АННОТАЦИЯ

Диссертационной работы Черикбаевой Ляйли на тему: «Разработка и исследование оптимальных алгоритмов групповых решений в задачах распознавания», представленной на соискание степени доктора философии (PhD) по специальности 6D070300 – Информационные системы

Актуальность темы исследования. Задача распознавания образов состоит в классификации объектов по нескольким классам (образам). Каждый объект характеризуется конечным набором признаков. В общей постановке задачи классы известны для всех объектов выборки, для распознавания подаются новые объекты, для которых требуется определить к какому классу они принадлежат (распознавание «с учителем», Supervised learning).

В данной работе рассматривается один из вариантов постановки задачи распознавания образов - задача полуконтролируемого обучения (Semi-supervised learning). В этой задаче для некоторых определенных объектов исходной выборки классы известны, а для остальных неизвестны. Эта задача актуальна по следующим причинам:

- неразмеченные данные доступны;
- размеченные данные часто бывает сложно получить;
- использование неразмеченных данных совместно с небольшим количеством размеченных может обеспечить значительный прирост качества обучения.

Существует множество алгоритмов и подходов к решению задачи полуконтролируемого обучения. Цель данной работы заключается в разработке нового подхода для решения задачи полуконтролируемого обучения, его теоретическом и экспериментальном обосновании. Новизна работы состоит в сочетании алгоритмов коллективного кластерного анализа и ядерных методов классификации.

Задачей кластерного анализа является разбиение выборки на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер представлял группу похожих объектов, а объекты в разных кластерах существенно различались. Решение задачи кластеризации не является однозначным по нескольким причинам:

- Не существует наилучшего критерия качества кластеризации. Известно большое число разумных эвристических критериев и алгоритмов, не имеющих явно заданного критерия, но осуществляющих достаточно качественную кластеризацию;
- Число кластеров очень часто неизвестно заранее и устанавливается либо вручную, либо в ходе работы алгоритма;
- Результаты кластеризации очень сильно зависят от метрики, которая выбирается экспертом и специфики прикладной области.

Кроме того, алгоритмы кластерного анализа не универсальны: каждый алгоритм имеет свою специфическую область применения. Например, некоторые алгоритмы направлены на «шарообразные» структуры данных, другие на

«ленточные» кластеры и т.д.

На основании этих особенностей был предложен коллективный подход к кластерному анализу. В настоящее время коллективный подход показывает наилучшие результаты по сравнению с отдельными алгоритмами и позволяет использовать преимущества и особенности сразу нескольких алгоритмов.

В системах интеллектуального анализа данных особое место занимает проблема классификации, поскольку необходимость в проведении классификации объектов встречается при решении широкого круга прикладных задач: при анализе кредитного риска, в медицинской диагностике, при распознавании рукописных символов (почерка), при категоризации текстов, при извлечении информации и т.п. Не менее актульно проведение качественной классификации данных в технических системах, например, при обработке изображений, полученных при дистанционном зондировании, при идентификации различных объектов (пешеходов, лиц и т.п.). На практике довольно часто приходится проводить декомпозицию объектов в случае, когда об их внутренних связях ничего неизвестно и заранее неизвестна группировка объектов, на основе которой можно определить принципы для их разделения. В таком случае в качестве первой задачи анализа, требующей решения, можно рассматривать задачу кластеризации, предполагающую выполнение обучения без учителя в машинном обучении с целью выявления внутренней структуры в данных. Следует отметить, что не существует универсальных алгоритмов и методов классификации кластерного анализа. Более того, применение различных алгоритмов классификации к одному и тому же набору объектов может привести к различным результатам. Это связано с тем, что в основу этих агоритмов заложены различные принципы классификации и используемые в них метрики, функции близости, критерии оптимальности, способы выбора начальных параметров и т.п. В связи с этим возникает необходимость в получении результирующего классификационного решения, объединяющего результаты разбиений, полученные при реализации нескольких декомпозиционных алгоритмов, и совершающего меньшее число ошибок, чем каждый из этих алгоритмов.

Существует несколько вариантов получения группового решения задачи кластерного анализа, в предлагаемой в данной работе используется усредненная матрица коассоциации и метод, основанный на выделении центральных объектов.

В настоящее время алгоритмом групповых решений занимаются ученые: Журавлев Ю.И., Рязанов В.В., Лбов Г.С., Бирюков А.С. (г. Москва), Мазуров В.Д. (г. Свердловск), Ивахненко А.Г. (г. Киев), [Айдарханов М.Б.], Мухамедгалиев И.А., Дюсембаев А.Е, Амиргалиев Е.Н. (г. Алматы) и другие. Задачей полуконтролируемого обучения: Загоруйко Н.Г., Пестунов И.А., Бериков В.Б. (г. Новосибирск).

Цель диссертационной работы: Целью диссертационной работы является исследование и разработка теоретических и практических основ построения эффективных групповых решений распознавания и классификации, основанных на выделении центральных объектов (ядер), алгоритма для решения задач полуконтролируемого обучения и создание информационной системы распознавания и классификации.

Задачи исследования. Для достижения целей исследования решаются следующие вопросы:

1. Исследование и анализ алгоритмов групповых решений в задачах классификации и распознавания;

2. Разработка алгоритмов групповых решений в задачах классификации и распознавания:

а) Разработка алгоритма полуконтролируемого обучения и разпознавания в рамках постановки задачи групповых решений.

б) Алгоритм группового решения, основанный на выделении центральных объектов в группе базовых алгоритмов.

3. Анализ и оценка результатов алгоритмов групповых решений.

4. Создание информационной системы распознавания на основе предложенных методов групповых решений с применением современных средств проектирования и разработки;

Объект исследования. Набор объектов, пространство признаков, метрика близости, классы (кластеры), функционал качества, средства проектирования информационных систем.

Предмет исследования. Методы, алгоритмы и программные средства распознавания и классификации.

Методы исследования. Системный анализ и теория систем, теория графов, теория принятия решений, технологии разработки программного обеспечения.

Научная новизна: Новизна работы заключается в следующих научно обоснованных результатах, полученных в ходе диссертационного исследования как в рамках алгоритмов распознавания, так и в алгоритмах групповых решений.

1. Исследован и предложен алгоритм формирования классификатора для решения задач полуконтролируемого обучения, на основе совместного использования алгоритмов группового кластерного анализа и ядерных методов классификации, позволяющий повысить эффективность анализа сложноструктурированных, зашумленных данных большого объема за счет более точного выявления структуры данных с помощью алгоритмов кластерного анализа, в сочетании со способностью ядерных методов обнаруживать сложные нелинейные границы классов, а также за счет снижения трудоемкости и требуемой памяти с помощью малорангового представления матрицы ядра.

2. Исследован и разработан эффективный алгоритм групповых решений на базе предложенных алгоритмов распознавания и классификации, ориентированный на выделение эталонных (ядер) объектов, представляющий корректное решение задачи распознавания по группе выбранных функционалов качества;

Теоретическое и практическое значение работы. Теоретическая значимость данной работы заключается в совершенствовании разработанных алгоритмов групповых решений, основанных на выделении центральных объектов и сочетании алгоритмов группового кластерного анализа и ядерных методов классификации.

Практическая значимость работы заключается в следующем:

Разработанные алгоритмы групповых решений в задачах распознавания и

классификации и информационная система могут быть успешно применены для решения многих научных и прикладных задач в различных областях знаний.

Основное положение, выносимое на защиту. Эффективность разработанных на базе кластерного ансамбля алгоритмов групповых решений, основанных на полуконтролируемом обучении и на выделения эталонных объектов (ядер), теоретически обоснована и подтверждена вычислительными экспериментами, а реализованная оптимизационная модель в рамках информационной системы показала значимость в прикладных задачах.

Объем и структура работы. Диссертация состоит из введения, из 3 разделов и заключения, списка использованной литературы и приложения. Общий объем диссертации составляет 101 страниц, 47 рисунков, 4 таблиц. Список литературы состоит из 70 наименований.

Во введении рассмотрены актуальность темы диссертационной работы, цели, а также задачи для достижения поставленной цели. Описаны результаты, полученные до настоящего времени, их научная новизна и значимость. Здесь был представлен список статей, опубликованных в соответствии с темой диссертации.

В первой части представлены основные понятия и принципы методов и алгоритмов распознавания, критерии определения сходства объектов классификации. В работе рассматривались методы определения (увеличения) кластеров с ограничениями расстояний между точками объектов, способы и алгоритмы формирования кластеров по заданному количеству групп.

Во второй части были даны основные определения групповых решений, приведена постановка задач групповых решений в задачах распознавания и классификации, а также описаны методы построения групповых решений. Рассмотрены несколько концепций построения групповых решений. Разработанные алгоритмы групповых решений используют введенное понятие объектной структуры матрицы групповых решений. Приведены методы группового решения, основанные на выделении центральных объектов (ядер) – эталонов будущих классов и алгоритм групповых решений, основанный на использовании усредненной матрицы коассоциации. Рассмотрена постановка задачи распознавания образов - задача полуконтролируемой классификации (semi-supervised classification). В этой задаче лишь для части объектов исходной выборки известны метки классов; требуется классифицировать либо имеющиеся неразмеченные объекты, либо сформировать решающее правило для распознавания новых объектов. Исследован новый подход к решению данной задачи, основанный на сочетании алгоритмов коллективного кластерного анализа и ядерных методов классификации. Идея, лежащая в его основе, состоит в рассмотрении матрицы коассоциации, полученной кластерным ансамблем, как матрицы попарного сходства объектов и использовании этой матрицы как матрицы ядра (например, в методе опорных векторов). Такая замена имеет ряд оснований. Во-первых, можно полагать, что объекты из плотной области (кластера) в пространстве признаков с большей вероятностью имеют общие метки классов, даже если данная область имеет сложную форму. С этой точки зрения такие объекты более похожи друг на друга, чем другие точки, удаленные друг от друга на такое же расстояние, но из разных кластеров. Во-вторых, известно, что

усредненная матрица коассоциации определяет полуметрику на пространстве наблюдений, а значит, частоты отнесения пар объектов одним и тем же кластерам можно рассматривать как показатели сходства между соответствующими точками. При этом полученная матрица зависит от выходов алгоритмов кластеризации и является менее зависимой от случайных выбросов, чем обычная матрица сходства. В главе показано, что численные эксперименты на тестовых задачах и реальном гиперспектральном изображении демонстрируют эффективность предложенного метода, в том числе при наличии зашумленных данных.

Привлечение алгоритмов групповых решений позволяет повысить устойчивость результатов кластерного анализа в случае неопределенности в структуре данных. В этой главе показано, что целесообразность применения такого подхода была подтверждена экспериментальными результатами, свидетельствующими о том, что использование усредненной матрицы коассоциации в роли матрицы сходства во многих случаях существенно повышает качество решений.

В работе предлагается метод анализа гиперспектральных изображений на основе полуконтролируемого обучения. Основная идея состоит в разделении процесса обучения на два этапа. Вначале с помощью ансамбля алгоритмов кластерного анализа строятся варианты сегментации изображения. Далее вычисляется усредненная коассоциативная матрица. На втором этапе по размеченным пикселям строится решающая функция с применением алгоритмов обучения по сходству, на вход которого подается полученная матрица. Описан пример применения разработанного метода для анализа гиперспектральных изображений. Показано, что предложенный алгоритм более устойчив к шуму.

Основная задача алгоритмов групповых решений классификации состоит в построении оптимального результирующего разбиения исследуемого объектного множества во множество разбиений, полученные каждым алгоритмом, из базового набора алгоритмов. Понятие оптимальности конкретизируется для реальных задач классификации по выбранным функционалам качества. Показаны экспериментальные исследования с предложенными алгоритмами, и их результаты, а также сравнение полученных результатов с результатами известных алгоритмов.

В третьем разделе рассматриваются вопросы проектирования и реализации информационной системы. Приведена концептуальная схема информационной системы. Рассмотрены подсистема ввода и предварительной обработки данных, подсистема управления системой. Показаны диаграммы, построенные при создании информационный системы, подсистемы групповых решений. Система позволяет исследователям осуществлять разбиение определенного набора объектов на классы, согласно алгоритмам классификации и распознавания образов, в том числе алгоритмами групповых решений. Для решения задачи распознавания разработаны групповые методы. Разработаны и реализованы алгоритмы, как групповых решений, так и отдельные алгоритмы классификации.

Информационная система, включающая в себя разработанные и реализованные алгоритмы, представляет платформу, на которой для разных типов исходных данных (изображение, объекты, описываемые множеством признаков)

решаются конкретные задачи распознавания и классификации:

1. Исходные данные представлены в виде спутниковых изображений конкретной определенной местности (на нашем примере в качестве исходных данных взяты спутниковые изображения «Национальной Академии Наук» и «Университета имени Сулеймана Демиреля» с прилегающих территорий). Необходимо обработать данные изображения с целью разпознавания и классификации объектов этого изображения. Для получения улучшенного распознавания использован алгоритм группового решения. На базе полуконтролируемого обучения вычислительные эксперименты проведены для различных вариантов (с учетом зашумленности) изображений.

2. В качестве исходных данных предложены объекты (образцы), описанные набором признаков. В качестве объектов взяты пробы, полученные в гидрогеологических исследованиях из Чу Илийского региона. В качестве признаков взяты лабораторные данные о физико-химических свойствах взятой пробы – объекта.

В рамках информационной системы реализованы алгоритмы классификации (бустинг) для обработки исходных данных с целью получения наилучшего результата.

Предложенный алгоритм группового решения на базе обработки результатов отдельных алгоритмов классификации, входящих в групповой ансамбль, дает наиболее лучшее решение в смысле выбранных функционалов качества.

В заключении приводятся основные результаты и выводы данной диссертационной работы.

Уровень достоверности и результаты апробации. Полученные научные результаты подтверждены вычислительными экспериментами при решении реальных прикладных задач, что обусловливает высокую степень достоверности и обоснованности каждого научного результата, выносимых на защиту, а также сравнением эффективности полученных результатов с уже имеющимися результатами, полученными известными алгоритмами распознавания и классификации.

Результаты диссертации были обсуждены на научном семинаре факультета информационных технологий и кафедры информационных систем КазНУ им. аль-Фараби, а также на следующих научно-методических конференциях:

1. III Международная научно-практическая конференция «Информатика и прикладная математика», посвященной 80-летнему юбилею профессора Бияшева Р.Г. и 70-летию профессора Айдарханова М.Б. (Алматы, 26-29 сентября 2018 г.).

2. XIII Balkan Conference on Operational Research (BALCOR 2018), Сербия, Белград;

3. The 7 th International Conference on “Optimization Problems and Their Applications (OPTA-2018)”, Russia, 2018 г.

По результатам анализа и результатам выполнения диссертационной работы опубликовано 13 статей и получено 1 авторское свидетельство. Среди них 4 (четыре) статьи в изданиях, рекомендованных комитетом по обеспечению качества в сфере образования и науки МОН РК, 4 (четыре) статьи, включенных в базу «Scopus», 5 (пять) статей в материалах международных конференций.

Научные публикации.

1. Berikov V. B., Amirgaliyev Y.N., Cherikbayeva L.Sh, Yedilkhan D., Tulegenova B. “Classification at incomplete training information: usage of group clustering to improve performance” Journal of Theoretical and Applied Information Technology. - 2019. - Vol.97. - № 19. – P. 5048-5060 (Процентиль по базе Scopus - 33).

2. Amirgaliyev Y., Berikov V., Cherikbayeva L., Latuta K., Bekturgan K. “Group approach to solving the tasks of recognition” // Yugoslav Journal of Operations Research, - 2018. – Volume 2. – P. 177-192 (Scopus).

3. Sh. Shamiluulu, B. Y. Amirgaliyev, L. Cherikbayeva. “ Critical analysis of scikit-learn ml framework and weka ml toolbox over diabetes patients medical data ” // News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences. - 2017. - Volume 6. - Number 426. - P. 231 – 236 (Scopus).

4. Berikov V., Cherikbayeva L. Searching for Optimal Classifier Using a Combination of Cluster Ensemble and Kernel Method // Optimization Problems and Their Applications (OPTA-2018), CEUR Workshop Proceedings, Omsk, Russia, Vol. 2098, P. 45-60 (Scopus).

Статьи, опубликованные в издании, рекомендованных комитетом по обеспечению качества в сфере образования и науки МОН РК

5. Амиргалиев Е.Н., Шамиль-улу Ш., Черикбаева Л.Ш., Кеншимов Ч.А. “О некоторых численных результатах распознавания с машинным обучением” // Вестник КазНИТУ, – 2017. – №2 (120). – С. 386-391.

6. Черикбаева Л.Ш. “Классификациялау және кластерлеу әдістері” // Вестник КазНИТУ, – 2017. – №2 (120). – С. 158-161.

7. Черикбаева Л.Ш., Байсылбаева Қ.Д. “Өзгермелі арақашықтық метрикасы негізіндегі алгоритмдер”//Вестник КазНИТУ, - 2018. №2, – С. 99 - 103.

8. Черикбаева Л.Ш. “Алгоритмдердің топтық шешімдерін пайдалана отырып тиімді классификаторларды іздеу” // Вестник КазНИТУ, - 2019. №2, – С. 289 - 292.

Статьи, опубликованные в международных конференциях:

9. Kalimoldayev M., Amirgaliyev Y., Berikov V., Chericbayeva L., Latuta K., Kalybek uulu B. One approach for the group synthesis of recognition and classification tasks // XIII Balkan Conference on Operational Research (BALCOR 2018), Belgrade. – P. 400-407.

10. Бериков В.Б., Амиргалиев Е.Н., Черикбаева Л.Ш. Полуконтролируемое обучение на основе кластерного ансамбля // Материалы II Международной научной конференции «Информатика и прикладная математика» 27-30 сентября 2017 года, Алматы, Казахстан, (Часть II), С. 65-76.

11. Черикбаева Л.Ш., Калдыбекұлы Б. Кластерлік талдауда топтық шешудің тиімді параметрлерін таңдау алгоритмдері // Материалы III Международной научной конференции «Информатика и прикладная математика» 26-29 сентября 2018 года, Алматы, Казахстан, (Часть II), С. 42-47.

12. Викентьев А.А., Серов М.С., Бериков В.Б., Черикбаева Л.Ш., Тулегенова Б.А. “Коллективные расстояние для кластеризации множеств формул N-значной логики”. // Материалы IV Международной научно-практической конференции

«Информатика и прикладная математика» 25-29 сентября 2019 года, Алматы, Казахстан, (Часть I), С. 219-234.

13. Черикбаева Л.Ш., Концепции построения распознающих и классифирующих систем // «Көліктегі инновациялық технологиялар: білі, ғылым, тәжірибе» атты XLI Халықаралық ғылыми-практикалық конференцияның материалдары, 3-4 сәуір 2017 ж., Алматы, Казахстан, (I том), 117-119 б.

Авторское свидетельство «Software Semi-Supervised learning based on cluster ensenble» 20.03.2019 ж. №6373.