

УДК 519.237.8

И. В. Аникин, А. П. Кирпичников, А. О. Назаров

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ МЕТОДА КОНЦЕПТУАЛЬНОЙ КЛАСТЕРИЗАЦИИ
ОБЪЕКТОВ, ХАРАКТЕРИЗУЕМЫХ НЕЧЕТКИМИ ПАРАМЕТРАМИ*Ключевые слова:* кластеризация данных, нечеткая логика, выделение пользовательских ролей.*В статье рассматривается аналог метода концептуальной кластеризации COBWEB, способный работать с объектами, характеризуемыми нечеткими параметрами. Значения данных параметров определяются в виде функций принадлежности. Разработан программный комплекс, позволяющий осуществлять концептуальную кластеризацию подобных объектов и решать различные практические задачи. Решена задача по автоматизации формирования пользовательских ролей в корпоративной информационной сети. Проведены экспериментальные исследования для сравнительной оценки точности кластеризации.**Keywords:* data clustering, fuzzy logic, user roles discovery.*We consider the method of conceptual clustering based on COBWEB, which is able to work with the fuzzy parameters. These parameters are described by the membership functions. We develop the software for fuzzy conceptual clustering. We can use this software in various applications. We used this software for user roles discovery in corporate networks. We present experimental results of the accuracy of clustering.*

Задача кластеризации является одной из важнейших задач интеллектуального анализа данных в различных проблемных областях [1,2]. Кластеризация является примером задачи обучения без учителя и сводится к разбиению исходного множества объектов на подмножества классов таким образом, что элементы одного класса были бы схожи между собой, а элементы различных классов были бы максимально различны. Основной сложностью применения классических методов кластеризации для решения практических задач является то, что многие реальные объекты могут быть описаны исключительно нечеткими параметрами. В связи с этим, для кластеризации подобных объектов в последнее время активно развиваются методы нечеткой кластеризации.

В настоящее время известно множество методов нечеткой кластеризации, таких как Fuzzy C-Means, FORTICS и др. [3]. Данные алгоритмы формируют кластеры, границы которых размыты, а объект может принадлежать более чем одному кластеру с различными степенями принадлежности. Однако следует отметить, что большинство алгоритмов нечеткой кластеризации работают с четкими значениями параметров объектов, формируя кластеры, например, на основе оценки расстояний между объектами и центром кластера [4]. Такой подход не позволяет эффективно осуществлять кластеризацию объектов с нечетко заданными значениями параметров и ограничивает практические области применения данных методов.

В связи с этим, актуальной является разработка и практическое применение методов кластерного анализа, способных учитывать нечеткую природу объектов, то есть работать с параметрами, заданными в нечеткой форме в виде функций принадлежности.

В работе [5] был предложен метод концептуальной кластеризации COBWEB для объектов с бинарными параметрами. Данный метод строит дерево классификации объектов с вероятностными описаниями концептов. Выбор возможного способа кластеризации объектов основан на значениях функции полезности кластеризации. Развивая данный метод, авторами статьи в [6] предложен его аналог, способный работать с объектами, характеризуемыми нечеткими параметрами. Значения данных параметров определяются в виде функций принадлежности. Разработанный метод предполагает реализацию классического метода концептуальной кластеризации в следующих условиях

1. Множество распознаваемых объектов $O = \{O_i\}_{i=1, \dots, r}$, характеризуется нечеткими параметрами $\tilde{A} = \{\tilde{A}_j\}_{j=1, \dots, m}$,

2. Значение параметра $\tilde{A}_j$ для объекта O_i определяется в виде функции принадлежности $\mu_{\tilde{A}_j}(x) \in \{0,1\}$.

3. Степень сходства двух функций принадлежности $\mu_{\tilde{A}_j}(x)$ и $\mu_{\tilde{A}_k}(x)$ определяется их наибольшей верхней границей, в виде:

$$v_{jk} = \sup_{x \in X} \min\{\mu_{\tilde{A}_j}(x), \mu_{\tilde{A}_k}(x)\} \in [0,1],$$

где $\mu_{\tilde{A}_j}(x)$ - функция принадлежности параметра $\tilde{A}_j$ для объекта O_i , а $\mu_{\tilde{A}_k}(x)$ - функция принадлежности параметра $\tilde{A}_k$ для объекта O_i .

4. Основываясь на формуле полезности кластеризации классического метода COBWEB и условиях 1-3, оценка полезности кластеризации осуществляется по модифицированной формуле:

$$CU^* = \frac{\sum_{k=1}^n P(C_k) \left[\sum_{j=1}^m \sum_{i=1, O_i \in C_k}^r \sum_{t=1, O_t \in C_k}^r v_{jit} / |C_k| - \sum_{j=1}^m \sum_{i=1}^r \sum_{t=1}^r v_{jit} / r \right]}{n},$$

где n – количество кластеров.

Предложенный метод может быть эффективно использован для решения задач концептуальной кластеризации объектов с нечеткой природой в различных проблемных областях - технических, естественнонаучных, социальных. Для практического решения данных задач был разработан программный комплекс в среде C# [7], позволяющий осуществлять кластеризацию, проводить численно-параметрические исследования метода, помогать интерпретировать полученные результаты.

С помощью разработанного программного комплекса решена практическая задача по автоматизации построения пользовательских ролей [8] в корпоративной информационной системе (КИС) для системы разграничения доступа пользователей [9]. Грамотное разграничение доступа пользователей к ресурсам информационной системы максимально снижает возможность выполнения несанкционированных действий со стороны ее пользователей, а ролевой подход к разграничению доступа считается наиболее эффективным для крупных КИС. При этом решение задачи формирования пользовательских ролей является нетривиальным для крупных КИС, включающих большое количество субъектов и объектов доступа. Ее решение в ручном режиме является сложным, ресурсоемким, затратным по времени, ведущим к возможным ошибкам в формировании профилей пользовательских ролей. Автоматизация процесса построения пользовательских ролей является актуальной задачей, а применение методов кластерного анализа является одним из эффективных механизмов такой автоматизации. Практическое решение данной задачи позволяет с одной стороны значительно упростить работу администратора информационной безопасности по формированию пользовательских ролей в КИС, а с другой стороны позволяет обнаруживать anomальное поведение пользователей в КИС, выявляя недобросовестных сотрудников, использующих информационные ресурсы и сервисы не только для выполнения своих функциональных обязанностей, но и в личных целях [10,11]. При этом поведенческую составляющую пользователя КИС следует рассматриваться как объект с нечеткими параметрами.

Архитектура системы, предназначенной для автоматизации построения пользовательских ролей, представлена на рисунке 1 и состоит из 4 модулей.

1. Модуль сбора данных, предназначенный для сбора статистических данных об объектах исследования. Для решения задачи автоматизации и формирования пользовательских ролей в КИС сбор данных осуществляется на основе анализа журналов событий.

2. Модуль подготовки данных на основе собранной статистики, позволяет с участием эксперта построить функции принадлежности для каждого объекта по каждому из параметров. Выходом данного модуля являются сформированные нечеткие описания объектов в виде функций принадлежности их параметров.

3. Модуль обработки данных реализует разработанный метод нечеткой концептуальной кластеризации.



Рис. 1 - Архитектура системы, предназначенной для автоматизации построения пользовательских ролей

Разработанный программный комплекс помогает интерпретировать эксперту полученные результаты кластеризации. Результаты работы программного комплекса можно увидеть в отдельном окне. В нем представляются результаты кластеризации, полученная иерархия кластеров.

В качестве примера задача автоматизации формирования пользовательских ролей была решена для информационной системе конкретной организации, структура которой представлена на рис. 2.

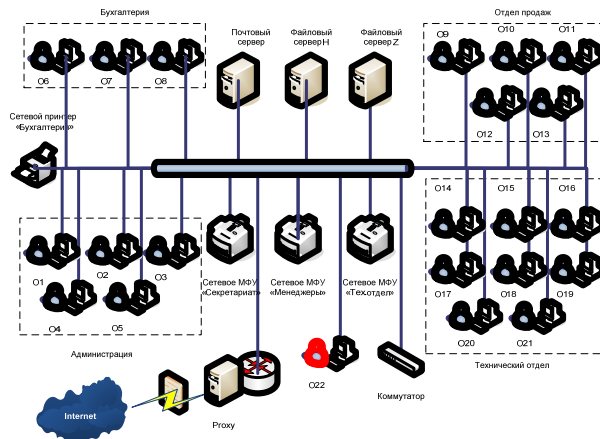


Рис. 2 - Структура КИС предприятия

Осуществлялась кластеризация 22 пользователей: $O = \{O_i\}_{i=1}^{22}$, распределенных по отделам следующим образом:

– Администрация (Генеральный директор – О1; Финансовый директор – О2; Технический директор – О3; Секретарь – О4; Офис-менеджер – О5)

– Бухгалтерия (Главный бухгалтер – О6; Зам. главного бухгалтера – О7; Бухгалтер – О8)

– Отдел продаж (Начальник отдела продаж – О9; Зам. начальника отдела продаж – О10; Менеджеры – О11 – О13)

– Технический отдел (Начальник технического отдела – О14; Зам. начальника технического отдела – О15; Технические специалисты – О16 – О21; Администратор ЛВС – О22).

Для описания поведения пользователей было выделено 18 параметров $\tilde{A} = \{\tilde{A}_j\}_{j=1}^{18}$, представленные в таблице 1. Функции принадлежности данных параметров были построены с участием эксперта на основе анализа статистики. Результаты экспериментальных исследований показали, что в качестве формы функций принадлежности более предпочтительно использовать кусочно-линейную форму.

Таблица 1 - Параметры пользователей

Параметр	Описание параметра (временной промежуток - сутки)
$\tilde{A}_1$	Количество обращений к почтовому серверу
$\tilde{A}_2$	Количество обращений к файловому серверу Н
$\tilde{A}_3$	Количество обращений к файловому серверу Z
$\tilde{A}_4$	Количество обращений к коммутатору
$\tilde{A}_5$	Количество обращений к сетевому принтеру «Бухгалтерия»
$\tilde{A}_6$	Количество обращений к сетевому МФУ «Секретариат
$\tilde{A}_7$	Количество обращений к сетевому МФУ «Менеджеры»
$\tilde{A}_8$	Количество обращений к сетевому МФУ «Тех.отдел»
$\tilde{A}_9$	Количество обращений к прокси-серверу
$\tilde{A}_{10}$	Количество принятых, отправленных писем через Microsoft Office Outlook
$\tilde{A}_{11}$	Количество обращений к «1С:Бухгалтерия 8»
$\tilde{A}_{12}$	Количество обращений к «1С:Документооборот»
$\tilde{A}_{13}$	Количество обращений к «1С:Предприятие 8»
$\tilde{A}_{14}$	Количество обращений к «Microsoft Navision 3.60»
$\tilde{A}_{15}$	Количество обращений к «1С:Зарплата и управление персоналом 8»
$\tilde{A}_{16}$	Количество обращений к модулю «Монитор сопровождения»

$\tilde{A}_{17}$	Объем внешнего сетевого трафика
$\tilde{A}_{18}$	Средний процент загруженности центрального процессора

На основании анализа поведения пользователей по выше перечисленным параметрам, осуществлялась кластеризация и распределение пользователей по кластерам. Результаты кластеризации представлены в таблице 2.

Таблица 2 - Результаты кластеризации

Кластер	Пользователь
C1	O1
C2	O2
C3	O3
C4	O4, O5
C5	O6, O7, O8
C6	O9, O10, O11, O13
C7	O12
C8	O14, O15, O16, O18, O19, O20, O21
C9	O17
C10	O22

Анализируя таблицу 2, можно сделать вывод, что программа сформировала 10 кластеров. Первые три кластера описывают действия пользователей руководящего состава генерального директора, финансового директора и технического директора соответственно. Кластер C4 описывает поведение объектов O4 и O5 – пользователи секретарь и офис-менеджер. В связи с особенностями работы администратора ЛВС, его действия в КИС отличны от действий других пользователей. Поэтому для администратора ЛВС (O22) был создан отдельный кластер. Так же были выделены кластеры, характеризующие поведение пользователей, входящих в различные структурные подразделения организации: бухгалтерия – C5, отдел продаж – C6, технический отдел – C8.

Пользователи, относящиеся к кластерам C7 и C9 были выделены в отдельные кластеры, что свидетельствует об их аномальном поведении. Проведя детальный анализ, было выявлено, что объем внешнего сетевого трафика объекта O12 превышает показатели трафика других пользователей отдела продаж, что и формирует подобную аномалию. Пользователь O17 обращался к программам, использование которых не является необходимым при выполнении функциональных обязанностей сотрудника технического отдела, что также сформировало соответствующую аномалию. Выявление подобных инцидентов позволило администратору безопасности своевременно отреагировать на них.

Вторая практическая задача, решенная с помощью разработанного метода, состояла в распределении животных по кластерам на основе нечетких значений их параметров. Для каждого семейства животных (медвежьи, зайцевые, кошачьи) взята выборка по 7 видов - $O = \{O_i\}_{i=1}^{21}$:

– Медвежьи (Большая панда – О1; Очковый медведь – О2; Бурый медведь – О3; Черный

медведь – О4; Белый медведь – О5; Гималайский медведь – О6; Губач – О2);

– Зайцевые (Лазяющий заяц – О8; Бушменов заяц – О9; Полосатый заяц – О10; Заяц-русак – О11; Заяц-беляк – О12; Калифорнийский кролик – О13; Дикий кролик – О14);

– Кошачьи (Гепард – О15; Обыкновенная рысь – О16; Пума – О17; Канадская рысь – О18; Дымчатый леопард – О19; Леопард – О20; Ирбис – О21).

Каждое животное было описано 3 параметрами в нечетком виде (длина тела, вес, скорость) – $A = \{A_j\}_{j=1}^3$.

В результате работы метода нечеткой концептуальной кластеризации, объекты были распределены по 3 кластерам в соответствии с семействами животных. При этом разработанный метод показал 100% точность кластеризации. Для сравнительного анализа, данная задача также была решена с помощью известных методов кластеризации EM и g-means. При этом выполнялась дефазификация параметров объектов, заданных в нечетком виде. Точность кластеризации определялась как отношение количества правильно проклассифицированных объектов к общему их числу. Точность решения задачи кластеризации методами EM и g-means составила, соответственно, 80,9% и 76,1%, что меньше точности, полученной в результате работы разработанного метода.

Таким образом, решенные практические задачи и результаты экспериментов показали, что разработанный метод позволяет эффективно осуществлять концептуальную кластеризацию объектов, характеризуемых нечеткими параметрами. Применение данного метода позволяет работать с объектами нечеткой природы, а также повысить точность кластеризации по сравнению с рядом известных методов.

Литература

1. Емалетдинова Л.Ю., Катасёв А.С., Кирпичников А.П. Нейронечеткая модель аппроксимации сложных объектов с дискретным выходом // Вестник Казанского технологического университета – 2014. – Т.17, № 1. – С. 295-299.
2. Кирпичников А.П., Осипова А.Л., Ризаев И.С. Повышение аналитических возможностей баз данных // Вестник технологического университета – 2012. – № 3. – С. 157-160.
3. Sato M., Sato Y., and Jain L. Fuzzy Clustering Models and Applications, Physica-Verlag, Heidelberg, 1997. – P. 135-148.
4. Вятчинин Д. А. Нечеткие методы автоматической классификации: Монография. – Мн.: УП «Технопринт», 2004 – 219 с.
5. Fisher D. Knowledge Acquisition Via Incremental Conceptual Clustering, 1987. – P. 142-153.
6. Назаров А.О., Аникин И.В. Распознавание поведения объектов методом нечеткой кластеризации данных // Вестник Казанского государственного технического университета им. А.Н. Туполева. - 2012. - № 4(1). - С.222-228.
7. Назаров А.О. КНЗ-1 // Свидетельство № 2013614934 о государственной регистрации программы для ЭВМ.
8. Назаров А.О., Аникин И.В. Формирование эталонных профилей поведения пользователей в корпоративных информационных системах // Вестник Казанского государственного технического университета им. А.Н. Туполева. - 2012. - № 3. - С.138-142.
9. Девянин П.Н. Модели безопасности компьютерных систем: Учеб. пособие для студ. высш. учеб. заведений. — М.: Издательский центр «Академия», 2005. — 144 с.
10. Аникин И.В. Технология интеллектуального анализа данных для выявления внутренних нарушителей в компьютерных системах // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. – 2010. – Т.6, № 113. – С. 112-117.
11. Аникин И.В., Лукоянов В.С. Метод и программный комплекс выявления внутренних нарушителей в корпоративных информационных сетях // Вестник Казанского государственного технического университета им. А.Н. Туполева. – 2009. - № 3. – С. 78-83.

© **И. В. Аникин** – канд. техн. наук, зав. каф. систем информационной безопасности КНИТУ-КАИ, anikinigor777@mail.ru; **А. П. Кирпичников** – д-р физ.-мат. наук, зав. каф. интеллектуальных систем и управления информационными ресурсами КНИТУ, kirpichnikov@kstu.ru; **А. О. Назаров** – соискатель каф. систем информационной безопасности КНИТУ-КАИ, sas4406@yandex.ru.

© **I. V. Anikin** - associate professor, KNRTU-KAI, anikinigor777@mail.ru; **A. P. Kirpichnikov** – Prof. KNRTU, kirpichnikov@kstu.ru; **A. O. Nazarov** - the competitor KNRTU-KAI, sas4406@yandex.ru.