# КОМПЬЮТЕРНЫЕ НАУКИ

УДК 004.93

## КЛАСТЕР-АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ РЕШЕНИЙ

*ГОФМАН Е.А., ОЛЕЙНИК А.А., СУББОТИН С.А.*

Рассматривается решение задачи кластерного анализа с использованием деревьев решений. Разрабатывается метод кластерного анализа, позволяющий выполнять разбиение пространства экземпляров на кластеры, при использовании которого отсутствует необходимость задания информации о количестве кластеров и их форме, что существенно расширяет возможность его применения на практике.

### Введение

При распознавании образов, классификации веб-контента, прогнозировании актуальной является задача кластерного анализа, которая заключается в разделении входной выборки данных на кластеры – компактные, непересекающиеся области (таксоны) в пространстве признаков. Известны различные методы кластерного анализа [1, 2]. Основным недостатком их является необходимость предварительного задания входных настраиваемых параметров (например, количество кластеров, которые должны быть выделены). Это усложняет их применение при обработке данных в реальных ситуациях, когда не имеется достаточной информации об исследуемом объекте, процессе или системе.

Поэтому актуальной является разработка новых методов кластеризации, свободных от указанных недостатков и обеспечивающих необходимую точность получаемых решений. Основным критерием, которому должны удовлетворять методы, применяемые для решения данной задачи, является возможность разделения пространства экземпляров на области со сходными характеристиками. К таким методам относится поиск на основе деревьев решений, которые из-за своей структуры выполняют разбиение пространства решений на области в зависимости от значений входных переменных [3–5]. В связи с этим в данной работе предлагается решать задачу кластерного анализа на основе построения деревьев решений.

Существуют различные методы идентификации деревьев решений (ID3, CART, CHAID, QUEST, C5.0). Однако они не учитывают особенностей решаемой задачи кластерного анализа, связанной с выделением таксонов, состоящих из объектов с наиболее сходными характеристиками [3–6].

Цель данного исследования – разработать метод кластерного анализа, основанный на построении деревьев решений, который позволит выполнять разбиение на кластеры путём введения равномерно распределённых точек пространства поиска и сократит требования к вычислительным ресурсам при выполнении кластерного анализа.

Для достижения поставленной цели необходимо выполнить следующие задачи:

– дать обзор существующих методов кластерного анализа и выявить их преимущества и недостатки;

– изучить основные понятия, принципы и особенности деревьев решений;

– модифицировать рассматриваемый метод в соответствии со спецификой решаемой задачи;

– сравнить разработанный подход с существующими методами кластерного анализа путём проведения экспериментов и анализа полученных результатов.

### 1. Постановка задачи

Пусть задано множество объектов O, каждый из которых характеризуется множеством значений признаков X. Тогда задача кластерного анализа заключается в том, чтобы на основании значений признаков X, разбить множество объектов O на m (m – целое) кластеров (подмножеств) $C_1$, $C_2$, …, $C_m$ так, чтобы каждый объект $O_i$ принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам, были разнородными.

### 2. Кластерный анализ

Кластерный анализ заключается в разбиении данных на группы схожих объектов. Каждая группа, называемая кластером, состоит из объектов, которые похожи между собой, но при этом отличаются от объектов других групп.

Существует несколько видов методов кластерного анализа, отличающихся между собой допущениями о форме кластеров, видом результирующего разбиения и параметрами, которые должны быть установлены (например, количеством кластеров).

Исключающая кластеризация: данные группируются путём исключения единиц данных. Если определённый объект принадлежит одному кластеру, то он не может быть включён в другой (к таким методам относится, например, метод k-средних). Основными недостатками такого подхода являются:

– необходимость указания количества кластеров, на которые необходимо разбить входную выборку;

– выполнение поиска кластеров только заданной формы.

Перекрывающаяся кластеризация: данные могут входить в два или более кластеров в зависимости от значения функции принадлежности. К таким методам

относится метод нечётких С-средних. При использовании их также необходимо задавать количество кластеров.

Иерархическая кластеризация: в начале кластеризации каждый объект рассматривается как отдельный кластер, после чего два ближайших объединяются в один и так далее. Метод заканчивает свою работу, когда все данные объединены в один кластер либо если выполнилось условие окончания работы. Основным недостатком такого подхода является существенная вычислительная сложность, что особенно заметно при обработке многомерных выборок большого объёма.

Вероятностная кластеризация имеет две разновидности:

– методы, основанные на смеси многомерных нормальных распределений;

– методы интеллектуальной оптимизации, построенные на моделировании коллективного интеллекта общественных живых существ.

Поскольку данный подход основан на вероятностном подходе, то есть возможность несходимости к оптимальному решению.

Как видно из приведенной классификации, каждый из рассмотренных методов обладает определёнными недостатками, основными из которых являются: необходимость задания количества формируемых кластеров, допущение о форме кластеров, большая вычислительная сложность. В связи с этим можно сделать вывод о том, что применение деревьев решений для кластерного анализа является перспективным, однако данную технику необходимо применять с учётом особенностей решаемой задачи кластерного анализа.

### 3. Деревья решений

Деревья решений представляют собой нисходящую систему, основанную на подходе “разделяй и властвуй”. Основная цель ее – разделение дерева на взаимно непересекающиеся подмножества [3, 5]. Каждое подмножество представляет собой подзадачу классификации.

Дерево решений описывает процедуру принятия решения о принадлежности определённого экземпляра к тому или иному классу.

Дерево решений является древовидной структурой, состоящей из внутренних и внешних узлов, связанных рёбрами [6]. Внутренние узлы – модули, принимающие решение, – рассчитывают значение функции решения, на основании чего определяют дочерний узел, который будет посещён далее. Внешние узлы (также называемые конечными узлами), напротив, не имеют дочерних узлов и описывают метку класса или значение, характеризующее входные данные. В общем случае, деревья решений используются следующим образом. Вначале передаются данные (обычно это вектор значений входных переменных) на корневой узел дерева решений. В зависимости от полученного значения функции решения, используемой во внутреннем узле, происходит переход к одному из дочерних узлов. Такие переходы продолжаются до тех пор, пока не будет посещён конечный узел, описывающий либо метку класса, либо значение, связанное со входным вектором значений признаков.

### 4. Кластеризация на основе построений деревьев решений

В предлагаемом методе кластеризации данных на основе построений деревьев решений в процессе синтеза деревьев использован традиционный подход, позволяющий разделить пространство поиска на несколько разных классов на основании функции приоритетности. Однако, поскольку при решении задачи кластеризации не заданы классы экземпляров, то предлагается вводить несуществующие равномерно распределённые экземпляры для проведения кластерного анализа. За счёт введения таких экземпляров можно условно разбить входную выборку, как минимум, на два класса: существующие экземпляры и несуществующие, за счёт чего можно выполнять классификацию с использованием деревьев решений. Такой подход позволяет выделить те области, которые представляют собой кластеры, поскольку в них больше реальных экземпляров, чем искусственно добавленных. Далее представлены основные особенности данного метода.

При построении дерева решений для каждого признака из n-мерного пространства (n – количество признаков, характеризующих обучающую выборку) метод рассчитывает индекс Джини для разделения дерева решений, используемый в качестве критерия приоритетности альтернативных возможных вариантов разбиения по признаку, т.е. текущий узел дерева разбивается по признаку, по которому получено лучшее (наименьшее) значение этого индекса.

В каждом узле происходит разбиение по определённому признаку на левую и правую ветви (области ограничены предыдущими разделениями). Таким образом, данный этап предполагает выполнение следующей последовательности действий:

– установить счётчик признаков в единицу: $i = 1$;

– для каждого конкретного значения признака $X_i$ рассчитать индекс Джини;

– установить: $i = i + 1$;

– если $i \leq n$, выполнить переход к расчету индекса Джини для следующего признака;

– сохранить лучшее разбиение для текущего узла;

– выполнить разбиение для левого потомка;

– выполнить разбиение для правого потомка.

Принципиальная особенность такого этапа вычисления состоит в том, что индекс Джини для разделения вычисляется для значений признаков и некоторых равномерно распределённых искусственно добавлен-

ных K точек. Каждое значение признака $X_i$ рассматривается как возможное разбиение, поэтому индекс Джини рассчитывается для каждого значения.

Пусть имеется множество M с соответствующей мощностью |M|. Пусть дополнительно к этому множеству добавляется множество K равномерно распределённых точек мощности |K| = |M| (количество дополнительных точек в дальнейшем наследуется от родительского узла). Каждое значение $x \in M$ разбивает множество на две области.

Пусть в левой области относительно текущей точки $x \in M$ находятся области из $k_{x-}$ и $m_{x-}$ точек, значения которых меньше заданного значения, в правой области находятся $m_{x+} = |M| - m_{x-}$ и $k_{x+} = |K| - k_{x-}$, соответственно. Тогда рассчитать $k_{x+}$ и $k_{x-}$ можно следующим образом:

$$k_{x-} = |M| - k_{x+} = \frac{|K|(x - \min(M))}{\max(M) - \min(M)},$$

$$k_{x+} = |M| - k_{x-} = \frac{|K|(\max(M) - x)}{\max(M) - \min(M)},$$

где x – конкретное значение признака; min(M) – минимальное значение из M, max(M) – максимальное значение в M. Такая формула означает, что если в пределах между min(M) и max(M) находится |K| равномерно распределённых точек, тогда в интервале между min(M) и текущим значением x находится $n_{x-}$ точек.

В общем случае индекс Джини для разделения по *x* можно рассчитать следующим образом:

$$g_x = \frac{k_{x-} + m_{x-}}{|K| + |M|} g_{x-} + \frac{k_{x+} + m_{x+}}{|K| + |M|} g_{x+},$$

где индексы Джини для подмножеств x– и x+ рассчитываются так:

$$g_{x*} = 1 - \frac{k_{x*}^2 + m_{x*}^2}{\left(k_{x*} + m_{x*}\right)^2},$$

где * обозначает соответствующее подмножество (+ или –).

После того, как получено лучшее разбиение, оно переносится в текущий узел, его правый и левый потомки наследуют множество K, включающее $n_{x-}$ и $n_{x+}$ точек, соответственно.

Вычисление разбиений продолжается до тех пор, пока:

– текущий узел содержит экземпляры в количестве |M|, большем заданного минимального значения, т.е. данный параметр является единственным настраиваемым входным параметром для предлагаемого метода;

– текущее множество данных содержит группы как минимум с двумя точками (при этом точки с одинаковыми значениями группируются на начальном этапе).

Таким образом, разбиение узла должно продолжаться при выполнении хотя бы одного из этих условий. В противном случае разбиение на данной ветке должно завершиться.

Из сказанного следует, что основной особенностью предложенного метода является введение дополнительных равномерно распределённых экземпляров, что позволяет выполнять классификацию, как минимум, для двух классов экземпляров. При этом основное преимущество предложенного метода состоит в том, что нет необходимости задавать информацию о количестве кластеров, их форме и т.п., что существенно расширяет возможность применения разработанного метода кластерного анализа на основе построения деревьев решений.

## 5. Эксперименты и результаты

Предложенный метод кластерного анализа на основе построения деревьев решений был программно реализован в среде пакета Matlab 7.0.

При помощи разработанного программного обеспечения и встроенных средств пакета Matlab 7.0 проводились эксперименты, которые заключались в разбиении на кластеры искусственно сформированных выборок с помощью разработанного метода, а также методов кластеризации: K-средних и агломеративного иерархического.

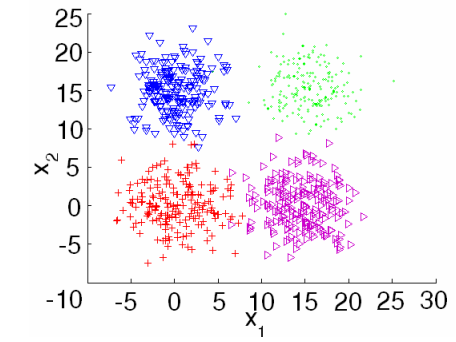
Выборки формировались случайным образом на основе нормального распределения с различными математическими ожиданиями и дисперсиями. Было сформировано четыре двумерные выборки, отличающиеся между собой степенью пересеченья кластеров. Параметры распределений, на основании которых формировались выборки, приведены в табл. 1.

Таблица 1. Параметры распределений выборок

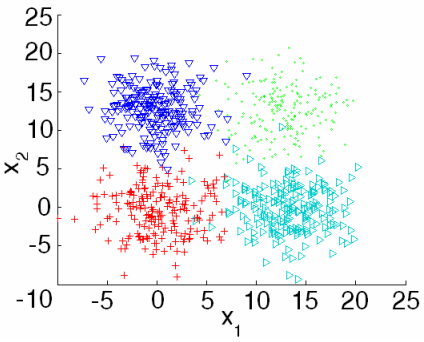
| Выборка | Кластер | $x_1$ | | $x_2$ | |
|---|---|---|---|---|---|
| | | M(X) | D(X) | M(X) | D(X) |
| 1 | 1 | 0 | 3 | 0 | 3 |
| | 2 | 15 | 3 | 15 | 3 |
| | 3 | 15 | 3 | 0 | 3 |
| | 4 | 0 | 3 | 15 | 3 |
| 2 | 1 | 0 | 3 | 0 | 3 |
| | 2 | 13 | 3 | 13 | 3 |
| | 3 | 13 | 3 | 0 | 3 |
| | 4 | 0 | 3 | 13 | 3 |
| 3 | 1 | 0 | 3 | 0 | 3 |
| | 2 | 0 | 3 | 25 | 4 |
| | 3 | 16 | 3 | 25 | 4 |
| | 4 | 25 | 5 | 0 | 3 |
| 4 | 1 | 0 | 3 | 0 | 3 |
| | 2 | 0 | 3 | 12 | 3 |
| | 3 | 12 | 3 | 0 | 3 |
| | 4 | 12 | 3 | 12 | 3 |

Распределение выборок 1–4 в пространстве переменных представлено на рисунке 1 а)–г). Каждая выборка состояла из четырёх кластеров, каждый из которых, в

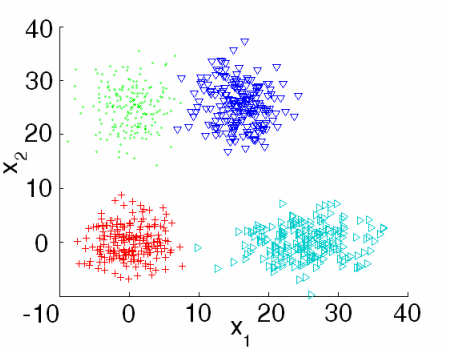
свою очередь, состоял из 200 экземпляров, характеризующихся двумя признаками. Как видно из табл. 1 и рисунка вторая и четвёртая выборки характеризуются большим пересечением кластеров по сравнению с первой и третьей.



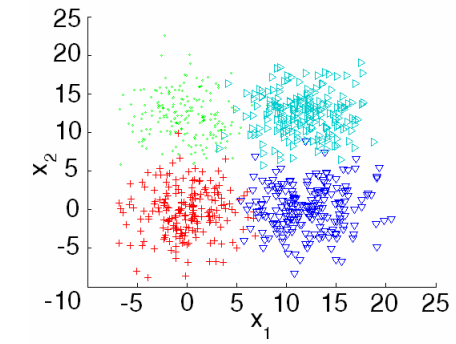

а




б




в




г

Графическое представление первой (а), второй (б), третьей (в) и четвёртой (г) выборок

В качестве критерия сравнения результатов работы исследуемых методов кластеризации использовалась ошибка классификации:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N} res_i ,$$

где $res_i = 1$, если $cluster_i^* \neq cluster_i$, в противном случае – $res_i = 0$; $cluster_i^*$ – номер кластера, к которому отнесён i-й объект при помощи заданного метода кластерного анализа, $cluster_i$ – номер кластера, к которому относится i-й объект в заданной обучающей выборке.

Результаты работы традиционных методов кластеризации и предложенного метода представлены в табл. 2.

Таблица 2. Результаты работы методов кластерногоанализа

| Метод | Значение ошибки | | | |
|---|---|---|---|---|
| | Выборка 1 | Выборка 2 | Выборка 3 | Выборка 4 |
| Метод K-средних | 0,0113 | 0,0288 | 0,0050 | 0,0288 |
| Иерархический агломеративный метод | 0,0138 | 0,0325 | 0,0075 | 0,0300 |
| Метод кластерного анализа на основе деревьев решений | 0,0043 | 0,0215 | 0,0041 | 0,0219 |

Исходя из результатов экспериментов, представленных в табл. 2, можно видеть, что предложенный метод характеризуется меньшей ошибкой классификации по сравнению с методами K-средних и иерархическим агломеративным. При этом наибольшая ошибка классификации наблюдалась для всех методов при анализе второй и четвёртой выборок, для которых характерна ощутимая пересекаемость кластеров.

Важно также отметить, что для работы предложенного метода не надо было задавать количество выходных кластеров, в отличие от рассматриваемых традиционных методов. При этом количество кластеров, на которое разбивал входную выборку разработанный метод, было правильным и составило четыре кластера для всех выборок.

**Выводы**

В статье решена актуальная задача кластеризации на основе использования деревьев решений.

*Научная новизна* исследования заключается в том, что разработан метод кластерного анализа, основанный на построении деревьев решений, который позволяет выполнять разбиение на кластеры путём введения равномерно распределённых точек пространства поиска и сокращает требования к вычислительным ресурсам при выполнении кластерного анализа. Кроме того, при использовании предложенного метода отсутствует необходимость задания информации о количестве кластеров и их форме, что существенно расширяет возможность его применения на практике.

**Литература: 1.** *Berkhin P.* Survey of clustering data mining techniques / P. Berkhin. San Jose : Accrue Software, 2002. 59 p. **2.** *Субботін С. О.* Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник / С. О. Субботін. Запоріжжя: ЗНТУ, 2008. 341 с. **3.** *Quinlan J. R.* Induction of decision trees / J. R. Quinlan / / Machine Learning. 1986. № 1. P. 81–106. **4.** *Rokach L.* Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. London : World Scientific Publishing Co, 2008. 264 p. **5.** *Classification* and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. California : Wadsworth & Brooks, 1984. 368 p. **6.** *Quinlan J. R.* C.4.5: Programs for machine learning / J. R. Quinlan. San Mateo : Morgan Kaufmann, 1993. 312 p.

Поступила в редколлегию 02.06.2011

**Рецензент:** д-р техн. наук, проф. Бодянский Е.В.

**Гофман Евгений Александрович**, аспирант кафедры программных средств Запорожского национального технического университета. Научные интересы: деревья решений. Адрес: Украина, 69063, Запорожье, ул. Жуковского, 64.

**Олейник Алексей Александрович**, канд. техн. наук, доцент кафедры программных средств Запорожского национального технического университета. Научные интересы: интеллектуальные системы поддержки принятия решений. Адрес: Украина, 69063, Запорожье, ул. Жуковского, 64.

**Субботин Сергей Александрович**, канд. техн. наук, доцент кафедры программных средств Запорожского национального технического университета. Научные интересы: нейронные сети, нечеткая логика, интеллектуальные системы поддержки принятия решений. Адрес: Украина, 69063, Запорожье, ул. Жуковского, 64.