений и целевой функции, отражающей одну из сторон содержательного определения кластера. Проанализированы недостатки и достоинства предложенного подхода, к числу которых относятся возможности точной формулировки критерия и получения решения, соответствующего глобальному оптимуму целевой функции.

Ключевые слова: кластерный анализ, целочисленное линейное программирование, технология моделирования, линейная модель кластерной структуры

Введение

Одним из наиболее общих и фундаментальных методов познания является метод группировки [1]. Он представляет собой одно из незаменимых средств снижения разнообразия и построения классификаций. Последние, как известно, составляют одну из важных ступенек построения научного знания. Человек классифицирует все, что встречается в окружающем его мире. Например, самих людей рассматривают в разрезе самых различных признаков. Общепринятыми являются деления по полу, возрасту, образованию, специальности, квалификации и т.д. Однако такой подход остается поверхностным и позволяет выявлять лишь некоторые сугубо внешние особенности явлений, не проникая глубже в их суть. Кластерный анализ как раз и призван выявлять внутреннюю структуру явлений, представляемых совокупностью объектов, которые рассматриваются в разрезе не одного, а целого ряда признаков [2, 3]. Причем эти признаки анализируются не в отдельности, а в их взаимной связи, что позволяет надеяться на познание сущности как некоторой целостности, обладающей внутренней структурой.

Исходные данные для выполнения процедур кластерного анализа принято представлять в виде таблиц «объект-свойство» [4], где строки соответствуют отдельным объектам, а столбцы – свойствам или признакам. На пересечении строки и столбца располагается значение свойства соответствующего объекта. Значения обычно являются числовыми и бывают выражены в какой-либо из шкал, адекватно представляющих спектр значений признака. В математической форме таблица записывается как матрица: $X = (x_{ij})$. Первый индекс – индекс объекта, второй – признака. И здесь мы сталкиваемся с первой проблемой.

Сколько признаков или свойств есть у объекта? Мы этого не знаем. Знаем только, что много. Причем больше того количества, которое мы способны наблюдать. И в этих условиях провозглашается, что кластерный анализ выявляет «объективную» внутреннюю структуру множества кластеризуемых объектов. С этих позиций мы можем утверждать, что степень объективности ограничена нашими возможностями по наблюдению за объектами. Но далее обычно следуют рекомендации по отбору признаков с точки зрения их существенности для кластеризации. Но в природе нет таких градаций в признаках. Существенность сама по себе не присуща тем или иным признакам. Признак может быть существенным либо несущественным лишь для достижения конкретной цели. Здесь уже затрагивается даже не семантический, а прагматический аспект, главным действующим лицом которого наряду с объектом является субъект с его устремлениями, желаниями и целями. Это означает, что аналитическая составляющая кластерного анализа в определенной мере подчинена решению задач синтеза, которые, хоть и в неявной форме, занимают определенное место в рамках данной дисциплины. Впрочем, так мы можем сказать практически о любой дисциплине, в названии которой есть слово «анализ».

Суть второй проблемы заключается в том, что различные признаки объектов измеряются в различных единицах измерения. Например, возраст – в годах, рост – в сантиметрах, а зарплата – в сомах. Как их сопоставлять и соизмерять? Признаки качественно различны и количественно несоизмеримы. Для математической обработки требуется выразить значения всех признаков в одних и тех же единицах измерения. Одним из решений, повсеместно используемых, является стандартизация исходных данных. Из значений признаков вычитают их среднее (операция центрирования) и делят на среднее квадратическое отклонение (операция нормирования):

$$z_{ij} = \frac{x_{ij} - \bar{x}}{s_x} \; .$$

Полученные стандартизованные данные (они обозначены буквой z) представлены в виде отклонений и выражены в единицах собственного среднего квадратического отклонения, то есть являются безразмерными. Это не всегда может быть удобным, но позволяет привлечь всю мощь математики. Правда иногда, для некоторых процедур, значения признаков центрируют с помощью минимального значения и выражают в единицах размаха вариации.

Третья проблема связана с выбором мер расстояния или сходства. Для простоты мы будем оперировать мерами расстояния. В кластерном анализе используются два типа мер: меры расстояния между объектами и меры расстояния между кластерами. Для первых будем использовать наиболее привычную форму:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{kj})^2} \; .$$

Здесь показано расстояние между объектами i и k, вычисленное по всем m признакам. Совокупность межобъектных расстояний (или значений мер сходства) собирается в таблицу типа «объект-объект», которая является квадратной и симметричной относительно главной диагонали. В матричном виде она может быть представлена как $D = (d_{ik})$.

Выбор межобъектной меры может оказывать влияние на результаты кластеризации, но еще большее влияние оказывает выбор межкластерной меры. К примеру, мера ближнего соседа выявляет цепочки связанных объектов, а дальнего соседа шаровые скопления. И это при использовании одних и тех же алгоритмов кластеризации.

Четвертая проблема связана моделью и с критериями кластерного анализа. Определение кластера, неразрывно связанное с определением кластерной структуры, формулируется на содержательном уровне так [3]:

1) в состав каждого кластера входят объекты, расположенные близко друг к другу либо схожие друг с другом;
2) объекты, входящие в состав различных кластеров, несходны либо далеки друг от друга;
3) количество кластеров существенно меньше количества объектов.

В этом проявляется одна из черт высокой степени фундаментальности кластерного анализа. Его модель формулируется в декларативных и предельно общих теоретико-множественных понятиях. Формально мы можем представить ее следующим образом:

$$o_i \in c_p \cap o_k \in c_p \Leftrightarrow d_{ik} < d^c, \quad i \neq k \; ;$$

$$o_i \in c_p \cap o_k \in c_q \Leftrightarrow d_{ik} \geq d^c, \quad i \neq k, \quad p \neq q \; ;$$

$$pow(o) > pow(c) \; .$$

Здесь: $o=\{o_i: i=1,\ldots,n\}$ – множество объектов; $c=\{c_p: p=1,\ldots,l\}$ – множество кластеров; d^c – критическое значение расстояния.

Данное формальное описание несколько условно, но вполне подходит для прояснения сути последующего изложения.

Как видим, в данной модели нет целевой функции. И критерий, и ограничения слиты воедино. С одной стороны, это обеспечивает определенный простор для трактовок в конкретных задачах, но с другой – не дает опоры для оценки степени достижения целей и оценки качества полученного решения.

Надо отметить, что первое и второе условия сильно связаны и в значительной мере дополняют друг друга. По этой причине, наверное, большинство методов и алгоритмов кластерного анализа оперируют одним из этих свойств. Чаще всего это свойство близости или сходства объектов одного кластера.

Пятая проблема кластерного анализа связана с его методами. Как правило, это методы, конечно, шаговые, переборные, находящие некоторые субоптимальные решения. При этом не может быть уверенности в том, что решение является хотя бы одним из наилучших. Не исключено, что критерий, неявно учитываемый в правилах поиска решения, имеет несколько экстремумов. Причем мы не знаем, что за экстремум достигнут в результате поиска.

По этим причинам в руководствах по пакетам кластерного (статистического) анализа известных и хорошо зарекомендовавших себя компаний-разработчиков можно встретить следующие рекомендации.

1. Массив исходных данных перед расчетами следует рандомизировать, так как многие алгоритмы чувствительны к порядку расположения данных.
2. Для получения хороших результатов следует повторить расчеты несколько раз, каждый раз изменяя порядок расположения данных, после чего выбрать приемлемый вариант решения.

Отметим сразу, что выполнение этих рекомендаций вовсе не гарантирует желаемого результата, так как общее количество решений обычно очень велико, а вид критериальной функции неизвестен.

Как уже упоминалось, мы попробуем представить задачу кластерного анализа в терминах целочисленного линейного программирования (ЦЛП). Как отмечает создатель этой дисциплины [5, 6, 7], модель состоит из линейной целевой функции, совокупности функциональных линейных ограничений в форме равенств и неравенств и параметрических ограничений на возможные значения переменных модели. Технология моделирования и решения задачи включает следующие этапы [5, 8].

1. Формализация положений теории, имеющей отношение к рассматриваемой задаче. Цель преобразуется в целевую функцию. Возможности и ограничения принимают форму системы ограничений, задающей область определения задачи.
2. Первоначальная математическая модель преобразуется в каноническую либо стандартную форму.
3. Выполняется «кодирование» задачи на языке пакета либо программы математического программирования. Используется матричная форма представления данных либо развернутая форма представления модели, включающая идентификаторы переменных и знаки операций.
4. Решение задачи с помощью программ и функций пакета программ.
5. Интерпретация полученного решения – перевод на язык предметной области.

Первый этап является наименее формализованным из всех, и далее мы уделим ему все внимание. Остальные этапы сводятся к выполнению хорошо известных и подчас рутинных операций.

Целевая функция

Как мы уже отмечали, большинство методов и алгоритмов кластерного анализа не используют, по крайней мере явно это понятие. Для описания задачи в терминах линейного программирования целевая функция необходима. За основу возьмем первое из свойств кластера. Мы можем представить его в виде:

$$y_p = \sum_{i,j \in Ip} d_{ij}, \quad p = \overline{1,l}\,. \qquad (1)$$

Здесь I_p – множество индексов объектов, входящих в кластер с индексом p. Тогда целевая функция запишется как:

$$Y = \sum_{p=1}^{l} y_p \to \min . \qquad (2)$$

Эта функция линейная, но она не может рассматриваться как целевая, так как в ее составе нет искомых переменных. Расстояния – это константы. Преобразуем полученные выражения, введя логические переменные x_{ij}, значения которых имеют следующий смысл: 0 – объекты с соответствующими индексами не входят в состав одного кластера; 1 – входят в состав одного и того же кластера.

$$y_p = \sum_{i,\, j \in Ip} d_{ij} x_{ij}, \quad p = \overline{1, l} . \qquad (3)$$

Из определения кластерной структуры и результата формализации первого из ее соотношений следует, что оптимум достигается при количестве кластеров, равном количеству объектов. При этом значение целевой функции равно нулю. Если мы рассмотрим второе соотношение, то сумма межкластерных расстояний окажется максимальной. В качестве ограничения мы можем взять только третье соотношение. Но каким должно быть это ограничивающее значение? Определение кластера не дает ответа на этот вопрос. Для этих целей должна быть привлечена и действительно привлекается дополнительная информация. Информация прагматического характера.

Сразу отметим, что предложенный вариант целевой функции не является единственно возможным. В частности, может представлять интерес целевая функция, минимизирующая количество кластеров в кластерной структуре. При необходимости, как известно, любое из функциональных ограничений может быть преобразовано в целевую функцию, если в этом есть смысл и необходимость.

Система ограничений

Первая группа функциональных ограничений определяет возможные связи объектов кластеризации с кластерами через логические переменные:

$$\sum_{k=1}^{l} x_i^k = 1, \quad i = \overline{1, n} . \qquad (4)$$

Здесь переменная *x* является логической, и ее единичное значение показывает, что объект с индексом *i* отнесен к кластеру с индексом *k*. Каждое из уравнений связывает один из объектов с одним из кластеров. Мы остановимся именно на такой базовой трактовке. Но вариантов может быть много, например, таких, как возможность отнесения объекта не к одному, а сразу к нескольким кластерам, если это соответствует специфике предметной области.

Следующая группа соотношений связывает переменные двух типов, а именно переменные, необходимые для выбора близких друг к другу объектов, с переменными, определяющими отношение объектов к кластерам. Это отношение характеризуется тем, что единичное значение переменной одного типа достигается только и только тогда, когда две соответствующие переменные другого типа имеют единичные значения: $x_{ij} = x_i^k \cap x_j^k$. В нотации линейного программирования это условие трансформируется в следующую систему двусторонних ограничений:

$$0 \le x_i^k + x_j^k - 2x_{ij} \le 1, \quad i = \overline{1, n-1}, \quad j = \overline{i+1, n}, \quad k = \overline{1, l} . \qquad (5)$$

Такой выбор индексов объектов обусловлен тем, что матрица расстояний избыточна. Мы выбираем только один из треугольников матрицы и исключаем из рассмотрения диагональные элементы.

При необходимости могут быть наложены ограничения на размеры кластеров. Сделать это можно с помощью выражений следующего вида:

$$\sum_{i=1}^{n} x_i^k \le M^k, \quad k = \overline{1, l} . \qquad (6)$$

Здесь M^k – предельная мощность кластера с соответствующим индексом.

Специфическая особенность задач или моделей линейного программирования в том, что при некоторых исходных данных часть ограничений может оказаться избыточной. Правда, это может никак не сказываться на получаемых решениях. Как правило, для избыточных ограничений оценки имеют нулевые значения.

Хуже обстоят дела, если функциональных ограничений недостаточно в том смысле, что система ограничений не адекватна моделируемой ситуации. В этом случае, на основе углубленного анализа решения требуется добавить новые ограничения в систему.

Система параметрических ограничений, накладываемых на значения переменных задачи, включает в свой состав ограничения двух типов, так как в ней наряду с целочисленными (логическими) переменными могут быть и переменные с вещественными значениями:

$$x_i^k \in \{0,1\}, \quad i=\overline{1,n}, \quad k=\overline{1,l}\,;$$

$$x_{ij} \in \{0,1\}, \quad i=\overline{1,n-1}, \quad j=\overline{i+1,n}\,;$$

$$y_p \ge 0, \quad p=\overline{1,l}\,.$$

Отметим, что компоненты целевой функции, представленные суммами для отдельных кластеров, могут размещаться и вычисляться в составе системы ограничений. Это не влияет на конечный результат и определяется предпочтениями, удобством и, может быть, некоторыми другими соображениями или факторами.

В дополнение к представленным функциональным ограничениям, в зависимости от выбранной целевой функции, могут быть добавлены ограничения на радиус выделяемых кластеров, на минимальные межкластерные расстояния, количество выделяемых кластеров и другие. Из изложенного уже понятно, как их можно представить в линейной модели.

Структура табличной (матричной) формы представления модели

Существенную помощь в уяснении сути задачи, а также проверки правильности построенной модели оказывает матричная или табличная форма. Ее широко использовали основатели дисциплины [5]. К сожалению, громоздкость предложенной линейной модели кластерного анализа не позволяет поместить ее в формат статьи в общем виде. Тем не менее определенное представление о связях блоков модели может дать таблица 1.

Таблица 1 – Табличная форма представления модели (простой пример)

	Коэффициенты целевой функции	Тип огр.	CType	B
C	0 0 0 0 0 0 0 0 0 0 1 1 3 4 1 3 4 3 4 1 1 1 3 4 1 3 4 3 4 1			
	Блок ограничений типа (4)			
1	**1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	=	S	1
2	0 0 **1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	=	S	1
3	0 0 0 0 **1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	=	S	1
4	0 0 0 0 0 0 **1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	=	S	1
5	0 0 0 0 0 0 0 0 **1 1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	=	S	1
	Блок ограничений типа (5)			
6	**1** 0 **1** 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
7	**1** 0 0 0 **1** 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
8	**1** 0 0 0 0 0 **1** 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
9	**1** 0 0 0 0 0 0 0 **1** 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
10	0 0 **1** 0 **1** 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
11	0 0 **1** 0 0 0 **1** 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
12	0 0 **1** 0 0 0 0 0 **1** 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
13	0 0 0 0 **1** 0 **1** 0 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
14	0 0 0 0 **1** 0 0 0 **1** 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0 0	≥	L	0
15	0 0 0 0 0 0 **1** 0 **1** 0 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0 0	≥	L	0
16	0 **1** 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0 0	≥	L	0
17	0 **1** 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **-2** 0 0 0 0 0 0 0 0	≥	L	0

18	0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0	≥	L	0
19	0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0	≥	L	0
20	0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0	≥	L	0
21	0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0	≥	L	0
22	0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0	≥	L	0
23	0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0	≥	L	0
24	0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0	≥	L	0
25	0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2	≥	L	0
26	1 0 1 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
27	1 0 0 0 1 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
28	1 0 0 0 0 0 1 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
29	1 0 0 0 0 0 0 0 1 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
30	0 0 1 0 1 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
31	0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
32	0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
33	0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
34	0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0 0	≤	U	1
35	0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0 0	≤	U	1
36	0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0 0	≤	U	1
37	0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0 0	≤	U	1
38	0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0 0	≤	U	1
39	0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0 0	≤	U	1
40	0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0 0	≤	U	1
41	0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0 0	≤	U	1
42	0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0 0	≤	U	1
43	0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0 0	≤	U	1
44	0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2 0	≤	U	1
45	0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -2	≤	U	1
	Блок ограничений типа (6)			
46	1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	5
47	0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	≤	U	5

Здесь расписан простейший демонстрационный пример, включающий пять объектов, образующих два кластера. Размерность признакового пространства равна двум. Модель приведена полностью, с сохранением нулевых значений. Цель – возможность копирования данных для вставки в скрипт подходящего математического пакета. Это для тех, кто пожелает проверить результаты (подход, все-таки нетрадиционный) или решит освоить этот материал на практике.

Первые десять элементов целевой функции имеют нулевые значения, так как соответствуют переменным, не участвующим в вычислении значений целевой функции. Это – переменные принадлежности объектов кластерам. В остальной части целевой функции располагаются два одинаковых фрагмента, содержащие в развернутом виде значения мер расстояния между объектами. Это – элементы верхней треугольной части матрицы расстояний (без диагональных элементов), развернутые в строку. Для того, чтобы таблица не вышла за пределы страницы, значения пришлось округлить до целых.

Ниже в таблице располагаются блоки коэффициентов системы функциональных ограничений. Первые десять столбцов и пять строк описывают условия (4) связи объектов с кластерами. Второй блок, включающий строки с шестой по двадцать пятую, а затем с двадцать шестой по сорок пятую, соответствует соотношениям (5) и связывает переменные первого блока с элементами матрицы расстояний, что находятся в целевой функции. В последнем блоке вычисляются мощности кластеров. Но в данном примере эти две строки фактически не задействованы, так как мощности кластеров, указанные в правых частях системы ограничений, соответствуют мощности всего множества объектов. Чтобы задействовать ограничения последнего блока, требуется уменьшить значения правых частей.

Для типов ограничений, наряду с обычными математическими, заданы также и буквенные обозначения пакета GNU Ovtave. В последнем столбце размещены правые части системы ограничений.

На рисунке 1 дано графическое представление исходных данных и искомых результатов. Ввиду того, что обе координаты в признаковом пространстве выражены в одних и тех же единицах, стандартизация данных не производилась.

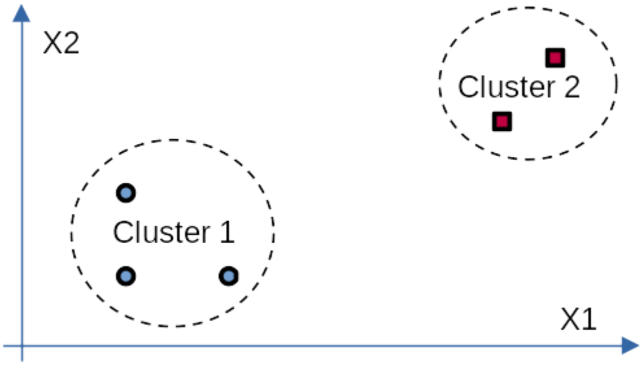

Рисунок 1 – Графическое представление данных демонстрационного примера

Результат расчетов, проведенных с помощью вычислительного пакета GNU Octave, таков: выделено два кластера: в первом три объекта, во втором – два; сумма внутрикластерных расстояний равна четырем.

Сравнение с методом кластеризации К-средних

Выбор метода кластеризации для сравнительного анализа очевиден. Целевая функция, заложенная явно или неявно в этот метод, должна быть известна и должна соответствовать нашей формализации (2). В противном случае сравнение некорректно. Из таких методов был выбран метод К-средних. Известно, что при использовании Эвклидовой метрики его критерий минимизации расстояний до центра кластеров аналогичен критерию (2). К тому же этот метод требует задания количества кластеров. В нашей формулировке задачи ЦЛП мы зафиксировали количество кластеров.

Исходные данные для сравнительных расчетов представлены таблицей 2. Для кластеризации взято двадцать объектов в двумерном признаковом пространстве.

Таблица 2 – Исходные данные

№ объекта	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Координата 1	5	2	7	9	1	9	6	6	6	3	26	23	26	29	27	24	26	27	23	23
Координата 2	1	2	4	9	7	8	8	4	5	8	22	24	23	27	25	21	22	25	26	29

Результаты расчетов представлены таблицей 3. На пересечениях строк и столбцов располагаются номера кластеров, к которым отнесены объекты. Их номера в шапке таблицы. Сбоку – номера вариантов кластеризации. Первые семь строк – это результаты, полученные с помощью метода К-средних. В последней строке под номером восемь – решение задачи ЦЛП.

Таблица 3 – Результаты кластеризации

		О б ъ е к т ы																				Значения целевой функции
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Вариант	1	2	4	2	3	4	3	3	2	2	4	1	1	1	1	1	1	1	1	1	1	22.268
	2	3	3	3	4	1	4	1	3	3	1	2	2	2	2	2	2	2	2	2	2	23.086
	3	1	1	1	1	1	1	1	1	1	1	2	2	2	4	2	2	2	2	3	3	29.017
	4	2	2	2	4	4	4	4	2	2	4	1	3	1	3	3	1	1	3	3	3	14.862
	5	1	1	3	3	1	3	3	3	3	3	4	4	4	4	4	4	4	4	2	2	19.598
	6	1	1	3	3	1	3	3	3	3	3	4	2	4	4	4	2	4	4	2	2	16.657
	7	1	1	1	3	3	3	3	1	1	3	2	2	2	4	2	2	2	2	2	4	17.147

ы	8	1	1	1	4	4	4	4	1	1	4	3	3	3	2	2	3	3	2	2	2	14.023

В последнем столбце располагаются значения целевой функции. Для обеспечения сопоставимости для решений, полученных методом К-средних, были рассчитаны суммарные внутрикластерные расстояния по формуле (2).

Результаты оказались неожиданными. Из семи решений лишь одно приблизилось к полученному нашим методом, но и оно не является глобальным минимумом. Кроме того, часть решений вовсе неудовлетворительна. Например, значение целевой функции для решения, полученного в третьем варианте, вдвое превышает значение глобального минимума.

Заключение

В работе были рассмотрены основы кластерного анализа и его проблемы. Для решения некоторых из них был предложен подход по трансформации модели кластерного анализа в модель задачи целочисленного линейного программирования, часть переменных которой может иметь вещественные значения. Предложены варианты построения целевой функции и системы функциональных ограничений. Показана структура табличной формы представления модели.

Одним из недостатков моделирования кластерных структур средствами линейного программирования является громоздкость получаемых моделей. Не исключено, что очень большие множества объектов могут порождать модели, размерности которых будут выходить за пределы возможностей пакетов линейного программирования.

К несомненным достоинствам предложенного подхода следует отнести наличие явно задаваемой целевой функции, что существенно облегчает процессы интерпретации решения, а также то, что получаемое решение соответствует глобальному экстремуму критериального показателя в рамках установленных ограничений. Это позволяет надеяться на возможности расширения областей успешного применения методов кластерного анализа, например, в управлении в организационно-экономическими системами [9], разработок в области техники и технологий [10], анализе транспортных проблем [11] и многих других.

Литература

1. Michael R. Anderberg. Cluster Anflysis for Applications. — New York, SanFracisco, London: Aadevic Press, 1973. — 360 p.
2. Дюран Б., Оделл П. Кластерный анализ. — М.: Статистика, 1977. — 128 с.
3. Мандель И.Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.
4. Айвазян С. А. и др. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1989. — 607 с.
5. George B. Dantzig. Linear Programming and Extensions // R-366-PR. — Santa Monica. Caligornia: RAND Corp., August 1963. — 521 p.
6. G.B. Dantzig. Notes on linear programming: part XXXV — Discrete-variable extremum problems // RM-1832. — Santa Monica. Caligornia: RAND Corp., 1957. — 29 p.
7. Данциг Дж. Линейное программирование, его применение и обобщения. — М.: Прогресс, 1966. — 600 с.
8. Яр-Мухамедов И.Г. Исследования, разработки и образование // Философские и методологические проблемы исследования российского общества: сборник трудов VII Международной научно-практической конференции (24 ноября 2023 года, г. Москва, Российская Федерация) / Под общ. ред. Г.В. Бариновой, С.Н. Климова, П.И. Вермишовой; М-во транспорта Рос. Федерации, Рос. ун-т транспорта (МИИТ), Рос. открытая акад. транспорта. — М.: РУТ (МИИТ): РОАТ, 2023. — С. 233—238.

9. Аскарбеков Р.Н., Орускулова Т.Р., Раззаков М.И. Информационные технологии в делопроизводстве учебного заведения / Р.Н. Аскарбеков, Т.Р. Орускулова, М.И. Раззаков // Проблемы автоматики и управления. – Бишкек: Илим, 2022, No 3(45). — С. 101—109.

10. Обозов А. Дж., Насирдинова С.М., Салбаев А.Н. Структурная схема управления БГУ с учетом графика нагрузки потребителя / А. Дж. Обозов, С.М. Насирдинова, А.Н. Салбаев // Проблемы автоматики и управления. — Бишкек: Илим, 2021, No1 (40). — С. 54—60.

11. Адиева Г.М. Анализ общественного транспорта города Ош / Г.М. Адиева // Проблемы автоматики и управления. – Бишкек: Илим, 2021, No1 (40). — С. 68—74.