УДК 519.7.007.004

**А. В. Ляховец**
Харьковский национальный университет радиоэлектроники
Проспект Ленина, 14, 61166 Харьков, Украина

# Кластеризация с помощью нейронной сети Кохонена и модифицированного алгоритма иерархической кластеризации Хамелеон в различных предметных областях

*Представлены результаты экспериментов и работы нейронной сети Кохонена для кластеризации клиентов банка и модифицированного алгоритма иерархической кластеризации Хамелеон для кластеризации пациентов со спинальным поясничным стенозом. Приведен анализ исследуемых алгоритмов кластеризации и их сравнение. Описаны экспериментальные выборки и результаты экспериментов.*

***Ключевые слова:*** *кластеризация, модифицированный алгоритм Хамелеон, графы, нейронная сеть Кохонена, кластеризация клиентов банка, кластеризация больных спинальным поясничным стенозом.*

С ростом объемов данных наблюдается рост потребности в интеллектуальном анализе данных (Data Mining). Существует множество задач анализа данных, применяемых в различных сферах человеческой деятельности. Одной из таких задач, требующей решения и оптимизации, является задача кластеризации.

Существует много различных методов кластеризации, которые по способу обработки данных делятся на иерархические методы (агломеративные методы AGNES — Agglomerative Nesting, дивизионные методы DIANA — Divisive Analysis) и неиерархические методы. Значимая часть неиерархических методов — это итеративные методы, которые по способу анализа данных делятся на четкие и нечеткие. По количеству применений алгоритмов кластеризации методы делятся на методы с одноэтапной кластеризацией и с многоэтапной кластеризацией. Методы по возможности расширения объема обрабатываемых данных делятся на масштабируемые и немасштабируемые. По времени выполнения кластеризации делятся на потоковые (on-line) и не потоковые (off-line) [1].

Среди всех этих методов для решения каждой конкретной задачи должен выбираться наиболее подходящий, основываясь на достоинствах и недостатках этих методов. Выбор оптимального метода позволит получить разделение наилучшего качества за наименьшее время для конкретной задачи и поставленных требований.

Целью данной работы является изучение, анализ и применение на реальных данных таких алгоритмов кластеризации как самоорганизующиеся карты Кохонена и модифицированный алгоритм Хамелеон, обоснование применения алгоритмов для решения конкретных практических задач и анализ результатов кластеризации.

## Анализ и сравнение методов кластеризации данных

В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие базы данных. В них основное внимание уделяется масштабируемости. Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами. К наиболее актуальным алгоритмам относятся: BIRCH, CURE, Хамелеон, ROCK, Кохонен [1, 2]. Сравнение данных методов представлено в таблице, где знаками «+» и «–» обозначено наличие или отсутствие описываемой характеристики, значение «+–» указывает на то, что в некоторых ситуациях характеристика присутствует, а в некоторых нет, что может зависеть от выборки, настроек алгоритма или других факторов.

Сравнение алгоритмов кластеризации

| | *BIRCH* | *CURE* | *Хамелеон* | *ROCK* | *Кохонен* |
|---|---|---|---|---|---|
| Большие объемы данных | + | + | + | + | + |
| Устойчивость к шуму | + | +– | +– | – | + |
| Масштабируемость | + | – | + | + | + |
| Определяет количество кластеров | + | – | + | – | +– |
| Кластеры произвольного размера и плотности | – | +– | + | – | +– |

На основании выполненного анализа существующего состояния развития методов и алгоритмов кластеризации и анализа входных данных можно сделать вывод, что реальные данные сильно отличаются по характеристикам исследуемой выборки, и для оптимального анализа необходимо обрабатывать различные выборки разными методами.

Искусственные нейронные сети, а именно карты Кохонена, являются конкурентами алгоритмов, построенных на графах, в частности модифицированного алгоритма Хамелеон, рассматриваемого в данной работе. Существенными недостатками карт Кохонена является необходимость предопределенности базового количества классов, а также зависимость от случайной инициализации нейронов и порядка обучающего множества. Последний указанный недостаток может привести к некорректности расположения классов в пространстве при разной плотности и удаленности классов [3]. Перечисленные недостатки карт Кохонена могут иметь большое значение при выборе метода кластеризации для определенных задач и показывают преимущество алгоритма Хамелеон, не обладающего ими.

## Кластеризация и классификация клиентов банка нейросетевыми методами

Для маркетингового исследования банка крайне важно знать свой сегмент рынка, в зависимости от этого формируется направление развития стратегии банка, строятся маркетинговые планы.

На основе нейронных сетей решается задача кластеризации для неоднозначно заданного числа классов. Требованиями к методу для решения задачи кластеризации в данном случае является возможность определения и уточнения количества классов, возможность обучения без учителя и возможность обучения на небольшой обучающей выборке. Для решения данной задачи с представленными требованиями наиболее приемлемым решением будет выбор самоорганизующихся карт (Self Organizing Maps — SOM) Кохонена, обучающихся без учителя [4].

Кластеризация проводилась на выборке из 144-х анкет клиентов банка, в каждой из которых присутствовало 12 наиболее важных для банка характеристик — личные данные, социальное положение, финансовая информация, информация об отношениях с банком и информация о предоставленной банком услуге. При решении задачи кластеризации произведено разбиение на классы.

Проанализировав результаты кластеризации, проведенной с помощью нейронной сети Кохонена, выявлено 4 кластера. Каждый кластер является отдельным сегментом рынка банковских услуг кредитования. В соответствии со статистическими характеристиками значений использованных переменных в каждом полученном кластере, предложены следующие названия каждого сегмента:

Сегмент 1 — «клиент ненадежный со средним достатком»;

Сегмент 2 — «клиент ненадежный, обеспеченный»;

Сегмент 3 — «клиент надежный со средним достатком»;

Сегмент 4 — «клиент средней надежности, малообеспеченный».

Вышеперечисленные результаты работы программы после интерпретации и анализа являются основоположными данными для предоставления определенных услуг и условий кредитования различным группам клиентов банка [5, 6]. Данные действия могут быть выполнены с помощью веб-версии реализации метода, что может быть существенным преимуществом в данной области [7].

Предложенная система была модифицирована для оценки целесообразности кредитования студента/абитуриента, основываясь не только на данных о клиенте, но и учитывая данные об учебном заведении и выбранной специальности. Анализ этих групп данных разными методами и применение не только скоринговых моделей, а и интеллектуальных методов, позволит увеличить точность оценки [8].

Для решения поставленной задачи недостатки карт Кохонена не были существенными, но метод не является универсальным при других исходных данных или условиях анализа.

## Исследование результатов применения модифицированного алгоритма Хамелеон в области лечения поясничного стеноза

Развитие современной медицины невозможно без внедрения в клиническую практику прогнозирования результатов лечения, которое дает возможность объек-

тивного выбора лечебной тактики, оценки эффективности и экономического обоснования целесообразности того или иного метода терапии, а также повышает надежность планирования ресурсов здравоохранения.

Сложность и многочисленность имеющихся в арсенале врача методов и подходов для лечения болезней позвоночника, разнообразие и индивидуальность вариантов течения дегенеративных заболеваний позвоночника, многочисленность анализируемых факторов и симптомов, отсутствие систематизированной оценки результатов требует привлечения математических методов, которые позволяют анализировать имеющийся материал и результативно использовать его для принятия решений при лечении конкретного пациента.

Целью исследования в данной области является постановка оптимальных методов и алгоритмов в рамках модифицированного алгоритма Хамелеон для оптимальной обработки выборки данных с характеристиками пациентов со стенозом поясничного канала и анализа полученных классов. Для достижения цели исследований необходимо решение следующих задач: подготовка экспериментальных данных, подготовка и выбор оптимальной конфигурации модифицированного алгоритма Хамелеон и анализ результатов исследования.

Были изучены результаты лечения 143-х пациентов, которые с 1997 по 2010 годы находились на стационарном лечении в отделении вертеброгии ИППС им. М.И. Ситенко и у которых клинико-рентгенологически диагностировался стеноз поясничного отдела позвоночного канала.

Исследование выполнено на основной группе из 143-х больных, консервативная терапия у 30-ти больных дала положительный результат, у 40-а — отрицательный, хирургическое лечение дало положительный результат для 38-ми пациентов и отрицательный для 35-ти.

Начальная выборка состояла из 32-х характеристик. Поиск специфических прогностических признаков, несущих прогностическую информацию, осуществлен при помощи различных статистических методов. Первым использованным методом для выделения наиболее важных признаков был метод ConsistencySubsetEval с применением различных поисковых методов (BestFirst, GeneticSearch, GreedyStepwise). В результате работы данных методов были выделены две наиболее важных характеристики — индекс дестабилизации (Destabilization index) и индекс массы тела (BMT).

Для анализа значимости всех атрибутов был использован метод InfoGainAttributeEval с использованием поискового метода Ranker. Результатом работы данного метода была градация всех атрибутов в соответствии с их уровнем значимости. В результате работы данного метода были выбраны атрибуты для дальнейшего анализа: индекс массы тела (BMT), индекс дестабилизации (Destabilization index) и щелочная фасфатоза (Alkalinephasfatoz).

Разделение данных будет производиться с помощью модифицированного алгоритма Хамелеон — нового иерархического алгоритма, который преодолевает ограничения существующих алгоритмов кластеризации. Данный алгоритм рассматривает динамическое моделирование в иерархической кластеризации.

В алгоритме можно выделить несколько этапов. Первый этап — построение графа, Хамелеон представляет объекты посредством часто используемого графа

*k*-ближайших соседей (*k*-nearest neighbor graph). На следующем шаге строится очередь из последовательно уменьшенных гиперграфов — стадия огрубления (Coarsening Phase). На третьей стадии выполняется разделение огрубленного графа таким образом, чтобы было удовлетворено ограничение баланса и оптимизирована функция разделения. На четвертом шаге выполняется восстановление графа. Разделение огрубленного графа проецируется на следующий уровень исходного графа, и выполняется алгоритм улучшения разделения (partitioning refinement algorithm). На последней итерации Хамелеона определяется показатель схожести между каждой парой кластеров. На основании данной меры наиболее близкие кластеры объединяются.

В процессе выбора конфигурации для модифицированного алгоритма Хамелеон было рассмотрено 15488 комбинаций алгоритмов [9].

Для оптимальной работы с модифицированным алгоритмом Хамелеон были выбраны оптимальные комбинации алгоритмов для разных типов выборок, в частности, для исследуемой медицинской выборки. Комбинации алгоритмов выбирались на основании критериев времени выполнения и качества кластеризации, которое оценивалось по 5-ти различным показателям. Анализ результатов кластеризации позволил интерпретировать полученные результаты:

— в отдельный класс были вынесены пациенты со средней и тяжелой тяжестью течения заболевания возрастом от 70 лет. Данной группе пациентов должно быть предложено щадящее консервативное лечение;

— группу больных, определяющими характеристиками которой было наличие радикулярного дефицита с суммой баллов больше 4 по шкале *Z*, следует наблюдать на наличие изменений в течении периода до двух лет, предоставляя консервативное лечение;

— большая группа больных, характеризуемая не ярко выраженными положительными значениями показателей и небольшим сроком течения заболевания, определена как группа пациентов, которой должно быть предложено хирургическое лечение;

— в отдельный класс были определены пациенты с существенным сроком течения заболевания и наличием неврологических нарушений и пареза, как первого симптома неврологического дефицита. Анализ данного результата может свидетельствовать о том, что со временем повышается вероятность появления таких нарушений.

Этот подход позволяет решить задачу анализа выборки медицинских данных более эффективно в сравнении с другими методами на основании критериев времени выполнения алгоритма и качества кластеризации, а также позволяет решать другие задачи со своими специфичными проблемами и потребностями.

## Выводы

В данной статье представлен обзор методов кластеризации данных. Проанализированы достоинства и недостатки наиболее актуальных методов. Приведены примеры работы методов карт Кохонена и модифицированного алгоритма кластеризации Хамелеон на примерах реальных данных. Приведены результаты кластеризации клиентов банка с помощью нейронной сети Кохонена и результаты клас-

теризации и анализа результатов выборки пациентов со спинальным поясничным стенозом с помощью модифицированного алгоритма кластеризации Хамелеон. Приведены достоинства и недостатки методов и специфика их применения. На примере исследования и кластеризации двух различных выборок реальных данных показана целесообразность применения конкурирующих методов в зависимости от специфики задачи. Таким образом, проведен анализ, представлены эксперименты на реальных данных и проанализированы результаты экспериментов, что соответствует поставленным целям работы.

В дальнейшем планируется изучение методов кластеризации, которые могут быть использованы для модификации алгоритма Хамелеон, проведение экспериментов для выявления зависимостей целесообразности использования вновь добавленных методов от статистических характеристик входных данных.

1. *Ляховец А.В.* Экспериментальные результаты исследования качества кластеризации разнообразных наборов данных с помощью модифицированного алгоритма Хамелеон / А.В. Ляховец // Вестник Запорожского национального университета. — 2011. — № 2. — С. 86–73.

2. *George Karypis.* Chameleon: Hierarchical Clustering Using Dynamic Modeling / George Karypis, Eui-Hong (Sam) Han, Vipin Kumar // Computer. — 1999. — Vol. 32, N 8. — P. 68–75.

3. *Николаев А.Б.* Нейросетевые методы анализа и обработки данных: учеб. пособ. / А.Б. Николаев, И.Б. Фоминых. — М.: МАДИ (ГТУ), 2003. — 95 с.

4. *Kohonen T.* Self-Organizing Maps / T. Kohonen. — Second Edition, Berlin: Springer-Verlag, 1997.

5. *Ляховец А.В.* Кластеризация клиентов банка / А.В. Ляховец, Н.С. Лесная, С.Д. Маковецкий // Сборник тезисов докладов по материалам 10-й юбилейной международной научной конференции «Теория и техника передачи, приема и обработки информации». Часть 2. — ХНУРЭ. — 2004. — С. 301–302.

6. *Лесная Н.С.* Методы кластеризации и классификации клиентов банка и их реализация с помощью нейронных сетей / Н.С. Лесная, А.В. Ляховец // Научно-технический журнал «Бионика интеллекта». — Харьков: ХНУРЭ, 2005.

7. *Lyakhovets Alyona.* Internet Oriented Crediting Process Automation System / Lyakhovets Alyona, Repka Victoriya, Vystavnyi Vitalii // Proc. of the Internation. Conf. for Internet Technology and Secured Transactions (ICITST 2007). — E-Centre for Informics, UK, 2007. — P. 69–75.

8. *Лесная Н.С.* Модель системы поддержки процессов кредитования образования / Н.С. Лесная, А.В. Ляховец // Научно-технический журнал «Вестник НТУ ХПИ» Тематическое издание «Системный анализ, управление и информационные технологии». — Харьков: ХПИ, 4'2009. — С. 36–45

9. *Ляховец А.В.* Исследование результатов применения модифицированного алгоритма хамелеон в области лечения поясничного стеноза / А.В. Ляховец // Восточно-европейский журнал передовых технологий. — 2012. — № 3/11(57). — С. 13–16.

Поступила в редакцию 30.01.2013