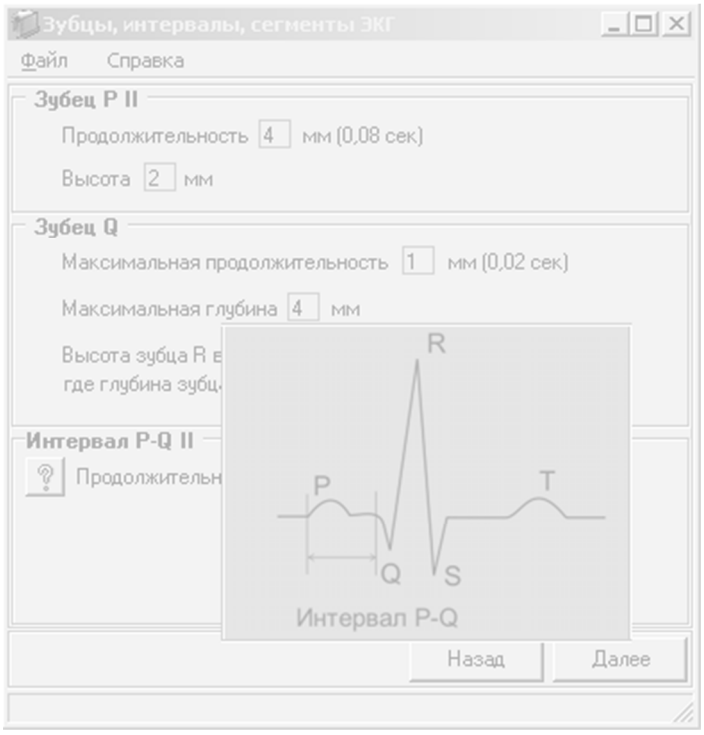

Рис. 1. Окно работы обучающей системы

На рис. 1 приведено окно работы обучающей системы. При необходимости ввод данных сопровождается справочной информацией о наиболее сложных параметрах ЭКГ, в виде схем и рисунков, а также подробной инструкцией о заполнении полей форм. Реализована простая и удобная навигация по заполненным и незаполненным формам посредством кнопок «Далее» и «Назад».

Предложена простая и удобная система навигации по правилам. Программой предусмотрена возможность проверки базы знаний на целостность, то есть в случае удаления, добавления правил осуществляется контроль на корректность внесенных изменений. Текущая база знаний включает в себя 98 правил. Предполагается дальнейшее расширение базы знаний, за счет детализации ЭКГ-симптомов.

ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ ДАННЫХ ДЛЯ ВЫЯВЛЕНИЯ ОБОБЩЕННЫХ ПРИЕМОВ СОВЕРШЕНСТВОВАНИЯ ЭКСПЛУАТАЦИОННЫХ ХАРАКТЕРИСТИК

В. М. Зарипова, И. Ю. Петрова, А. А. Пучкова
Астраханский государственный
архитектурно-строительный университет
Астраханский государственный университет

В статье приведено описание проблемы выявления обобщенных приемов совершенствования эксплуатационных характеристик устройств. Приведена крат-

кая классификация существующих методов кластерного анализа, проведен их анализ, выявлены достоинства и недостатки каждого метода. Статья содержит формулы наиболее часто используемых метрик (мер близости) между объектами выборки. Охарактеризована предлагаемая методика кластеризации патентной информации для выявления обобщенных приемов совершенствования, приведена соответствующая диаграмма активности. Разработанная методика реализована в подсистеме “Patent search” программно-технического комплекса «Интеллект». Статья содержит пример использования процедуры кластеризации для выявления приемов совершенствования калориметрических биосенсоров. В заключении сделан вывод об эффективности выбранных решений и о необходимости дальнейшего развития механизма кластеризации, в частности, реализации автоматической формулировки выявленных обобщенных приемов.

***Ключевые слова**: кластерный анализ, прием совершенствования, элемент информационно-измерительных и управляющих систем, мера близости, Data Mining.*

APPLICATION OF CLUSTERING DATA TO DETERMINE THE GENERALIZED METHODS OF OPERATIONAL CHARACTERISTICS IMPROVEMENT

V. M. Zaripova, I. Yu. Petrova, A. A. Puchkova

Astrakhan State University of Architecture and Civil Engineering

Astrakhan State University

The article describes the determination problem of generalized methods of device operational characteristics improvement. The existing methods of cluster analysis were briefly classified and analyzed, this analysis revealed the advantages and disadvantages of each method. The article contains the formulas of commonly used metrics (proximity measures) between sampling objects. The proposed technique of patent information clustering to identify the generalized methods of improvement is described, the chart shows the relevant activity diagram. Developed technique was implemented in the "Patent search" subsystem of "Intellect" software and technical package. This article contains an example of clustering procedures using to identify methods of improvement of calorimetric biosensors. Finally, it was made a conclusion about the efficiency of the chosen solutions, and about the need for further development of the clustering technique, in particular, the implementation of automatic formulation of the revealed generalized methods.

***Key words:** cluster analysis, method of improvement, element of information-measuring and management systems, proximity measure, Data Mining.*

Введение

Кластерный анализ – одно из направлений интеллектуального анализа данных (Data Mining), представляющий собой автоматическую классификацию объектов. Главной целью кластерного анализа является выделение в исходных многомерных данных набора однородных подмножеств таким образом, чтобы объекты внутри одной группы обладали как можно большей схожестью и при этом максимально отличались от объектов других групп. Схожесть при этом понимается как близость объектов в многомерном пространстве признаков, в таком

случае задача кластерного анализа состоит в выделении в этом пространстве естественных скоплений объектов, которые и являются однородными группами [1].

Проблема выявления обобщенных приемов совершенствования

Применительно к выявлению обобщенных приемов совершенствования эксплуатационных характеристик элементов информационно-измерительных и управляющих систем задача кластерного анализа может быть сформулирована следующим образом. Прием совершенствования – набор изменений в технологии изготовления, конструкции или составе технического устройства, позволяющих достичь положительного эффекта по сравнению с прототипом. Приемы совершенствования эксплуатационных характеристик могут быть объединены в обобщенные приемы и применены к другому прототипу, что в большом числе случаев может привести к созданию более совершенного технического решения. Обобщенные приемы могут быть выделены на основе анализа существующих технических решений, подавляющее большинство которых на сегодняшний день представлено в виде патентов на изобретения и полезные модели. Следовательно, возможно проведение кластерного анализа документов, содержащих исходный текст патента, для выявления групп однородных объектов, каждая из которых будет представлять устройства, использующие один обобщенный прием.

Таким образом, пусть в результате отбора элементов генеральной совокупности имеется некая выборка текстовых документов патентов $S_D = \{d^{(1)},..., d^{(M)}\}$. Для выявления обобщенных приемов необходимо сформировать *K* однородных групп элементов выборки. При этом *K* не может быть определено заранее, поскольку число используемых в выборке обобщенных приемов не известно исследователю. Каждый объект генеральной совокупности может быть описан при помощи набора переменных $P = \{P_1,...,P_n\}$, включающего переменные различных типов. Множество значений переменной P_i обозначим как V_i. Тогда для каждого элемента выборки *e* может быть составлен набор значений $p = p(e) = p_1(e),..., p_n(e)$, где $p_i(e)$ – значение переменной P_i объекта *e*. Все полученные наборы значений могут быть сведены в матрицу *A* размером *M* x *N*. В результате кластерного анализа мы получим разбиение выборки S_D на *K* групп: $R = \{G^{(1)},...,G^{(k)}\}$, при этом $G^{(i)} = \{e^{(j_1)},..., e^{(j_c)}\}$, где $e^{(j_r)}$ – объект выборки S_D, *c* – число объектов в i-й группе, $i \in [1, K]$. Каждая из групп разбиения *R* представляет собой один кластер. Таким образом может быть введена группировочная функция f, представляющая собой отображение f: $S_D \rightarrow \{1,..., K\}$.

Анализ алгоритмов кластеризации

Результаты анализа существующих на сегодняшний день алгоритмов кластеризации позволили выявить несколько основных подходов к кластеризации. Ниже приведены краткие описания каждого из них.

- Плоские методы. К данной группе относятся такие методы, как K means, spherical K-means [2]. Для их работы необходимо априорное определение числа кластеров *K*. Далее производится первоначальное разбиение выборки, и для каждой группы определяется центр тяжести, представляющий собой вектор средних значений переменных объектов, входящих в группу. Затем итеративно производятся повторные разбиения выборки таким образом, чтобы внутригрупповое рассеяние объектов было минимальным. Внутригрупповое рассеяние может быть определено по формуле (1):

$$d(R) = \sum_{i=1}^{K}\sum_{r=1}^{c}\sum_{l=1}^{N}(e_l^{(j_r)} - \frac{1}{c}\sum_{r=1}^{c} e_l^{(j_r)})^2 \qquad (1)$$

- Методы, использующие теорию графов. Выборка представляется в виде графа с *M* вершин, вес ребер которого определяется как расстояние между двумя объектами выборки [3]. Для сформированного графа производится построение минимального остовного дерева. Результирующие кластера могут быть получены из остовного дерева путем удаления ребер с максимальной длиной.

- Иерархические методы подразделяются на две категории: агломеративные и дивизимные. Итоговое разбиение представляется в виде дерева, корнем которого является вся выборка, а листьями – отдельные ее объекты. В случае агломеративных методов построение дерева ведется от листьев к корню, в случае дивизимных – от корня к листьям. Результирующие кластера в обоих случаях представляют собой вложенную иерархию подгрупп. Примерами подобных методов могут служить Single Link, Complete Link и др. [2].

- Нейросетевые методы. К ним относятся, в частности, самоорганизующиеся сети Кохонена [4]. Это однослойная нейронная сеть, в которой каждый нейрон соответствует одному кластеру. Число входов одного нейрона соответствует *N* – числу переменных из набора *P*. Обучение происходит на основе обучающей выборки путем перераспределения весов нейронов.

- Семантические методы. Применимы только для кластеризации текстовых данных. Для каждого объекта выборки строится суффиксное дерево, представляющее собой все суффиксы входной строки. На их основе формируется итоговое суффиксное дерево, строящееся из фраз исходных документов. Узлами этого дерева являются результирующие кластера [5].

- Эволюционные методы. К ним относится генетический алгоритм [6]. Каждый вариант разбиения *R* представляет собой одну особь

в популяции. В ходе работы алгоритма родительские особи скрещиваются между собой, что приводит к формированию новых особей. Дочерние особи могут случайным образом подвергаться мутации, что повышает разнообразие рассматриваемых вариантов разбиения. После формирования каждого нового поколения в результате естественного отбора происходит уничтожение наименее приспособленных особей. Критерием приспособленности в данном случае может служить фитнес-функция, равная критерию внутригруппового рассеяния.

Методы каждой из вышеописанных групп обладают своими достоинствами и недостатками, которые приведены в таблице 1.

Таблица 1

Результаты анализа методов кластеризации

Группа	*Достоинства*	*Недостатки*
Плоские	Высокая скорость	Необходимость задания числа кластеров. Высокая чувствительность к выбросам. Высокая чувствительность к начальным значениям центров масс кластеров
Нейросетевые	Возможность реализации параллельных вычислений. Высокая скорость работы	Необходимость задания числа кластеров. Необходимость наличия обучающей выборки
Эволюционные	Возможность реализации параллельных вычислений. Наиболее быстрое достижение квазиоптимального решения на больших объемах данных	Не гарантируют нахождения оптимального решения. Требуют большого размера популяции с высокой степенью вариативности особей
Семантические	Высокая скорость. Отсутствие необходимости задания числа кластеров	Высокая стоимость составления дерева в случае передачи исходных текстовых документов по сети
Иерархические	Высокая точность. Структура разбиения	Высокая алгоритмическая сложность (низкая скорость)
На основе теории графов	Высокая точность	

Меры близости между объектами выборки

Под мерой близости (метрикой) между двумя объектами выборки понимается функция $h(e^{(i)}, e^{(j)})$, удовлетворяющая ряду условий:

- $i,j \in [1, M] : h(e^{(i)}, e^{(j)}) \geq 0$,
- $h(e^{(i)}, e^{(j)}) = h(e^{(j)}, e^{(i)})$,
- $h(e^{(i)}, e^{(j)}) \leq h(e^{(i)}, e^{(k)}) + h(e^{(k)}, e^{(j)})$.

Существуют различные меры близости между двумя наборами значений переменных, в таблице 2 приведены наиболее часто используемые из них [1].

Таблица 2

Меры близости между объектами выборки

Название	*Формула*	*Тип шкалы*
Расстояние Евклида	$h\left(e^{(i)}, e^{(j)}\right) = \sqrt{\sum_{k}^{N}(e_k^{(i)} - e_k^{(j)})^2}$	Количественные переменные
Расстояние Спирмена	$h\left(e^{(i)}, e^{(j)}\right) = \sum_{k}^{N}(e_k^{(i)} - e_k^{(j)})^2$	
Манхэттенское расстояние	$h\left(e^{(i)}, e^{(j)}\right) = \sum_{k}^{N}\lvert e_k^{(i)} - e_k^{(j)}\rvert$	
Расстояние Чебышева	$h\left(e^{(i)}, e^{(j)}\right) = \max_{k=1,N}(\lvert e_k^{(i)} - e_k^{(j)}\rvert)$	
Коэффициент Хэмминга	$h\left(e^{(i)}, e^{(j)}\right) = \frac{\sum_k^N f(e_k^{(i)}, e_k^{(j)})}{M}$, где $f\left(e_k^{(i)}, e_k^{(j)}\right) = \begin{cases} 1, e_k^{(i)} = e_k^{(j)} \\ 0, e_k^{(i)} \neq e_k^{(j)} \end{cases}$	Номинальные переменные
Коэффициент Рао	$h\left(e^{(i)}, e^{(j)}\right) = \frac{\sum_k^N f(e_k^{(i)}, e_k^{(j)})}{M}$, где $f\left(e_k^{(i)}, e_k^{(j)}\right) = \begin{cases} 1, e_k^{(i)} = e_k^{(j)} = 1 \\ 0, \text{в других случаях} \end{cases}$	

Наибольшее распространение среди них получили расстояния Евклида и Спирмена, а также коэффициент Хэмминга.

Предлагаемая методика кластеризации патентной информации

Задача по выявлению обобщенных приемов эксплуатационных характеристик элементов информационно-измерительных и управляющих систем должна быть решена в рамках программно-технического комплекса «Интеллект» [7], который представляет собой web-решение, поэтому семантические методы в данном случае неприменимы. Поскольку невозможно заранее по набору патентов определить число использованных в них обобщенных приемов, также в чистом виде неприменимы плоские, эволюционные и нейросетевые методы. Высокая длительность операций не позволяет применять и иерархические методы, а также методы на основе теории графов.

Следовательно, была предложена новая комплексная методика для кластеризации патентной информации, сочетающая в себе иерархический агломеративный и плоский подходы. Она включает в себя два

варианта использования (полный и быстрый) и состоит из семи основных этапов. Соответствующая диаграмма активности для быстрого варианта приведена на рис. 1. В полном варианте вместо плоской кластеризации используется иерархическая агломеративная кластеризация, что позволяет достичь большей точности разбиения, но приводит к значительному увеличению длительности процедуры.

Выявление приемов производится для указанного пользователем набора близких физико-технических эффектов (ФТЭ). Система производит отбор патентов, релевантных указанным ФТЭ, под релевантностью патента некоторому ФТЭ понимается вероятность использования данного ФТЭ в этом патенте. Релевантность патента конкретному ФТЭ равна 1 в случае верификации его паспорта экспертом и подтверждения факта использования. В качестве меры близости между двумя кластерами было выбрано расстояние между центрами масс (в полном варианте – невзвешенное попарное расстояние), а в качестве меры близости между двумя объектами – расстояние Спирмена (как наиболее чувствительное к выбросам).

Методика была реализована в подсистеме “Patent Search” комплекса «Интеллект» [8]. Для апробации методики была произведена кластеризация документов со следующими условиями: анализируемые патенты должны были использовать пироэлектрический ФТЭ, а также в тексте должно было присутствовать одно из слов: «биосенсор», “biosensor”. В случае корректной работы подсистема должна была выявить кластер калометрических биосенсоров. Поскольку для системы «Интеллект» в данный момент ведется разработка модуля синтеза биосенсоров [9], подсистема “Patent Search” также должна поддерживать операции над соответствующими патентами.

В результате отбора документов на русском и английском языках согласно введенным условиям подсистемой была сформирована следующая выборка патентов: {US4829003 A, US20100028969 A1, US 5108576, RU 2266959, US4551425, US20110182776 A1, US 20050196322 A1, US20130052632 A1, WO 1990013017 A1}. В результате кластеризации система выявила следующие обобщенные приемы:

Использование в конструкции материалов, обладающих низкой теплопроводностью и при этом высокой теплоемкостью, что позволяет минимизировать потери теплового сигнала, возникающего при реакции субстрата фермента, что, в свою очередь, приводит к увеличению чувствительности устройства.

Пиро-оптическое детектирование. Пленка пироэлектрика облучается импульсами света определенной длины волны. Пленка покрыта пленочными электродами, на поверхность которых нанесен иммобилизованный реагент, способный при облучении связываться с анализируемым материалом, что приводит к выделению дополнительного

тепла, преобразуемого в электрический сигнал. Такая конструкция приводит к улучшению соотношения сигнал/шум.

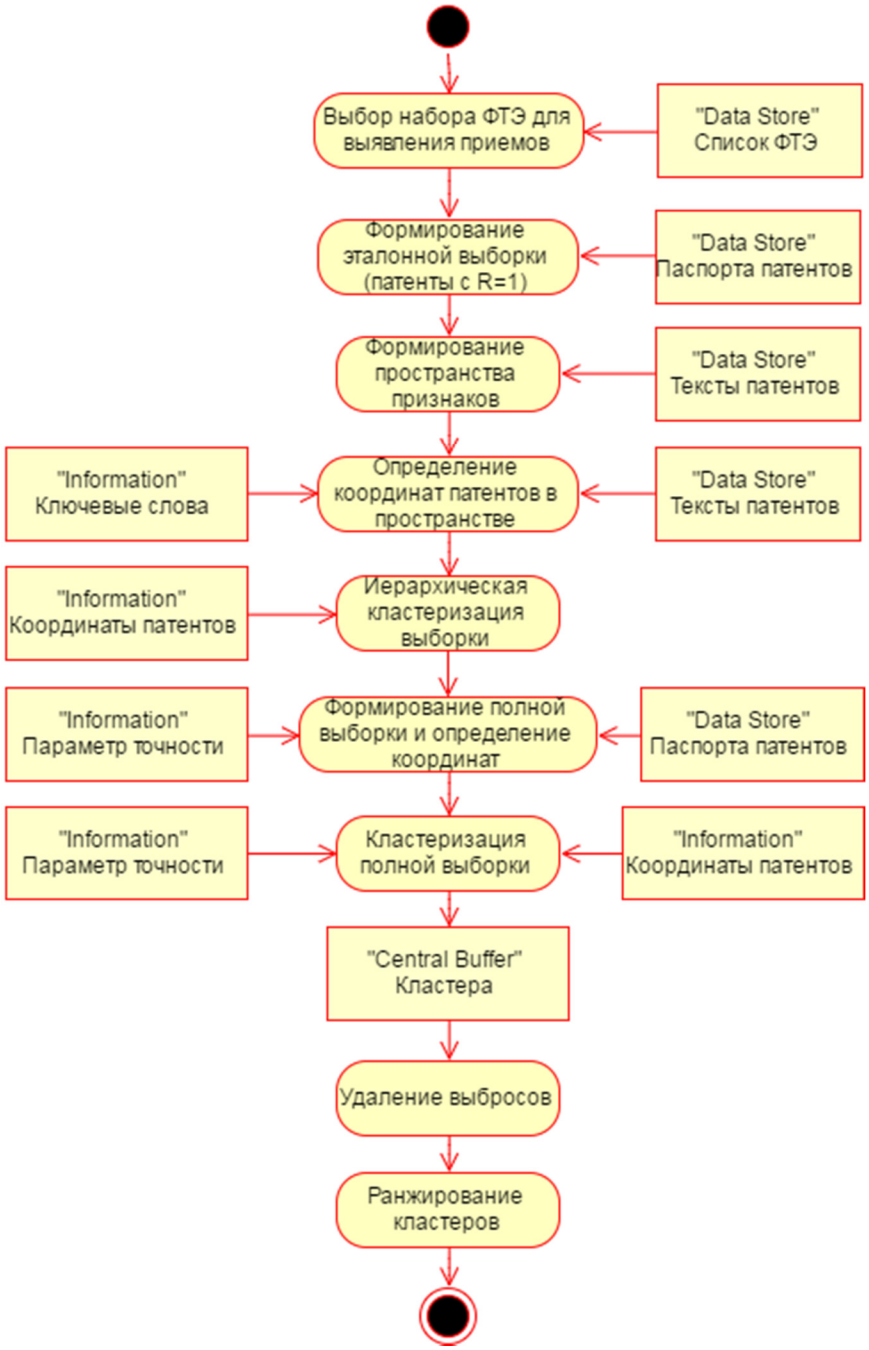

Рис. 1. Диаграмма активности быстрого варианта кластеризации

Включение тонкопленочных пироэлектриков с помощью дифференциальной схемы, при этом один из пироэлектриков покрывается эпоксидной пленкой, к которой при помощи фотохимических реагентов могут быть присоединены белки.

Выявленные системой в результате процедуры кластера отражают истинные обобщенные приемы, использованные в патентах из выборки, что свидетельствует об адекватности разработанной методики.

Заключение

В результате анализа достоинств и недостатков существующих алгоритмов кластеризации была разработана методика автоматического выявления обобщенных приемов совершенствования эксплуатационных характеристик элементов информационно-измерительных и управляющих систем. Созданная методика была реализована в подсистеме Patent Search программно-технического комплекса «Интеллект», по результатам опытной эксплуатации был сделан **вывод** об эффективности принятых алгоритмических решений. В дальнейшем планируется дальнейшее развитие разработанной подсистемы с целью автоматического формулирования выявленных обобщенных приемов.

Исследование было выполнено частично при поддержке РФФИ (грант №116-37-00258/16).

Список литературы

1. Мандель И. Д. Кластерный анализ. М. : Финансы и статистика, 1988.
2. Jain A. Clustering methods and algorithms. Prentice-Hall Inc., 1988.
3. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters // IEEE Trans. Comput., 1971. C-20. P. 68–86.
4. Kohonen T. Self-Organization and Associative Memory. 3rd ed. Springer information sciences series. Springer-Verlag, New York, 1989.
5. Zamir O. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results // A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. University of Washington, 1999.
6. Goldberg D. E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.
7. Зарипова В. М., Цырульников Е. С., Киселев А. А. «Интеллект» для развития навыков инженерного творчества // Alma Mater (Вестник высшей школы). 2012 (1). С. 58–61.
8. Автоматизированная система выявления приемов улучшения эксплуатационных характеристик на основе кластеризации патентной информации : патент № 2016613179 от 18.03.2016/ А. А. Пучкова, И. Ю. Петрова.
9. Петрова И. Ю., Зарипова В. М., Лежнина Ю. А., Сокольский В. М., Митченко И. А. Энергоинформационные модели биосенсоров // Вестник АГТУ. Серия управление, вычислительная техника и информатика. 2015. № 3. С. 35–48.

ЭКСПЕРТНАЯ СИСТЕМА ДЛЯ ВЫБОРА СМАРТФОНА БРЕНДА SAMSUNG

Л. С. Смирнов, Т. Л. Тен

Карагандинский экономический университет Казпотребсоюза

В данной статье рассмотрены теоретические и практические основы построения экспертной системы, реализованной на языке программирования c#. Рассмотрены и предложены решения по ключевым вопросам, касающимся программного проектирования достаточно качественных элементов экспертной системы, приме-