УДК 004.415:658

Е. С. ЯШИНА, М. А. ЩЕРБАК

Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина

ИСПОЛЬЗОВАНИЕ МЕТОДА КЛАСТЕРИЗАЦИИ В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ

Статья посвящена применению методов кластеризации данных при разработке информационно-аналитических систем. Проводится обзорный анализ существующих методов кластеризации. Предлагаются алгоритмы, построенные на основе модифицированных статистического (метод k-средних) и иерархического методов кластеризации. Разработана архитектура информационно-аналитической системы в виде web-приложения, использующего предложенные методы и алгоритмы для обработки данных, полученных из различных источников. На основе этих алгоритмов произведена разработка приложения для кластеризации наборов объектов.

Ключевые слова: *анализ, кластеризация, кластерный анализ, иерархическая кластеризация, метод k-средних.*

Введение

Аналитическая обработка данных становится одним из важнейших направлений развития информационных систем. Однако применение существующих методов и математических моделей для работы с большими объемами данных требует высокого уровня квалификации, что делает необходимым создание новых алгоритмов, программного обеспечения и инструментов средств, позволяющих эффективно использовать существующие методы работы с данными [1, 2].

Особый интерес к методам анализа данных возник в связи с развитием средств сбора и хранения данных, позволившим накапливать большие объемы информации. Перед специалистами из разных областей знаний встал вопрос об обработке собираемых данных и превращение их в знания.

Среди методов интеллектуального анализа данных особое место занимают классификация и кластеризация. Классификация, при известной заранее группировке данных на подмножества (классы), устанавливает закономерность, по которой данные группируются именно таким образом. Кластеризация же, основываясь на установленном отношении схожести элементов, устанавливает подмножества (кластеры), в которые группируются входные данные [3, 4]. В дальнейшем может быть выполнен углублённый анализ каждого из выделенных подмножеств с учётом сходства и различия объектов, установленных в ходе кластеризации.

Задача кластеризации состоит в разбиении исходной выборки, представленной матрицей «объект-свойство», на группы (кластеры) таким образом, чтобы внутри каждого кластера находились похожие друг на друга объекты, а объекты разных кластеров существенно различались. Если данные представить как точки в многомерном признаковом пространстве, то задача кластеризации сводится к отысканию компактных сгустков точек [5].

Рассмотрим подробнее задачу кластеризации данных.

Основные задачи кластерного анализа

Кластерный анализ (Data clustering) – задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов. Сходство объектов определяется на основе метрики, выбранной в соответствии с критерием кластеризации.

Таким образом, кластерный анализ (англ. cluster analysis) представляет собой многомерную статистическую процедуру, выполняющую сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

Можно выделить следующие основные задачи кластерного анализа [6]:

- разработка типологии или классификации;
- исследование полезных концептуальных схем группирования объектов;
- порождение гипотез на основе исследования данных;
- проверка гипотез для определения, действи-

тельно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Применение кластерного анализа предполагает следующие этапы:

- отбор выборки для кластеризации;
- определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства;
- вычисление значений той или иной меры сходства (или различия) между объектами;
- применение метода кластерного анализа для создания групп сходных объектов;
- проверка достоверности результатов кластерного решения.

При применении кластеризации, как и других методов статистического исследования, данные должны обладать однородностью и полнотой. Однородность требует, чтобы все кластеризуемые сущности были одной природы и описывались сходным набором характеристик. Полнота выборки обозначает это наличие всех элементов генеральной совокупности в основе выборки, обладающих существенными особенностями и характеристиками. При невыполнении этих требований выборка окажется нерепрезентативной, а результаты кластеризации могут быть некорректными.

Спектр применений кластерного анализа очень широк: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, социологии и других дисциплинах. Однако универсальность применения привела к появлению большого количества, методов и подходов, затрудняющих использование и непротиворечивую интерпретацию кластерного анализа [7 – 9].

Практическое применение кластеризации и других методов анализа больших объёмов данных требует разработки и применения специализированного программного обеспечения. Целью данной работы является проектирование web-приложения кластеризации данных, полученных из различных источников. Это потребует исследования и модификации существующих методов кластеризации и разработки алгоритмов применения методов кластеризации для анализа данных. Разрабатываемое web-приложение может быть частью архитектуры информационно-аналитической системы.

Формальная постановка задачи кластеризации

Пусть X – множество объектов, Y – множество номеров (имен меток) кластеров. Задана функция расстояния между объектами $\rho(x,x')$. Имеется конечная обучающая выборка объектов

$$Xm = \{x_1, \ldots, x_m\} \in X.$$

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X_m$ приписывается номер кластера y_i.

Алгоритм кластеризации – это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации [10 – 12].

Кластеризация отличается от классификации тем, что метки исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

- не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию. Все они могут давать разные результаты. Следовательно, для определения качества кластеризации требуется эксперт предметной области, который бы мог оценить осмысленность выделения кластеров;
- число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. Это справедливо только для методов дискриминации, так как в методах кластеризации выделение кластеров идёт за счёт формализованного подхода на основе мер близости;
- результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Указанная неоднозначность требует гибкой настройки инструментов обработки данных, что необходимо учитывать при разработке соответствующего программного обеспечения.

Методы кластеризации

Условно все методы классификации можно разделить на три группы. Отметим наиболее известные методы каждой группы:

1. Графовые методы кластеризации:
   - алгоритм кратчайшего незамкнутого пути;

- алгоритм FOREL.

2. Статистические методы кластеризации:

- EM-алгоритм;
- метод k-средних.

3. Иерархическая кластеризация (таксономия):

- агломеративная иерархическая кластеризация;
- дендрограмма и свойство монотонности;
- свойства сжатия, растяжения и редуктивности.

Рассмотрим приведенные выше методы более подробно.

Алгоритм кратчайшего незамкнутого пути. Данный алгоритм строит граф из $\ell-1$ рёбер так, чтобы они соединяли все ℓ точек и обладали минимальной суммарной длиной [1].

Число кластеров K в этом алгоритме задаётся как входной параметр. Его можно также определять графически, если упорядочить все расстояния, образующие каркас, в порядке убывания и отложить их на графике. Резкий скачок вниз где-то на начальном (левом) участке графика покажет количество наиболее чётко выделяемых кластеров.

Алгоритм FOREL. Пусть задана некоторая точка $x_0 \in X$ и параметр R. Выделяются все точки выборки $x_i \in X_l$, попадающие внутрь сферы $\rho(x_i, x_0) \le R$, и точка x_0 переносится в центр тяжести выделенных точек. Эта процедура повторяется до тех пор, пока состав выделенных точек, а значит и положение центра, не перестанет меняться [4, 5].

Доказано, что эта процедура сходится за конечное число шагов. При этом сфера перемещается в место локального сгущения точек. Центр сферы x_0 не является объектом выборки, потому и называется формальным элементом.

Для вычисления центра необходимо, чтобы множество объектов X было не только метрическим, но и линейным векторным пространством. Это требование естественным образом выполняется, когда объекты описываются числовыми признаками.

EM-алгоритм. Предполагается, что данные в каждом кластере подчиняются нормальному распределению. С учетом этого предположения можно определить параметры – математическое ожидание и дисперсию, которые соответствуют закону распределения элементов в кластере, наилучшим образом «подходящему» к наблюдаемым данным [1, 4].

Объекты выборки X_l появляются случайно и независимо согласно вероятностному распределению, представляющему собой смесь распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \quad \sum_{y \in Y} w_y = 1,$$

где $p_y(x)$ – функция плотности распределения кластера y, w_y – неизвестная априорная вероятность появления объектов из кластера y.

Объекты описываются n числовыми признаками $f_1(x), \ldots, f_n(x)$, $X = R^n$. Каждый кластер $y \in Y$ описывается n-мерной гауссовской плотностью $p_y(x) = N(x; \mu_y, \Sigma_y)$ с центром $\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ и диагональной ковариационной матрицей

$$\Sigma_y = \mathrm{diag}(\sigma^2_{y1}, \ldots, \sigma^2_{yn}).$$

Метод k-средних. Данный метод можно назвать модифицированным симбиозом алгоритма FOREL и EM-алгоритма [1, 3].

Метод k-средних является упрощением EM-алгоритма. Главное отличие в том, что в EM-алгоритме каждый объект x_i распределяется по всем кластерам с вероятностями $g_{iy} = P\{y_i = y\}$. В алгоритме k-средних (k-means) каждый объект жёстко приписывается только к одному кластеру. Второе отличие в том, что в k-means форма кластеров не настраивается.

Данный метод похож на базовую процедуру поиска центра кластера в алгоритме FOREL. Отличие в том, что в FOREL кластер – это шар заданного радиуса R, тогда как в k-means объекты относятся к кластерам по принципу ближайшего соседа.

Иерархическая кластеризация. Алгоритмы кластеризации, относящиеся к этой группе, называют также алгоритмами таксономии. Они строят не одно разбиение выборки на непересекающиеся классы, а систему вложенных разбиений. Результат таксономии обычно представляется в виде таксономического дерева – дендрограммы [6, 9].

Сначала необходимо произвести расчет евклидового расстояния между всеми объектами, а затем последовательно объединять объекты от наименьшего расстояния до наибольшего.

Выбор метода кластеризации

Для разрабатываемого программного приложения было решено выбрать два метода кластеризации, которые бы разносторонне смогли отобразить ее результаты.

Для графического представления результатов кластеризации был выбран метод иерархической кластеризации, так как он позволяет следить за процессом кластеризации и получать наглядный результат, на котором графически можно отобразить иерархию кластеров и принадлежность объектов к каждому из них [9].

Для получения статистической информации о кластерах был выбран алгоритм k-means (метод k-средних).

Данному алгоритму чаще всего отдают предпочтение, когда стоит задача кластеризации больших объемов информации.

Кроме того, что этот алгоритм позволяет разбить выборку на кластеры, с помощью него можно:

- составить описательные характеристики каждого из кластеров;
- выделить «центр масс» каждого кластера, который будет являться его показательным объектом;
- выделить аномальные элементы из исходного множества.

В совокупности данные методы смогут дать общую картину кластеризации выборки, а также предоставить довольно детальную статистическую информацию по каждому кластеру.

Архитектура web-приложения для анализа данных

Для реализации выбранного метода кластеризации необходимо разработать программный продукт, позволяющий получать данные от пользователя и выполнять их анализ. При проектировании архитектуры web-приложения необходимо обеспечить возможность обмена данными с пользователем путём импорта-экспорта файлов различных форматов. Это позволит обрабатывать данные, полученные из различных источников.

Web-приложения представляют собой особый тип программ, построенных по архитектуре «клиент-сервер». Особенность их заключается в том, что само web-приложение находится и выполняется на сервере – клиент при этом получает только результаты работы. Работа приложения основывается на получении запросов от пользователя (клиента), их обработке и выдачи результата.

Каждое web-приложение, разрабатываемое на основе технологии ASP.NET состоит из информационной части, программного кода и сведений о конфигурации.

Информационная часть содержит статические и динамические элементы страницы и реализуется в виде web-форм. Статические элементы представляют собой типичные элементы языка HTML, динамические же компонуются программным кодом приложения во время его выполнения.

Программный код реализует логику, определенную в процедурах обработки данных, которые определяют реакцию приложения на запросы пользователя. Программный код исполняется сервером и взаимодействует с динамическими элементами информационной части для формирования отклика приложения (рис. 1).

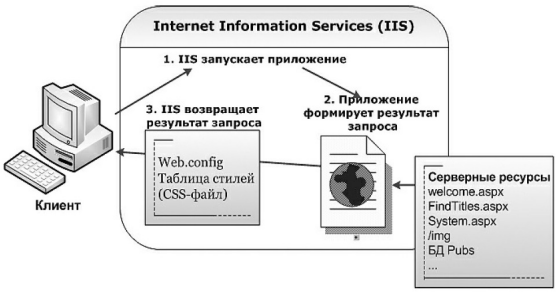

Рис. 1. Сценарий взаимодействия элементов web-приложения с клиентом

Сведения о конфигурации представляют собой файлы, содержащие параметры, определяющие способ исполнения приложения на сервере, параметры безопасности, реакцию приложения на возникающие ошибки и т.д. (рис. 2).

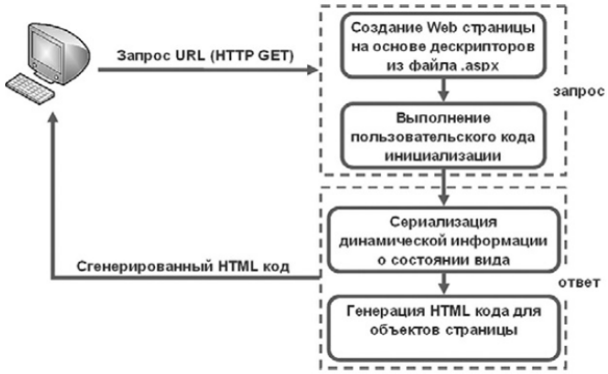

Рис. 2. Сценарий взаимодействия элементов web-приложения с клиентом при первом запросе

На рисунке 3 видно, что для того, чтобы произвести кластеризацию объектов пользователю изначально необходимо:

- импортировать данные об объектах из файла исходных данных;
- выбрать метод кластеризации;
- произвести предварительные настройки, которые будут использоваться в процессе кластеризации.

Представленное web-приложение может быть частью информационно-аналитической системы обработки данных. Получение исходных данных в виде файлов позволяет работать с наборами данных, полученных из различных источников.

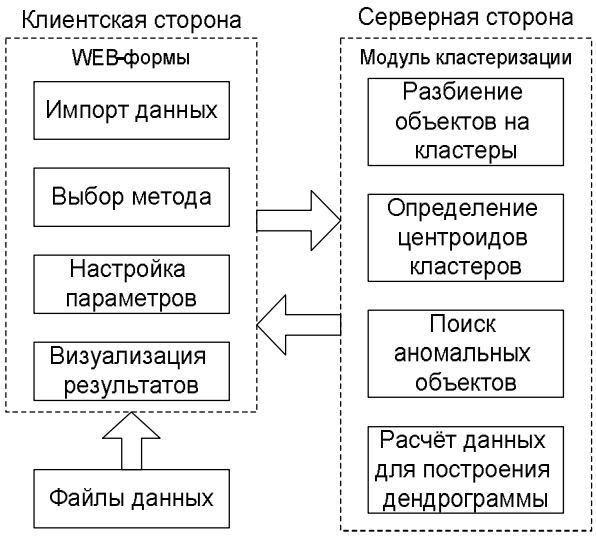

Рис. 3. Архитектура web-приложения анализа данных

Алгоритм работы метода k-средних

Для осуществления кластеризации данных методом k-средних, пользователю необходимо задать количество кластеров, а также максимальное число итераций.

Основные шаги алгоритма представлены на рис. 4. Каждый объект представлен последовательностью атрибутов. Предполагается, что данные уже нормализированы, то есть значения всех параметров (атрибутов) приведены к интервалу [0,1]. Это позволяет избежать искажения результатов из-за влияния размерности исходных данных, выраженных в разных единицах измерения.

После этого, всем объектам случайным образом задаются начальные номера кластеров.

На следующем шаге создается массив, в котором каждому номеру кластера k соответствует набор атрибутов m_{kj}. Значения для этих атрибутов рассчитываются путем вычисления среднего значения каждого атрибута среди всех объектов, попавших в данный кластер

$$m_{kj} = \frac{1}{N_k} \sum_{i=0}^{N_k - 1} x_{ij},$$

где N_k – число последовательностей в кластере k, x_{ij} – индивидуальные значения атрибутов последовательности.

Затем производится поиск центра масс (центроида) для каждого из кластеров. Этот поиск осуществляется путем нахождения евклидова расстояния от каждого объекта до средних значений атрибутов его кластера

$$Dist_{ki}^{mean} = \sqrt{\sum_{j=0}^{n-1} (m_{kj} - x_{ij})^2},$$

где n – количество атрибутов.

Рис. 4. Алгоритм работы метода k-средних

Центроидом c_k становится тот объект, атрибуты которого оказались ближе всего к средним зна-

чениям атрибутов кластера, т.е. тот элемент, для которого выполняется условие

$$\mathrm{Dist}_{ki}^{mean} = \min_{i}\left\{\mathrm{Dist}_{ki}^{mean}\right\}.$$

На следующем шаге необходимо найти расстояние от каждого объекта ко всем центроидам

$$\mathrm{Dist}_{ki}^{center} = \sqrt{\sum_{j=0}^{n-1}(c_{kj} - x_{ij})^2},$$

где c_{kj} – значение атрибута j центроида кластера k.

Объекту назначается тот кластер, к центроиду которого он оказался ближе всего.

Описанные операции будут выполняться до тех пор, пока не будет достигнуто максимальное число допустимых операций. Также условием выхода из цикла является достижение состояния, когда объектам перестанут назначаться новые кластеры.

В идеале, на выходе алгоритма можно получить набор объектов, которые разбиты на кластеры, центроиды которых равноудалены друг от друга.

Поиск аномальных объектов проводится после окончания процесса кластеризации. Для этого необходимо определить расстояния от центроида кластера до каждого объекта, который вошел в этот кластер. Самая удаленная последовательность может считаться аномальной.

Если несколько удаленных объектов имеют незначительную разницу в расстоянии, то можно считать, что в данном кластере отсутствуют аномальные объекты.

Результат работы метода k-средних

В качестве примера для испытаний работы разработанного алгоритма были выбраны данные группы пациентов. Данная группа состоит из 16 человек. В качестве характеристик рассматривались возраст этих пациентов, а также результаты их анализа крови: количество эритроцитов, тромбоцитов, лейкоцитов.

Входные данные имеют следующий вид:

[0] 64,00 3,40 198,00 1,10 61,00
[1] 39,00 3,90 141,00 7,90 130,00
[2] 45,00 3,90 141,00 6,30 106,00
[3] 33,00 4,00 175,00 7,90 137,00
[4] 63,00 3,90 141,00 7,10 106,00
[5] 57,00 3,30 209,00 1,10 65,00
[6] 56,00 4,10 175,00 7,10 133,00
[7] 83,00 3,80 173,00 6,10 106,00
[8] 50,00 3,70 186,00 6,60 116,00
[9] 54,00 3,30 232,00 0,40 61,00
[10] 55,00 3,80 164,00 6,30 102,00
[11] 39,00 3,80 130,00 7,10 109,00
[12] 41,00 3,90 164,00 7,20 123,00
[13] 38,00 3,60 107,00 5,10 92,00
[14] 50,00 3,90 164,00 7,70 116,00
[15] 52,00 3,70 130,00 7,10 120,00
[16] 20,00 4,00 186,00 8,40 133,00

В качестве параметров количества конечных кластеров и итераций заданы значения 3 и 30 соответственно.

На первом шаге данные были нормализированы в диапазоне от 0 до 1. Далее этот набор был обработан при помощи представленного выше алгоритма.

В результате пациенты были разделены на три кластера. Для каждого кластера были выделены их центроиды. Выделение таких центроидов снижает размерность набора, выделяя тем самым наиболее показательные элементы, изучение которых позволяет получить дополнительную информацию обо всём кластере.

Разделение на кластеры выглядит следующим образом:

- кластер 1 включает объекты [0], [5], [9], центроидом является [9];
- кластер 2 включает объекты [1], [3], [6], [8], [12], [14], [16], центроидом является [16];
- кластер 3 включает объекты [2], [4], [7], [10], [11], [13], [15], центроидом является [7].

Результаты кластеризации могут быть учтены врачами для определения их диагноза. Пациенты, попавшие в один кластер, скорее всего, имеют схожие заболевания.

Также могут быть определены аномальные данные, то есть выделены пациенты с нетипичными показателями анализа крови. Такие пациенты нуждаются в более детальном исследовании.

При выборе иерархического метода рассчитывались расстояния между результатами анализов по пациентам. Эти расстояния ранжированы по возрастанию от наименьшего. Они могут быть использованы для построения дендрограммы для визуального представления иерархической кластеризации.

Заключение

Рассмотрены основные направления и цели анализа данных. Проведен анализ методов кластеризации и два из них (метод k-средних и метод иерархической кластеризации) выбраны для разработки программного приложения.

Разработана архитектура разрабатываемого web-приложения. Описана его структура, а также сценарии его работы с клиентом.

Путём модификации метода k-средних разработан алгоритм для программной реализации в информационно-аналитической системе. Также был разработан алгоритм для расчета расстояний, которые могут использоваться для построения дендрограммы.

Было разработано web-приложение, в котором реализованы предложенные методы и алгоритмы. Также были проведены экспериментальные исследования по работе приложения на примере набора пациентов с указанием их медицинских показателей. В результате работы приложения набор пациентов был разделен на кластеры. Были выделены показательные и аномальные элементы для каждого из кластеров.

В результате работы иерархического метода был получен ранжированный список, состоящий из расстояний между пациентами. Он может использоваться для построения дендрограммы.

Таким образом, в работе выпалена модификация существующих методов кластеризации для применения их в разработке программного обеспечения информационно-аналитических систем. Это позволяет расширить область применения методов анализа данных для решения различных прикладных задач.

Предложенные в данной статье модели, алгоритмы и архитектура web-приложения могут быть использованы при проектировании и построении информационно-аналитической системы для обработки данных из различных источников.

Литература

1. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний [Текст] / Н. Г. Загоруйко. – Новосибирск : ИМ СО РАН, 1999. – 270 с.

2. Борисова, И. А. Методы решения задач распознавания образов комбинированного типа [Текст] : дисс. .. канд. техн. наук : 05.13.17 / И. А. Борисова. – Новосибирск, 2008. – 126 с.

3. Мендель, И. Д. Кластерный анализ [Текст] / И. Д. Мендель. – М. : Финансы и статистика, 1988. – 176 с.

4. Kaufman, L. Finding groups in data : an introduction to cluster analysis [Text] / L. Kaufman., P. J. Rousseeuw. – John Wiley & Sons, 2009. – 368 p.

5. Миркин, Б. Г. Методы кластер-анализа для поддержки принятия решений : обзор [Текст] / Б. Г. Миркин. – М. : Изд. дом Нац. иссл. ун-та «Высшая школа экономики», 2011. – 88 с.

6. Жамбю, М. Иерархический кластер-анализ и соответствия [Текст] / М. Жамбю. – М. : Финансы и статистика, 1988. – 345 с.

7. Zhou, F. Hierarchical aligned cluster analysis for temporal clustering of human motion [Text] / F. Zhou, F. De la Torre, J. K. Hodgin // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2013. – Vol. 35, No. 3. – P. 582-596.

8. Patient Characteristic Cluster Analysis Predicts Response To Therapy To Oral Treprostinil In Pulmonary Arterial Hypertension [Text] / Y. Rao, K. Shen, S. Rajagopal, K. S. Parikh // D53. The Promised Land : Clinical Studies In Pulmonary Hypertension. – American Thoracic Society, 2016. – P. A7340-A7340.

9. Tsitsimpelis, I. Partitioning of indoor airspace for multi-zone thermal modelling using hierarchical cluster analysis [Text] / I. Tsitsimpelis, C. J. Taylor // 2015 European Control Conference (ECC 2015). – IEEE, 2015. – P. 410-415.

10.Лагутин, М. Б. Наглядная математическая статистика [Текст] / М. Б. Лагутин. – М. : П-центр, 2003. – 210 с.

11.Yano, K. Labeling Feature-Oriented Software Clusters for Software Visualization Application [Text] / K. Yano, A. Matsuo // 2015 Asia-Pacific Software Engineering Conference (APSEC). – IEEE, 2015. – P. 354-361.

12.Журавлёв, Ю. И. Распознавание. Математические методы. Программная система. Практические применения [Текст] / И. Ю. Журавлёв, В. В. Рязанов, О.В. Сенько. – М. : ФАЗИС, 2006. – 176 с.

References

1. Zagoruiko, N. G. *Prikladnye metody analiza dannykh i znanii* [Applied methods of data analysis and knowledge]. Novosibirsk, IM SO RAN Publ., 1999. 270 p.

2. Borisova, I. A. *Metody resheniya zadach raspoznavaniya obrazov kombinirovannogo tipa: diss. .. kand. tekhn. nauk : 05.13.17* [Methods for solving problems of pattern recognition combined type: the dissertation .. candidate of technical sciences]. Novosibirsk, 2008. 126 p.

3. Mendel', I. D. *Klasternyi analiz* [Cluster analysis]. Moscow, Finansy i statistika Publ., 1988. 176 p.

4. Kaufman, L., Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons Publ., 2009. 368 p.

5. Mirkin, B.G. *Metody klaster-analiza dlya podderzhki prinyatiya reshenii: obzor* [Methods of cluster analysis for decision support: a review]. Moscow, Izd.

dom Nats. issl. un-ta «Vysshaya shkola ekonomiki» Publ., 2011. 88 p.

6. Zhambyu, M. *Ierarkhicheskii klaster-analiz i sootvetstviya* [Hierarchical cluster analysis and compliance]. Moscow, Finansy i statistika Publ., 1988. 345 p.

7. Zhou, F., De la Torre, F., Hodgins, J. K. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2013, vol. 35, no. 3, pp. 582-596.

8. Rao, Y., Shen, K., Rajagopal, S., Parikh, K. S. Patient Characteristic Cluster Analysis Predicts Response To Therapy To Oral Treprostinil In Pulmonary Arterial Hypertension. *D53. The Promised Land: Clinical Studies In Pulmonary Hypertension*. American Thoracic Society Publ., 2016, pp. A7340-A7340.

9. Tsitsimpelis, I., Taylor, C. J. Partitioning of indoor airspace for multi-zone thermal modelling using hierarchical cluster analysis. *2015 European Control Conference (ECC 2015),* IEEE Publ., 2015, pp. 410-415.

10.Lagutin, M. B. Naglyadnaya matematicheskaya statistika [Transparent Mathematical Statistics]. Moscow, P-tsentr Publ., 2003. 210 p.

11.Yano, K., Matsuo, A. Labeling Feature-Oriented Software Clusters for Software Visualization Application. *2015 Asia-Pacific Software Engineering Conference (APSEC)*. IEEE Publ., 2015, pp. 354-361.

12. Zhuravlev, I. Yu., Ryazanov, V. V., Sen'ko, O. V. *Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya* [Recognition. Mathematical methods. Software system. Practical applications]. Moscow, FAZIS Publ., 2006. 176 p.

Поступила в редакцию 11.04.2016, рассмотрена на редколлегии 12.05.2016

ВИКОРИСТАННЯ МЕТОДУ КЛАСТЕРИЗАЦІЇ В ІНФОРМАЦІЙНО-АНАЛІТИЧНІЙ СИСТЕМІ

О. С. Яшина, М. О. Щербак

Стаття присвячена застосуванню методів кластеризації даних при розробці інформаційно-аналітичних систем. Проводиться оглядовий аналіз існуючих методів кластеризації. Запропоновано алгоритми, побудовані на основі модифікованих статистичного (метод k-середніх) і ієрархічного методів кластеризації. Розроблено архітектуру інформаційно-аналітичної системи у вигляді web-додатку, що використовує запропоновані методи і алгоритми для обробки даних, отриманих з різних джерел. На основі цих алгоритмів проведено розробку програми для кластеризації наборів об'єктів.

Ключові слова: аналіз, кластеризація, кластерний аналіз, ієрархічна кластеризація, метод k-середніх.

APPLYING OF CLUSTERING METHODS IN THE INFORMATION-ANALYTICAL SYSTEM

H. S. Yashina, M. A. Shcherbak

The article is dedicated to the applying of data clustering methods in the development of information-analytical systems. Review of the existing clustering methods is executed. The algorithms that are based on the modified statistical (k-means) and hierarchical clustering methods are offered. The architecture of information-analytical system as a web-application that uses the proposed methods and algorithms for processing data received from various sources is developed. On the basis of these algorithms is made development of applications for sets of objects clustering.

Keywords: analysis, clustering, cluster analysis, hierarchical clustering, k-means.

Яшина Елена Сергеевна – канд. техн. наук, доц., доц. каф. информационно управляющих систем, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.

Щербак Марина Александровна – магистр каф. информационно управляющих систем, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.

Yashina Helena Sergeevna – candidate of technical science, docent of Information and control system chair of National Aerospace University Kharkov Aviation Institute, Kharkov, Ukraine.

Shcerbak Marina Aleksandrovna – master of Information and control system chair of National Aerospace University Kharkov Aviation Institute, Kharkov, Ukraine.