На правах рукописи

Шкаберина Гузель Шарипжановна

МОДЕЛИ И АЛГОРИТМЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ПРОДУКЦИИ

Специальность: 05.13.01 – Системный анализ,
управление и обработка информации
(космические и информационные технологии)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2020

Работа выполнена в ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева», г. Красноярск.

Научный руководитель: доктор технических наук, доцент **Казаковцев Лев Александрович**

Официальные оппоненты: **Еремеев Антон Валентинович**, доктор физико-математический наук, доцент, Омский филиал Института математики им. С.Л. Соболева СО РАН, главный научный сотрудник

Фридман Александр Яковлевич, доктор технических наук, профессор, Институт информатики и математического моделирования Федерального Исследовательского Центра «Кольский Научный Центр Российской академии наук», ведущий научный сотрудник

Ведущая организация: Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук»

Защита состоится «12» февраля 2021 года в 10:00 часов на заседании диссертационного совета Д 212.249.05, созданного на базе ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» по адресу: 660037, г. Красноярск, проспект имени газеты Красноярский рабочий, 31

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» и на сайте: https://www.sibsau.ru.

Автореферат разослан «____» _____ 2020 г.

Ученый секретарь
диссертационного совета,
канд. техн. наук, доцент Панфилов Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Современные средства сбора данных позволяют аккумулировать большие объемы многомерной информации об объекте исследования. Эта информация становится ценным источником знаний об объекте при ее обработке соответствующими методами и алгоритмами интеллектуального анализа. Задачи автоматической классификации называют задачами автоматической группировки, задачами кластерного анализа (обучение без учителя). Методы автоматической группировки (АГ) являются частью машинного обучения, актуальность которого возрастает с каждым годом.

АГ предполагает разделение множества объектов на подмножества (группы) так, чтобы объекты из одного подмножества были более похожи друг на друга, чем на объекты из других подмножеств по какому-либо критерию. Методы АГ на сегодняшний день применяются во многих сферах деятельности, например, в биологии, биоинформатике, медицине, маркетинге, в производстве при проверке качества промышленных изделий и т.д. Существуют отрасли с повышенными требованиями к качеству продукции, где решение задач АГ требует получения максимально точного и стабильного результата. Под точностью подразумеваем снижение доли ошибок АГ, а под стабильностью – повторяемость результата при многократных запусках алгоритма.

Многие широко используемые модели АГ являются моделями теории размещения. Для непрерывных задач размещения на сегодняшний день разработаны алгоритмы лишь для наиболее распространенных мер расстояния (евклидово, манхэттенское). Однако, учет особенностей пространства признаков конкретной практической задачи при выборе меры расстояния может привести к повышению точности АГ объектов.

Поиск алгоритма АГ объектов, обладающего одновременно высокой точностью и стабильностью результата, и при этом высокой скоростью работы, является одной из проблем АГ объектов. В представленной работе рассматриваем задачу автоматической группировки, а также задачу классификации. Диссертационная работа посвящена исследованию и разработке новых алгоритмов автоматической группировки объектов, которые позволяют повысить точность и стабильность результата решения практических задач.

Степень разработанности темы. Одной из наиболее известных моделей кластерного анализа является модель k-средних, которая была предложена Г. Штейнгаузом и в дальнейшем алгоритмически реализована С. Ллойдом. В работе Л. Кауфмана и П. Дж. Руссива представлены близкая модель k-медоид. Весомый вклад в задачи автоматической группировки и размещения объектов внесли Б. Дюран и П. Оделл. Ц. Дрезнером, Х. Хамахером, П. Хансеном, Ю. А. Кочетовым и Н. Младеновичем представлен метод поиска с чередующимися окрестностями для задач с большой размерностью данных.

О. Алпом, Э. Эркутом и Ц. Дрезнером, а в дальнейшем А.Н. Антамошкиным и Л.А. Казаковцевым предложены генетические и другие алгоритмы с жадной эвристической процедурой для задачи автоматической группировки на основе моделей теории размещения.

Задачи автоматической группировки на основе разделения смесей вероятностных распределений исследованы С. Ньюкомбом, К. Пирсоном, Б. Эвериттом, Д.М. Титтерингтоном, Дж. Маклаланом, В.Ю. Королевым, Дж. Гримом, О.К. Исаенко, В.Ю. Урбахом, А.П. Демстером, Н.М. Лейрда, Д.Б. Рубиным. Д.В. Сташков распространил подход Казаковцева-Антамошкина на такие задачи. В работах И.П. Рожнова и др. предложены алгоритмы поиска с чередованием жадных эвристических процедур для задач k-средних, k-медоид**.**

Улучшить результат АГ объектов с повышенными требованиями к точности и стабильности результата известными алгоритмами на текущий момент крайне трудно без значительного увеличения временных затрат. При решении практических задач АГ объектов, например, при решении задач выделения однородных партий промышленной продукции, вопросы вызывает адекватность моделей и, как следствие, точность АГ промышленной продукции. Таким образом, существует запрос на разработку новых моделей. Все же возможна также и разработка алгоритмов, обеспечивающих дальнейшее улучшение результата на основе выбранной модели, например, модели k-средних.

Объектом диссертационного исследования являются задачи автоматической группировки многомерных данных. **Предметом исследования** – модели и алгоритмы их решения.

Целью исследования является повышение точности и стабильности результата решения задач автоматической группировки объектов.

Поставленная цель достигается путем решения следующих **задач**:

1. сравнительный анализ известных алгоритмов решения задач автоматической группировки объектов с повышенными требованиями к точности разделения объектов на группы и стабильности результата при многократных запусках алгоритмов;
2. построение модели автоматической группировки объектов на основе модели k-средних с расстоянием Махаланобиса, а также с применением методов факторного анализа для предварительного снижения размерности исходных данных, которая позволяет снизить долю ошибок автоматической группировки промышленной продукции по сравнению с известными моделями;
3. разработка генетического алгоритма для задачи k-средних с перекрестной мутацией, позволяющего повысить точность и стабильность решения по достигаемому значению целевой функции за фиксированное время выполнения по сравнению с известными алгоритмами автоматической группировки объектов;
4. разработка алгоритма обучения двухслойной сигмоидальной искусственной нейронной сети с регуляризацией, повышающего точность решения задачи классификации промышленных изделий на однородные производственные партии при наличии малоинформативных признаков.

Научная новизна:

1. Предложена новая модель для решения задач автоматической группировки промышленной продукции на основе модели k-средних с расстоянием Махаланобиса с применением метода главных компонент. Применение новой модели позволяет повысить точность решения (индекс Рэнда) для задачи

выделения однородных производственных партий изделий по данным тестовых испытаний.

2. Предложен новый алгоритм автоматической группировки объектов, основанный на оптимизационной модели k-средних с мерой расстояния Махаланобиса и средневзвешенной ковариационной матрицей, рассчитанной по обучающей выборке. Алгоритм позволяет снизить долю ошибок (повысить индекс Рэнда) при выявлении однородных производственных партий продукции по результатам тестовых испытаний.

3. Разработан новый генетический алгоритм для задачи k-средних с применением единой жадной агломеративной эвристической процедуры в качестве оператора скрещивания и оператора мутации. Применение данного алгоритма позволяет статистически значимо повысить точность результата (улучшить достигаемое значение целевой функции в рамках выбранной математической модели решения задачи автоматической группировки), а также его стабильность, за фиксированное время, по сравнению с известными алгоритмами автоматической группировки.

4. Разработан новый алгоритм обучения двухслойной сигмоидальной искусственной нейронной сети с регуляризацией, демонстрирующий более высокую точность классификации промышленной продукции по данным тестовых испытаний в сравнении с методами обучения таких нейронных сетей при известных методах регуляризации.

Теоретическая значимость работы. Предложенный комплекс алгоритмов и моделей дополняет список эффективных методов решения задач АГ объектов. Принцип использования единой процедуры в качестве оператора скрещивания и мутации создает основу для синтеза новых эффективных алгоритмов для более широкого круга NP-трудных задач.

Практическая значимость работы. Предложенные модели автоматической группировки могут применяться для задач разделения объектов как при производстве и тестировании продукции, так и в других отраслях, где требуется классификация изделий с особыми требованиями качества. Использование предложенных алгоритмов решения задач автоматической группировки и классификации в различных системах, предназначенных для решения таких задач за фиксированное время, позволяет существенно повысить точность получаемых решений в рамках выбранной модели по получаемому значению целевой функции. Диссертационное исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания № FEFE-2020-0013 «Развитие теории самоконфигурирующихся алгоритмов машинного обучения для моделирования и прогнозирования характеристик компонентов сложных систем».

Методология и методы исследования. Методологической базой являются работы по методам кластеризации и классификации. Использованы методы системного анализа, исследования операций, теории оптимизации, прикладной статистики, корреляционного анализа, факторного анализа, интеллектуального анализа данных.

Положения, выносимые на защиту:

1. Модель автоматической группировки промышленной продукции на основе модели k-средних с расстоянием Махаланобиса с применением метода главных компонент в комбинации с отбором информативных признаков позволяет повысить точность решения (индекс Рэнда) для задачи выделения однородных производственных партий изделий по данным тестовых испытаний.

2. Алгоритм автоматической группировки объектов, основанный на оптимизационной модели k-средних с мерой расстояния Махаланобиса и средневзвешенной ковариационной матрицей, рассчитанной по обучающей выборке и учитывающей обобщенные корреляционные зависимости между признаками объектов в однородных группах, позволяет повысить точность решения (по индексу Рэнда) при выявлении однородных производственных партий продукции по результатам тестовых испытаний.

3. Генетический алгоритм для задачи k-средних с применением единой жадной агломеративной эвристической процедуры в качестве оператора скрещивания и оператора мутации, за счет увеличения разнообразия в популяции, повышает точность результата (улучшает достигаемое значение целевой функции в рамках выбранной математической модели решения задачи автоматической группировки), а также его стабильность, за фиксированное время, по сравнению с известными алгоритмами автоматической группировки.

4. Алгоритм обучения двухслойной сигмоидальной искусственной нейронной сети с регуляризацией позволяет достичь более высокой точности классификации промышленной продукции по данным тестовых испытаний по сравнению с методами обучения таких нейронных сетей с известными методами регуляризации.

Степень достоверности результатов. Достоверность результатов подтверждается корректным применением современных методов исследования, которые были применены в большом наборе экспериментов.

Апробация результатов. Основные положения и результаты докладывались на международных конференциях и семинарах: Mathematical Optimization Theory and Operations Research (MOTOR, 2019, г. Екатеринбург и 2020 г., Новосибирск), «Advanced Technologies in Material Science, Mechanical and Automation Engineering» (04-06 апреля 2019 г., Красноярск*)*, 2019 International Conference on Information Technologies (InfoTech, 2019 и 2020 гг., Болгария), 2019 International Russian Automation Conference (RusAutoCon, 2019 г., Сочи), «Advanced Technologies in Aerospace, Mechanical and Automation Engineering» - MIST: Aerospace (2019, г. Красноярск), Workshop on science of DataScience (2019, Trieste, Italy), Математические модели принятия решений (2020, Новосибирск), Математическое моделирование и дискретная оптимизация (2020, Омск), Проблемы математического и численного моделирования (2020, Красноярск), Математическое моделирование (2020, г. Москва).

Публикации. Основные теоретические и практические результаты диссертации изложены в 16 публикациях, среди которых 4 работы в ведущих рецензируемых журналах, рекомендуемых действующим перечнем ВАК, 11 – в международных изданиях, индексируемых в системах цитирования Web of

Science и Scopus. Имеется свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, 4 глав, заключения и приложений. Она изложена на 222 листах машинописного текста, содержит список литературы из 312 наименований.

Основное содержание работы

Во введении обоснована актуальность темы исследования, сформулированы цель и задачи исследования, основные положения научной новизны, теоретическая и практическая значимость работы, изложены методология и методы исследования, а также положения, выносимые на защиту.

Первая глава посвящена анализу текущего состояния проблем, связанных с АГ объектов, и методов их решения.

В процессе группировки объектов некоторого множества на определенные группы (подмножества) учитываются общие признаки объекта и методы, с помощью которых происходило разделение.

Для вектора наблюдений X алгоритм k-средних предназначен для определения k центров и назначения точек данных (объектов) каждому центру для формирования кластеров C_j, $j = 1..k$, при этом сводя к минимуму различие объектов внутри кластера. В работах А.К. Джаина и Дж. Маккуина базовый алгоритм k-средних состоит из итеративного повторения двух шагов:

Дано: k исходных кластерных центров (центроидов).

Шаг 1. Создание нового кластера C_j, назначением каждой точки данных ближайшему центру кластера (центроиду).

Шаг 2. Вычисление новых кластерных центров.

Повторение шагов 1 и 2, пока внутри каждого кластера изменения не прекратятся.

В алгоритме k-средних необходимо первоначально спрогнозировать количество групп (подмножеств). Кроме того, полученный результат зависит от начального выбора центров.

В отличие от алгоритма k-средних, в качестве центров в алгоритме k-медиан выступают медианы, а в алгоритме k-медоид – кластеризуемые объекты (медоиды). Во всех этих моделях могут применяться различные меры расстояния для расчета минимального расстояния между каждой точкой данных и ближайшим центроидом (медианой, медоидой). Функции расстояния и их определение играют важную роль в проблеме разделения исследуемого множества на группы. EM-алгоритм, как и алгоритм k-средних, основан на повторении двух шагов. EM-алгоритм, k-средних, k-медиан и k-медоид являются алгоритмами локальной оптимизации.

В своей работе О. Алп, Э. Эркут и Ц. Дрезнер представили простой генетический алгоритм с особой процедурой рекомбинации. С помощью процедуры объединения двух решений получаем решение с избыточным числом кластеров. Затем последовательно удаляем центры кластеров. Из рассмотрения удаляем тот центр, который незначительно ухудшает значение целевой функции.

В работах Ф. Гальтона, К. Пирсона, Ч. Спирмена, К. Хользингера, Л. Терстоуна, Р. Кателла, Д. Гилфорда основная цель факторного анализа сведена

к определению факторов, выявляющих взаимосвязи между входными характеристиками объекта, и понижении размерности входных характеристик. Мы используем факторную модель $X_i = \sum_{j=1}^{f} a_{ij}F_j + u_i$, где X_i – вектор значений i-й характеристики $(i = 1..M)$, F_j – первичные факторы $(j = 1..f)$, a_{ij} – коэффициенты, называемые факторными нагрузками, u_i – характерные (специфические) факторы, описывающие ту часть характеристики, которая не входит ни в один фактор.

В представленной работе в качестве меры точности кластеризации используем индекс Рэнда (Rand Index, RI), а для оценки качества алгоритмов классификации – количественный показатель AUC (area under the curve) ROC-кривой (receiver operating characteristic).

Анализ литературы показал, что существующие решения в области АГ объектов обладают либо высокой точностью, либо обеспечивают стабильность результата при многократных запусках алгоритма, либо высокой скоростью работы, но не всеми этими качествами одновременно. При этом на сегодняшний день разработаны алгоритмы лишь для наиболее распространенных мер расстояния. В области классификации существует проблема, связанная с точностью в условиях малого количества данных, большого количества признаков и наличия помех.

Предложенные модели и алгоритмы, изложенные в главах 2-4, позволяют повысить точность разделения объектов на группы и стабильность результата при многократных запусках алгоритмов.

Вторая глава посвящена проблеме понижения размерности данных с применением методов факторного анализа, и разработке оптимизационной модели АГ объектов, основанной на модели k-средних, позволяющей повысить качество АГ (по индексу Рэнда).

Для исследования эффективности практического применения ФА к задаче АГ объектов использованы данные об образцах промышленной продукции, принадлежность которых к одной из однородных партий заранее известна.

Значения факторов, полученные с помощью процедуры ортогонального вращения, рассматриваются как входные данные для алгоритмов кластеризации. Показано, что применение методов факторного анализа способствует снижению доли ошибок автоматической группировки при применении различных моделей кластеризации. Например, методом EM доля верно кластеризованных примеров составляет для выборки из трех партий 0,93, для выборки из двух партий 1.

Показано, что факторная модель позволяет уменьшить размерность исходных данных и повысить точность кластеризации объектов до 1 для задачи разбиения на две однородные партии, но с увеличением количества однородных партий точность кластеризации уменьшается. Но остается невозможным получить универсальный набор с небольшим количеством признаков для разделения смешанной партии, состоящей из произвольного числа однородных партий продукции: для надежного разделения методами кластерного анализа использование многомерных данных при решении задачи АГ объектов неизбежно. Методы ФА показали наличие линейных статистических

зависимостей (корреляций) между характеристиками объекта в однородной партии.

На сегодняшний день разработаны алгоритмы k-средних и k-медиан лишь для наиболее распространенных мер расстояния (евклидово, манхэттенское). Однако, учет особенностей пространства признаков конкретной практической задачи при выборе меры расстояния может привести к повышению точности АГ объектов. Задача АГ решается как задача k-средних в многомерном пространстве. Целью является нахождение k точек (центров) $X_1,\dots X_k$ в d-мерном пространстве, таких, чтобы сумма квадратов расстояний от известных точек (векторов данных) $A_1,\dots,A_N$ до ближайшей из искомых точек (центров) достигала минимума:

$$\arg\min F(X_1,\dots,X_k) = \sum \min_{j\in\{1,k\}} \left\|X_j - A_i\right\|^2. \qquad (1)$$

Использование корреляционных зависимостей может быть задействовано путем перехода от поиска в пространстве с евклидовой или манхэттенской метрикой к поиску в пространстве с метрикой Махаланобиса. Квадрат расстояния Махаланобиса D_M определяется следующим образом:

$$D_M(X) = (X-\mu)^T \mathrm{C}^{-1}(X-\mu), \qquad (2)$$

где X – вектор значений измеренных характеристик, μ – вектор средних значений (например, центр кластера), C – ковариационная матрица.

Результаты экспериментов по АГ промышленной продукции с моделями k-медоид и k-средних, в которых применена метрика Махаланобиса, показывают некоторое повышение точности кластеризации при АГ на 2-4 кластера и малом количестве объектов и информативных признаков. Тем не менее, применение метрики Махаланобиса в задаче АГ промышленной продукции во многих случаях (при использовании многомерных данных) приводит к снижению точности в сравнении с результатами, достигаемыми с евклидовой метрикой. Также эксперимент показал, что точность кластеризации k-средних с метрикой Махаланобиса с обучением (кластеризация сборной партии на две однородные партии при ковариационной матрице C, обученной по предварительно размеченным данным шести партий) выше в сравнении с метрикой Махаланобиса без обучения (кластеризация на две партии на ковариационной матрице, рассчитанной непосредственно по исходным данным). Но все же точность кластеризации остается ниже по сравнению с результатами, достигаемыми с евклидовой или манхэттенской метрикой.

Предложено вместо ковариационной матрицы из (2) по обучающей выборке рассчитать среднюю ковариационную матрицу для однородных партий изделий (по предварительно размеченным данным):

$$C = \frac{1}{N}\sum_{j=1}^{k} C_j\, n_j, \qquad (3)$$

где n_j – количество объектов (изделий) в j-й партии, N – общий размер выборки, C_j– ковариационные матрицы отдельных партий изделий.

В данной работе предложен алгоритм АГ объектов, основанный на оптимизационной модели k-средних с подстройкой параметра меры расстояния Махаланобиса (матрицы C) по обучающий выборке:

Алгоритм 2.1 k-средних с мерой расстояния Махаланобиса с усредненной оценкой ковариационной матрицы

Шаг 1. Методом k-средних с евклидовыми расстояниями разделить выборку на некоторое число k кластеров (здесь k – некоторая экспертная оценка возможного числа однородных групп, не обязательно точная).

Шаг 2. Для каждого кластера рассчитать центроид μ_i (вектор размерностью M). Центроид определяется как среднее арифметическое всех точек в кластере C_i, (i=1,..,k):

$$\mu_i = \frac{1}{m_i}\sum_{j\in C_i} X_j, \qquad (4)$$

где m_i – количество точек в i-м кластере, X_j – вектор значений измеряемой характеристики размерностью M.

Шаг 3. Рассчитать усредненную оценку ковариационной матрицы (3). Если усредненная оценка ковариационной матрицы является вырожденной, то перейти к Шагу 4, в противном случае перейти к Шагу 5.

Шаг 4. Увеличить число кластеров (k+1) и повторить шаги 1 и 2. Сформировать новые кластеры, используя квадрат Евклидова расстояния:

$$D\left(X_j, \mu_i\right) = \sum_{i=1}^{M}\left(X_j - \mu_i\right)^2, \qquad (5)$$

где M – количество характеристик. Вернуться к шагу 3 с новым обучающим примером (набором).

Шаг 5. Сопоставить каждый вектор данных ближайшему центроиду, используя квадрат расстояния Махаланобиса с усредненной оценкой ковариационной матрицы (3) для формирования новых кластеров.

Шаг 6. Повторять алгоритм с Шага 2, пока продолжают изменяться кластеры.

В работе представлены результаты трех групп экспериментов по данным образцов промышленной продукции: *Группа I*. Обучающая выборка соответствует рабочей выборке, для которой проводилась кластеризация; *Группа II*. Обучающая и рабочая выборки не совпадают. На практике в качестве обучающей выборки тестовый центр может использовать ретроспективные данные поставок и тестирования изделий того же типономинала; *Группа III*. Обучающая и рабочая выборка также не совпадают, но в качестве обучающей выборки использовались результаты работы автоматической группировки изделий (k-средних в режиме мультистарта с евклидовой метрикой). В каждой из групп экспериментов для каждой рабочей выборки алгоритм k-средних запущен по 30 раз с каждой из пяти исследуемых моделей кластеризации: Модель DM1 – k-средних с метрикой Махаланобиса, матрица C рассчитана для всей обучающей выборки; Модель DC – k-средних с метрикой, аналогичной расстоянию Махаланобиса, но используется корреляционная матрица вместо ковариационной матрицы; Модель DM2 – k-средних с метрикой Махаланобиса, с усредненной оценкой матрицы C; Модель DR – k-средних с манхэттенским расстоянием; Модель DE – k-средних с евклидовым расстоянием. Для каждой модели рассчитаны минимальное (Min), максимальное (Max), среднее значения (Average), среднеквадратичное отклонение (Std.Dev) индекса Рэнда (RI) и целевой функции, а также значения коэффициентов вариации (V) и размаха (R) целевой функции (таблица 1).

Установлено, что новая модель DM2 показывает лучшую точность среди представленных моделей практически во всех сериях экспериментов по индексу Рэнда (RI) и во всех случаях превосходит модель DE, где используется евклидово расстояние. Принимая во внимание более высокое среднее значение RI, оптимизационная модель DM2 АГ объектов по однородным партиям, основанная на модели k-средних с усредненной оценкой ковариационной матрицы, имеет преимущество перед моделями с евклидовой и манхэттенской мерами расстояния.

Таблица 1 – Результаты вычислительного эксперимента над данными микросхемы 1526IE10_002(3987 векторов данных размерностью 68), обучающая выборка из 10 партий, группа III, рабочая выборка из 7 партий изделий)

Серия V	индекс Рэнда (RI)					Целевая функция				
	Модель					Модель				
	DM1	DC	DM2	DR	DE	DM1	DC	DM2	DR	DE
Max	0,767	0,658	0,749	0,740	0,735	255886	379167	281265	18897	6494,62
Min	0,562	0,645	0,696	0,703	0,705	250839	36997	274506	17785	5009,42
Average	0,632	0,650	**0,725**	0,714	0,719	252877	37178	277892	18240	5249,95
Std .Dev	0,047	0,003	0,016	0,008	0,006	1164,5	152,8	2358,9	452,7	366,5
V						0,461	0,411	0,849	2,482	**6,981**
R						5047	920	**6759**	1112	1485

Также в экспериментах показано, что в большинстве случаев коэффициент вариации значений целевой функции выше для модели DE, где используется евклидова мера расстояния, и коэффициент размаха значений целевой функции имеет самые высокие значения для модели DM2, где используется мера расстояния Махаланобиса с усредненной оценкой ковариационной матрицы. Следовательно, для получения устойчиво хороших значений целевой функции требуются множественные попытки запуска алгоритма k-средних или использование других алгоритмов, основанных на модели k-средних, таких как j-means или же алгоритмов метода жадных эвристик.

Третья глава посвящена повышению эффективности алгоритма автоматической группировки в рамках выбранной модели. Результатом главы стала разработка генетического алгоритма с перекрестной мутацией для задачи k-средних. Новый алгоритм позволяет повысить точность решения задачи k-средних и стабильность результата за фиксированное ограниченное время выполнения. В данной главе под точностью алгоритма будем понимать исключительно достигнутую величину целевой функции, не учитывая показатели адекватности модели и соответствие результатов алгоритма фактическому (реальному) разделению объектов, если таковое известно.

Для генетических алгоритмов решения задачи k-средних с вещественным кодированием решений известно очень ограниченное множество возможных операторов мутации. Например, Маулик и др.[1] кодируют решения (хромосомы) в своих ГА как наборы центроидов, представленных их координатами (векторами действительных чисел) в многомерном пространстве. Каждая хромосома подвергается мутации с фиксированной вероятностью ω.

[1] Maulik, U. Genetic Algorithm-Based Clustering Technique / U. Maulik, S. Bandyopadhyay // Pattern Recognition Journal. – 2000. – Vol. 33(9). – P. 1455–1465. DOI: 10.1016/S0031-3203(99)00137-5.

Алгоритм 3.1 Процедура исходной мутации ГА для задачи k-средних

Шаг 1. Генерация случайного числа $b \in (0,1]$ с равномерным распределением.

Шаг 2. ЕСЛИ $b<\omega$, ТО хромосома мутирует. Если положение текущего центроида υ, то после мутации оно становится:

$$\upsilon \leftarrow \begin{cases} \upsilon \pm 2 \times b \times \upsilon, \upsilon \neq 0, \\ \upsilon \pm 2 \times b, \upsilon = 0. \end{cases}$$

Знаки «+» и «–» имеют одинаковую вероятность. Координаты центроида смещены случайным образом.

В своей работе мы заменили эту процедуру мутации для задачи k-средних следующей процедурой.

Алгоритм 3.2 Процедура перекрестной мутации ГА для задачи k-средних

Шаг 1. Генерация случайного начального решения $S = \{X_1 \dots X_k\}$.

Шаг 2. Применить алгоритм k-средних к S для получения локального оптимума S'.

Шаг 3. Применить простую перекрестную процедуру для мутированного индивида S' из популяции и S для получения нового решения S''.

Шаг 4. Применить алгоритм k-средних к S'' для получения локального оптимума S''.

Шаг 5. ЕСЛИ $F(S'')<F(S')$, **ТО** $S' \leftarrow S''$.

Предложенная процедура используется с вероятностью мутации равной 1, после каждого оператора скрещивания.

Результаты запуска исходного алгоритма 3.1, описанного с вероятностью мутации 0,01, и его версия с алгоритмом 3.2 в качестве оператора мутации представлена на рисунке 1 (численность популяции N_{POP} = 20). Новая процедура мутации является быстрой и эффективной по сравнению с исходной мутацией генетического алгоритма, показана высокая скорость сходимости целевой функции.

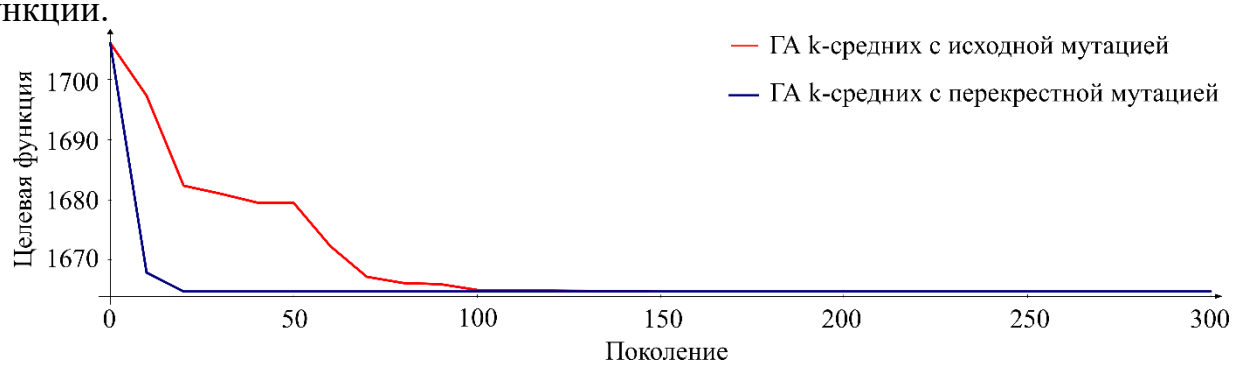

Рисунок 1 – Результаты для набора данных Mopsi-Joensuu (6014 векторов данных размерностью 2), 300 кластеров, ограничение времени 3 минуты

Далее проведено сравнение генетических алгоритмов с жадными эвристическими процедурами скрещивания, с их модификациями, которые включают перекрестные мутации с хромосомами, закодированными действительными числами (алфавит действительных чисел).

Генетические алгоритмы метода жадных эвристик и многие другие эволюционные алгоритмы для задачи k-средних, обходятся без мутации. Идея жадной агломеративной эвристической процедуры заключается в объединении

двух известных решений в одно недопустимое решение с избыточным числом центроидов и далее последовательно уменьшается количество центроидов. На каждой итерации удаляется центроид, удаление которого показывает наименьшее увеличение значения целевой функции (1).

Алгоритм 3.3 Базовая жадная агломеративная эвристическая процедура

Дано: начальное число кластеров K, необходимое количество кластеров k, $k<K$, начальное решение $S = \{X_1, \dots, X_K\}$, где $|S| = K$.

Шаг 1. Улучшить начальное решение алгоритмом k-средних.

Пока $K>k$

Цикл для каждого $i' \in \overline{\{1, K\}}$ выполнить:

Шаг 2. $S' \leftarrow S \backslash \{X_i'\}$. Улучшить решение S' алгоритмом k-средних и сохранить соответствующие полученные значения целевой функции (1), как переменные $F_{i'}' \leftarrow F\{S'\}$.

Конец Цикла

Шаг 3. Выбрать подмножество S_{elim} из центров n_{elim}, $S_{elim} \in S$, $|S_{elim}| = n_{elim}$ с минимальным значением соответствующих переменных $F_{i'}'$. $n_{elim} = max\{1, 0{,}2 \cdot (|S| - k)\}$.

Шаг 4. Получить новое решение $S \leftarrow S \backslash S_{elim}$; $K = K - 1$. Улучшить решение алгоритмом k-средних.

Конец Пока

Также начальное решение S можно получить объединением двух известных решений. Алгоритмы 3.4 и 3.5 модифицируют начальное решение вторым известным решением. Фактически Жадная процедура 1 дополняет первое множество поочередно каждым элементом из второго множества. Жадная процедура 2 объединяет оба множества.

Алгоритм 3.4 Жадная процедура 1 с частичным объединением

Дано: Два набора центров кластеров $S' = \{X_1', \dots, X_k'\}$ и $S'' = \{X_1'', \dots, X_k''\}$.

Цикл для каждого $i' \in \overline{\{1, K\}}$

Шаг 1. Объединить S' и один элемент из набора $S'': S \leftarrow S \cup \{X_{i'}''\}$.

Шаг 2. Запустить Алгоритм 3.3 с решением S и сохранить полученный результат.

Конец Цикла

Шаг 3. Вернуть лучшие решения, сохраненные на Шаге 2.

Алгоритм 3.5 Жадная процедура 2 с полным объединением множеств

Дано: Два набора центров кластеров $S' = \{X_1', \dots, X_k'\}$ и $S'' = \{X_1'', \dots, X_k''\}$.

Шаг 1. Объединить два набора центров кластеров $S \leftarrow S' \cup S''$.

Шаг 2. Запустить Алгоритм 3.3 с решением S.

Базовый генетический алгоритм (ГА, GA) для задач k-средних описывается следующим образом:

Алгоритм 3.6 ГА с алфавитом действительных чисел для задачи k-средних

Дано: Начальная численность популяции $N_{POP.}$

Шаг 1. Выбрать N_{POP} начальных решений $S_1, \dots, S_{N_{POP}}$, где $|S_i| = k$, и

$\{S_1, \dots, S_{N_{POP}}\}$ – случайно выбранное подмножество набора векторов данных. Улучшить каждое начальное решение алгоритмом k-средних и сохранить соответствующие полученные значения целевой функции (1), как переменные $f_k \leftarrow F(S_k),\ k = \overline{1, N_{POP}}$.

Цикл

Шаг 2. Если условие остановки выполнено, **тогда** СТОП. Вернуть решение $S_i, i \in \{\overline{1, N_{POP}}\}$ с минимальным значением f_i.

Шаг 3. Случайным образом выбрать 2 индекса $k_1, k_2 \in \{\overline{1, N_{POP}}\}$, $k_1 \neq k_2$.

Шаг 4. Запустить процедуру скрещивания: $S_C \leftarrow Crossingover(S_{k_1}, S_{k_2})$.

Шаг 5. Запустить процедуру мутации: $S_C \leftarrow Mutation(S_C)$.

Шаг 6. Запустить выбранную процедуру отбора для изменения набора популяции.

Конец Цикла

На Шаге 6 предложен следующий алгоритм:

Алгоритм 3.7 Процедура отбора (Алгоритм 3.6, Шаг 6)

Шаг 1. Случайным образом выбрать 2 индекса $k_4, k_5 \in \{\overline{1, N_{POP}}\}, k_4 \neq k_5$.

Шаг 2. Если $f_{k_4} > f_{k_5}$, **тогда** $S_{k_4} \leftarrow S_c,\ f_{k_4} \leftarrow F(S_c)$, **иначе** $S_{k_5} \leftarrow S_c, f_{k_5} \leftarrow F(S_c)$.

ГА с жадной эвристикой для задач p-медиан и k-средних можно описать следующим образом.

Алгоритм 3.8. ГА с жадной эвристикой для задач p-медиан и k-средних (модификации GA-FULL, GA-ONE и GA-MIX)

Дано: Численность популяции N_{POP}.

Шаг 1. Установить $N_{iter} \leftarrow 0$. Выбрать набор начальных решений $\{S_1, \dots, S_{N_{POP}}\}$, где $|S_i| = k$. Улучшить каждое начальное решение алгоритмом k-средних и сохранить соответствующие полученные значения целевой функции (1), как переменные $f_k \leftarrow F(S_k), k = \overline{1, N_{POP}}$. В данной работе начальное значение популяции N_{POP}=5.

Цикл

Шаг 2. Если условие остановки выполнено, **тогда** СТОП. Вернуть решение $S_i, i \in \{\overline{1, N_{POP}}\}$ с минимальным значением f_i, **иначе** установить численность популяции следующим образом: $N_{iter} \leftarrow N_{iter} + 1;\ N_{POP} \leftarrow max\ \{N_{POP}, \lceil\sqrt{1 + N_{iter}}\rceil\}$; **если** N_{POP} изменился, **тогда** сгенерировать новый $S_{N_{POP}}$, как описано в Шаге 1.

Шаг 3. Случайным образом выбрать 2 индекса $k_1, k_2 \in \{\overline{1, N_{POP}}\}, k_1 \neq k_2$.

Шаг 4. Запустить Алгоритм 3.4 (для GA-ONE*) или Алгоритм 3.5 (для GA-FULL*) с решениями S_{k_1} и S_{k_2}. Для GA-MIX* Алгоритм 3.4 или Алгоритм 3.5 выбираются случайным образом с равной вероятностью. Получить новое решение S_c.

Шаг 5. $S_C \leftarrow Mutation(S_C)$. По умолчанию процедура мутации не используется.

Шаг 6. Запустить Алгоритм 3.7.

Конец Цикла

*GA-ONE – генетический алгоритм с жадной эвристикой с частичным объединением, GA-FULL – генетический алгоритм с жадной эвристикой с полным объединением, GA-MIX – случайный выбор алгоритмов 3.4 или 3.5

Этот алгоритм использует динамически растущую популяцию. В нашей новой версии шага 5 оператор перекрестной мутации выглядит следующим образом.

Algorithm 3.9 Оператор перекрестной мутации для Шага 5 Алгоритма 3.8 (модификации GA-FULL-MUT, GA-ONE-MUT и GA-MIX-MUT)

Шаг 1. Запустить алгоритм k-средних для случайно выбранного начального решения, чтобы получить решение S'.

Шаг 2.Запустить Алгоритм 3.4 (для GA-ONE) или Алгоритм 3.5 (для GA-FULL) с решениями S_c и S''. Получить новое решение S'_c.

Шаг 3. Если $F(S'_c) < F(S_c)$, **тогда** $S_c \leftarrow S'_c$.

Проведены вычислительные эксперименты с наборами данных из репозитариев Machine Learning Repository, Basic Benchmark repositories, а также с данными образцов промышленной продукции (Таблица 1). Новые модификации трех ГА (GA-FULL-MUT, GA-ONE-MUT и GA-MIX-MUT) сравнивались с известными алгоритмами j-means и k-средних (в режиме мультистарта), ГА без мутации (GA-FULL, GA-ONE и GA-MIX), алгоритмы АГ для задачи k-средних с комбинированным применением алгоритмов поиска с чередующимися рандомизированными окрестностями, образованными применением жадных агломеративных эвристических процедур (k-GH-VNS1, k-GH-VNS2, k-GH-VNS3), а также для задачи j-means (j-means GH-VNS1, j-means GH-VNS2). Для всех наборов данных выполнено 30 попыток запуска каждого алгоритма.

Лучшие значения новых алгоритмов (*) выделены жирным шрифтом, лучшие значения известных алгоритмов указаны курсивом, подчеркнуты наиболее достигнутые значения целевой функции (таблица 2). Для подтверждения статистической значимости преимущества (↑⇑) новых алгоритмов над известными алгоритмами использованы критерий Манна-Уитни (↑) и критерий Стьюдента (⇑).

Таблица 2 – Результаты вычислительных экспериментов для набора данных Europe (169309 векторов данных размерностью 2), 30 кластеров, 4 часа

Алгоритм	Значение целевой функции			
	минимум	максимум	среднее	σ
j-means	7,51477E+12	7,60536E+12	7,56092E+12	29,764E+9
k-means	7,54811E+12	7,57894E+12	7,56331E+12	13,560E+9
k-GH-VNS1	*7,49180E+12*	7,49201E+12	*7,49185E+12*	*0,073E+9*
k-GH-VNS2	7,49488E+12	7,52282E+12	7,50082E+12	9,989E+9
k-GH-VNS3	7,49180E+12	7,51326E+12	7,49976E+12	9,459E+9
j-means-GH-VNS1	7,49180E+12	7,49211E+12	7,49185E+12	0,112E+9
j-means-GH-VNS2	7,49187E+12	7,51455E+12	7,4962E+12	8,213E+9
GA-FULL-MUT*	7,49293E+12	7,49528E+12	7,49417E+12	0,934E+9
GA-MIX-MUT*	7,49177E+12	7,49211E+12	7,49186E+12	0,117E+9
GA-ONE-MUT*↑⇑	**7,49177E+12**	7,49188E+12	**7,49182E+12**	**0,042E+9**

Проведенные вычислительные эксперименты показывают, что ГА с жадным агломеративным оператором скрещивания с новой идеей процедуры мутации превосходят ГА без мутации по полученному значению целевой функции.

В четвертой главе предлагается следующая идея организации работы тестового центра: при поступлении в тестовый центр первых закупленных партий изделий путем проведения кластеризации выделяются однородные партии, каждая из которых сертифицируется. Обучающее множество (размеченная выборка) формируется из сертифицированных однородных партий. Затем новые партии изделий выделяются с помощью обученного классификатора. На ряду с моделью k-средних сигмоидальные искусственные нейронные сети (ИНС, ANN) представляют собой простые и легко интерпретируемые модели, при этом обеспечивающие высокое качество решения задач классификации (по критерию AUC ROC-кривой, площадь под кривой ошибок).

Глава посвящена разработке алгоритма обучения двухслойной сигмоидальной нейронной сети с регуляризацией, а также новому подходу к нахождению начального приближения ИНС и сохранению его положительных свойств при обучении с избыточностью описания в виде чрезмерно большого числа нейронов и одновременной недостаточной аппроксимацией данных в определенных частях области данных.

Предварительное обучение производится посредством минимизации по W квадратичной ошибки и негладкого сглаживающего функционала:

$$W = \arg\min_W E(\alpha, W, D), \tag{6}$$

$$E(\alpha, W, D) = \sum_{x,y\epsilon D}\left(y - f(x, W)\right)^2 + \alpha L(W),$$

где $D = \left\{\left(x^i, y_i\right) | x^i \in R^p, y_i \in R^1, i = 1, \dots, N\right\}$ – данные наблюдения, α – параметры регуляризации, *f(x,W)* – аппроксимирующая функция, $x \in R^p$ – вектор данных, $W \in R^n$ – вектор настраиваемых параметров, p и n – их размерности. $L(W)$ – регуляризатор. Регуляризатор $L(W)$ подавляет компоненты вектора W.

Нами использованы следующие виды регуляризации:

1. Квадратичная регуляризация А.Н. Тихонова $L2(W) = \sum_{i=1}^{k} w_{ij}^2$; (7)

2. Смешанная регуляризация $L\gamma(W, a_1, a_2) = a_1 \times L\gamma 2(W) + a_2 \times L2(W)$, (8)

где $L\gamma$2(W) – негладкая регуляризация:

$$L\gamma 2(W) = \sum_{i=1}^{k}\left(\left|w_{ij}\right| + \varepsilon\right)^{\gamma} (\varepsilon = 10^{-6}, \gamma = 0.7);$$

3. Модульная линейная регуляризация (Лассо Тибширани)

$$L1(W) = \sum_{i=1}^{k}\left|w_{ij}\right|. \tag{9}$$

Эти же виды регуляризации представлены в логистической регрессии.

Обучили двухслойную сигмоидальную нейронную сеть следующего вида:

$$f(x, W) = w_{i0}^{(2)} + \sum_{i=1}^{m} w_{ij}^{(2)} \varphi(s_i), \tag{10}$$

где $\varphi(s_i)$ – функция активации нейрона вида

$$\varphi(s_i) = 1/\left(1 + exp(-s_i)\right), \tag{11}$$

$$s_i = w_{i0}^{(1)} + \sum_{j=1}^{p} x_j w_{ij}^{(1)}, i = 1, \dots, m. \tag{12}$$

где x_j – компоненты вектора $x \in R^p$, W=(($w_{ij}^{(2)}$, i=0,…,m), ($w_{ij}^{(1)}$, j=0,…,p, i=1,…,m)) – набор неизвестных параметров, которые оцениваются оцениваемые с помощью (6), m – число нейронов.

Мы предлагаем составной алгоритм обучения ИНС. Алгоритм IA предназначен для выбора начального приближения параметров ИНС и основан на первоначальной оценке параметров нейрона, связанных с выбранными центрами, обеспечивая равномерное покрытие области данных рабочими областями нейронов. Регуляризация, даже с чрезмерным количеством параметров по сравнению с объемом данных, позволяет получить приемлемое решение.

Алгоритм 4.1. Алгоритм нахождения начального приближения ИНС (Алгоритм IA)

Требуется: исходные данные $D = \{(x^i, y_i) | x^i \in R^p, y_i \in R^1, i = 1, \dots, N\}$, начальное количество нейронов m (m<N).

Шаг 1. Разделить набор исходных данных D на обучающее (DO) и тестовое множество (DT). Принять D = DO.

Шаг 2. Выбрать регуляризатор (7), (8), (9) и установить его начальные параметры.

Шаг 3. На наборе данных D установить начальные центры рабочих областей нейронов c_i, $i = 1, \dots, m$. Центры рабочих областей c_i определяются одним из алгоритмов кластеризации. Это гарантирует, что нейроны будут находиться в областях с высокой плотностью данных.

Шаг 4. Исключаем свободный член нейрона $w_{ij}^{(0)}$ из (12) и используем выражение:

$$s_i = \sum_{j=1}^{p} (x_j - c_i) w_{ij}^{(1)}, i = 1, \dots, m. \quad (13)$$

Шаг 5. Определить начальные параметры $w_{ij}^{(1)}$ для (13) и $w_i^{(2)}$ для (10) с помощью датчика равномерно распределенных случайных чисел для каждого нейрона.

Шаг 6. Решить задачу (6) для (10), (11), (13) с фиксированными центрами, используя регуляризатор (7), (8), (9). Это позволяет на первом этапе покрыть всю область данных рабочими областями.

Шаг 7. Вернуться к исходному описанию ИНС (10), (11), (12) для определения свободного члена $w_{ij}^{(0)}$, оставив параметры без изменений: $w_{i0}^{(1)} = -\sum_{j=1}^{p} c_i w_{ij}^{(1)}$.

Шаг 8. Окончательный набор параметров обозначаем W^0.

В основном алгоритме поочередно выполняется удаление малозначимых нейронов, затем выполняется обучение ИНС:

Алгоритм 4.2. Алгоритм обучения ИНС

Дано: исходные данные $D = \{(x^i, y_i) | x^i \in R^p, y_i \in R^1, i = 1, \dots, N\}$, начальное количество нейронов m (m<N).

Шаг 1. Выбрать начальное приближение параметров сети W^0, применив Алгоритм 4.1 (IA).

Шаг 2. Выбрать модель окончательного приближения ИНС. Установить h=m.

Цикл для k=0,1,…,m-1

Шаг 3. Решить задачу (6) для (10), (11), (12)

Шаг 4. Установить $S_k = S(W^k, D)$.

$S(W^k, D)$ – величина среднеквадратичной погрешности определяется:

$$S(W^k, D) = \sqrt{\sum_{x,y \in D} \left(y - f(x, W^k)\right)^2 / N}.$$

Шаг 5. Если $S_k < (1 + \varepsilon ps)S_0$, где $\varepsilon ps = 0{,}1$, **тогда** удалить нейрон для которого выполняется $S_k < (1 + \varepsilon ps)S_0$, h=h-1

Шаг 6. Если не произошло удаление нейронов**, тогда** удалить нейрон, который приводит к наименьшему росту S, h=h-1.

Шаг 7. Если $h \leq q$**, тогда** выход из цикла.

Конец Цикла

Шаг 8. Выбираем $f(x, W^k)$ с числом параметров, не превосходящих N, и имеющих наименьшее значение показателя S_k, в качестве модели окончательного приближения и решить задачу (6) для (10), (11), (12).

Проведена серия экспериментов с наборами данных из репозиториев Machine Learning Repository и с результатами тестирования образцов промышленной продукции. Рассмотрены различные виды моделей классификации, в том числе с применением различных типов регуляризации. Для набора данных об образцах промышленной продукции с помощью обученного классификатора идентифицируем новые партии изделий, поступившие в тестовый центр. В данном случае в каждой смешанной партии изделий сформировано обучающее и контрольное множество (выборка) в соотношении 2/3 и 1/3 в порядке поступления в тестовый центр (Серия I) и случайным образом (Серия II). В таблице 3 приведены доли верно классифицированных объектов.

Таблица 3 – Результаты вычислительных экспериментов для Серии I (микросхема 1526IE10_002, Полная выборка (3987 объектов размерностью 68), обучающее множество 2567, тестовое множество 1330 объектов)

Партия / Модель	1 n=24	2 n=39	3 n=623	4 n=417	5 n=48	6 n=37	7 n=142	Среднее
LDA	1,00	0,95	0,74	0,29	1,00	1,00	0,64	0,57
NBA	0,98	0,95	0,70	0,00	0,92	1,00	0,88	0,53
SVM	0,98	0,97	0,85	0,73	0,95	0,97	0,89	0,83
ANN_Lγ	0,98	0,67	0,89	0,72	0,94	0,95	0,90	0,83
ANN_L2	0,94	0,77	0,86	0,67	0,90	0,97	0,89	0,80
ANN_NL	0,66	0,68	0,82	0,72	0,94	0,82	0,89	0,79
ANN_L1	0,98	0,73	0,87	0,61	0,96	0,97	0,89	0,79
LR_Lγ	0,98	0,97	0,85	0,82	0,95	0,97	0,89	0,87
LR_L2	0,98	0,96	0,85	0,72	0,95	0,98	0,90	0,82
LR_NL	0,80	0,86	0,85	0,78	0,85	0,55	0,90	0,82
LR_L1	0,98	0,97	0,85	0,71	0,95	0,97	0,89	0,82

LDA – линейный дискриминантный анализ, NBA – наивный байесовский классификатор, SVM – метод опорных векторов, LR – логистическая регрессия; ANN – искусственная нейронная сеть; Lγ – смешанная регуляризация; L2 – квадратичная регуляризация; NL – без регуляризации; L1 – модульно линейная регуляризация

Результаты вычислительных экспериментов над данными промышленной продукции свидетельствуют об эффективности предложенного алгоритма. В случае с небольшим количеством однородных партий в выборке (2 – 4) практически все рассмотренные методы классификации позволяют решать задачу классификации с высокой точностью (0,95 – 1,00). При использовании

нейросетевых моделей классификации и моделей логистической регрессии наилучшие результаты достигаются с применением новых алгоритмов.

ЗАКЛЮЧЕНИЕ

В диссертационной работе предложены модели и алгоритмы, которые позволяют повысить точность (по индексу Рэнда) автоматической группировки промышленной продукции и получить более стабильные (по коэффициенту вариации) значения целевой функции в сравнении с известными моделями и алгоритмами автоматической группировки.

Цель диссертации достигнута путем решения поставленных задач, а именно:

1. Сравнительный анализ известных алгоритмов решения задач автоматической группировки объектов с повышенными требованиями к точности разделения, показал, что существующие решения не позволяют одновременно получать высокую точность разделения объектов, получать стабильные результаты при многократных запусках алгоритмов, а также иметь высокую скорость работы алгоритмов. В области классификации существует проблема, связанная с точностью классификации в условиях малого количества данных и большого количества признаков.

2. Построена новая модель автоматической группировки объектов на основе модели k-средних с расстояниями Махаланобиса и обучаемой ковариационной матрицей. Предложенный новый алгоритм, основанный на оптимизационной модели k-средних с расстояниями Махаланобиса и обучаемой ковариационной матрицей, позволяет снизить долю ошибок (повысить индекс Рэнда) при выявлении однородных производственных партий продукции. А также продемонстрировано, что модель с применением методов факторного анализа позволяет уменьшить размерность исходных данных и уменьшить долю ошибок кластеризации до нуля для задачи разбиения на две однородные партии.

3. Разработан новый генетический алгоритм для задачи k-средних с оригинальной идеей использования одной процедуры в качестве оператора скрещивания и оператора мутации, который демонстрирует более точный и стабильный результат значения целевой функции за фиксированное время выполнения.

4. Разработан алгоритм обучения двухслойной сигмоидальной искусственной нейронной сети с регуляризацией. Предложен новый подход нахождения начального приближения искусственной нейронной сети и сохранение его положительных свойств при обучении с избыточностью описания в виде чрезмерно большого числа нейронов. Полученные результаты демонстрируют высокую точность классификации с применением предложенных в работе алгоритмов регуляризации при наличии малоинформативных признаков.

Таким образом, в диссертации решена задача повышения точности работ систем классификации промышленной продукции за счет применения усовершенственных математических моделей и алгоритмов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Шкаберина, Г.Ш. Факторный анализ с использованием матрицы Спирмена в задаче автоматической группировки электрорадиоизделий по производственным партиям / Г.Ш. Шкаберина, В.И. Орлов, Е.М. Товбис, Л.А. Казаковцев // Системы управления и информационные технологии. – 2019. – № 1(75). – С. 91 – 96.

2. Шкаберина, Г.Ш. Применение алгоритмов кластеризации с особыми мерами расстояния для задачи автоматической группировки электрорадиоизделий / В.И. Орлов В.И, Г.Ш. Шкаберина, И.П. Рожнов, А.А. Ступина, Л.А. Казаковцев // Системы управления и информационные технологии. – 2019.– № 3(77). – С. 42 – 46.

3. Шкаберина, Г.Ш. Оптимизационная модель для автоматической группировки многомерных данных / Г.Ш. Шкаберина // Системы управления и информационные технологии. – 2020. – № 3(81). – С. 94–99.

4. Шкаберина, Г.Ш. Модели и алгоритмы автоматической группировки объектов на основе модели k- средних / Г.Ш. Шкаберина, Л.А. Казаковцев, Ж. Ли // Сибирский журнал науки и технологий. – 2020. – Т. 21, № 3. – С. 347–354.

Публикации в изданиях, входящих в Web of Science и Scopus:

5. Shkaberina, G. Automatic classification model and algorithms based on the minimum sum-of-squared errors model / G. Shkaberina, L. Kazakovtsev // Industrial Engineering & Management Systems. – 2020. – Vol. 19, № 4. – P. 730–737.

6. Shkaberina, G. Sh. New Method of Training Two-layer Sigmoid Neural Networks Using Regularization / V. N. Krutikov, L. A. Kazakovtsev, G. Sh. Shkaberina, V. L. Kazakovtsev // IOP Conference Series: Materials Science and Engineering 2019. – 2019. – Vol. 537, Issue 4. – Article ID 042055. – 6 P. DOI: 10.1088/1757-899X/537/4/042055.

7. Shkaberina, G. S. New Methods of Training Two-Layer Sigmoidal Neural Networks with Regularization / V. N. Krutikov, G. S. Shkaberina, M. N. Zhalnin, L. A. Kazakovtsev // 2019 International Conference on Information Technologies (InfoTech), St. St. Constantine and Elena resort (near the city of Varna), Bulgaria, 2019. – 2019. – P. 1–4. DOI: 10.1109/InfoTech.2019.8860890.

8. Shkaberina, G.S. Identification of the Optimal Set of Informative Features for the Problem of Separating of Mixed Production Batch of Semiconductor Devices for the Space Industry / G.S. Shkaberina, V.I. Orlov, E.M. Tovbis, L.A. Kazakovtsev // In: Bykadorov I., Strusevich V., Tchemisova T. (eds) Mathematical Optimization Theory and Operations Research. MOTOR 2019. Communications in Computer and Information Science. – 2019. – Vol 1090. – P. 408 – 421. DOI: 10.1007/978-3-030-33394-2_32.

9. Shkaberina, G. Algorithms for Reduction of Input Space Dimensionality in Regression-Based Classification Models / V. Krutikov, G. Shkaberina, L. Kazakovtsev // 2019 International Russian Automation Conference (RusAutoCon), Sochi, Russia. – 2019. – P. 1 – 5. DOI: 10.1109/RUSAUTOCON.2019.8867674.

10. Shkaberina, G. Sh. Efficiency of distance measures in the automatic grouping of electronic radio devices by k-means algorithm / G. Sh. Shkaberina, E. M. Tovbis, L. A. Kazakovtsev, A. O. Shemyakov, A. P. Romanov, N. V. Lukonin // IOP

Conference Series: Materials Science and Engineering 2019. – 2020. – Vol. 734. – Article ID 012136. – 5 P. DOI: 10.1088/1757-899X/734/1/012136.

11. Shkaberina, G. Sh. Detection of homogeneous production batches of semiconductor devices by greedy heuristic clustering algorithms with special distance metrics / G. Sh. Shkaberina, I. P. Rozhnov, V. P. Popov, L.A. Kazakovtsev, E. V. Lapunova // IOP Conference Series: Materials Science and Engineering 2019. – 2020. – . Vol. 734. –Article ID 012104. – 5 P. DOI: 10.1088/1757-899X/734/1/012104.

12. Shkaberina, G. Sh. On the optimization models for automatic grouping of industrial products by homogeneous production batches / G. Sh. Shkaberina, V. I. Orlov, E. M. Tovbis, L.A. Kazakovtsev // In: Kochetov Y., Bykadorov I., Gruzdeva T. (eds) Mathematical Optimization Theory and Operations Research. MOTOR 2020. Communications in Computer and Information Science. – 2020. – Vol. 1275. – P. 421–436. DOI: 10.1007/978-3-030-58657-7_33.

13. Shkaberina, G. Genetic Algorithms with the Crossover-Like Mutation Operator for the k-Means Problem. / L. Kazakovtsev, G. Shkaberina, I. Rozhnov, R. Li, V. Kazakovtsev // In: Kochetov Y., Bykadorov I., Gruzdeva T. (eds) Mathematical Optimization Theory and Operations Research. MOTOR 2020. Communications in Computer and Information Science. – 2020. – Vol. 1275. – P. 350–362. DOI: 10.1007/978-3-030-58657-7_28.

14. Shkaberina, G. Sh. Regularization methods for neural network models and logistic regression models in the problem of classifying industrial products into homogeneous batches / V.N. Krutikov, G. Sh. Shkaberina, E.M. Tovbis, L.A. Kazakovtsev // 2020 International Conference on Information Technologies (InfoTech), Varna, Bulgaria 2020. – 2020. – P. 1–4. DOI: 10.1109/InfoTech49733.2020.9211013.

15. Shkaberina, G. K-means genetic algorithms with greedy genetic operators / L. Kazakovtsev, I. Rozhnov, G. Shkaberina, V. Orlov // Mathematical Problems in Engineering. – 2020. – Article ID 8839763. DOI: 10.1155/2020/8839763

В других изданиях:

16. Shkaberina, G. Sh. Correlation between the time taken to master the competency and the rank of competence evaluated on the basic educational programme G. Shkaberina, L. Baranovskaya // Journal of Economics and Social Sciences. – 2016. – № 8. – P. 23–26.

Свидетельство о государственной регистрации программы для ЭВМ:

17. Шкаберина, Г.Ш. Программный комплекс решения задач автоматической группировки объектов с применением массивно-параллельных систем / В.И. Орлов, Л.А. Казаковцев, И.П. Рожнов, Г.Ш. Шкаберина, И.С. Масич // М.: РОСПАТЕНТ. – 2020. Свидетельство № 2020663109 от 22.10.2020.

Шкаберина Гузель Шарипжановна
Модели и алгоритмы автоматической классификации продукции
Автореферат
Подписано к печати 04.12.2020. Формат 60x84/16
Уч. изд. л. 1 Тираж 100 экз. Заказ № _____
Отпечатано в отделе копировальной и множительной техники
СибГУ им. М.Ф. Решетнева.
660037, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31