Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Базовая кафедра вычислительных и информационных технологий

УТВЕРЖДАЮ
Заведующий кафедрой
_____ / В.В. Шайдуров
«16» июня 2016 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 02.03.01 Математика и компьютерные науки

РАЗРАБОТКА ГЕНЕТИЧЕСКОГО АЛГОРИТМА ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ ДАННЫХ

Научный руководитель
кандидат физико-математических наук,
доцент _____ / И.В. Баранова
16.06.2016

Выпускник _____ / М.Г. Семенов
16.06.2016

Красноярск 2016

РЕФЕРАТ

Бакалаврская работа по теме «Разработка генетического алгоритма для решения задачи кластеризации данных» содержит 55 страниц текста, 1 приложение, 25 использованных источников.

КЛАСТЕР, КЛАСТЕРИЗАЦИЯ, ГЕНЕТИЧЕСКИЙ АЛГОРИТМ, МНОГОМЕРНЫЕ ДАННЫЕ, АЛГОРИТМ К-СРЕДНИХ, FOREL.

Цель работы – разработка генетического алгоритма, предназначенного для решения задачи кластеризации многомерных данных.

В результате исследования были изучены основные алгоритмы кластеризации многомерных данных. Реализованы методы кластеризации k-средних и FOREL. Разработан генетический алгоритм для решения задачи кластеризации многомерных данных с заданным количеством кластеров и алгоритм для решения задачи кластеризации с заданным размером кластеров. Создано программное приложение, реализующее работу предложенных алгоритмов кластеризации. Проведено сравнение изученных и предложенных методов по их вычислительной сложности и результатам работы. Решена практическая задача кластеризации 50 российских банков по показателям их деятельности. Проведено сравнение результатов, полученных в результате работы каждого метода.

СОДЕРЖАНИЕ

Введение……..….....3
1 Задача кластеризации данных…….……....5
1.1 Понятие кластеризации……..…….5
1.2 Постановка задачи кластеризации……..….....6
1.3 Метрики для задания кластеров…….……..9
1.4 Критерии кластеризации…….…10
1.5 Типы кластерных структур……..……...…13
1.6 Методы кластеризации…….……..15
2 Классические методы решения задачи кластеризации данных……20
2.1 Метод k-средних……20
2.2 Решение задачи кластеризации методом FOREL…….....23
3 Генетические алгоритмы……..……26
3.1 Основные понятия генетических алгоритмов…….……27
3.2 Классический генетический алгоритм…….……...28
3.3 Генетический алгоритм для решения задачи кластеризации…….35
4 Сравнение предложенных алгоритмов со стандартными методами кластеризации ……42
5 Решение практической задачи кластеризации…….....45
5.1 Описание статистики……..46
5.2 Решение задачи кластеризации методом k-средних и генетическим алгоритмом кластеризации с детерминированным числом кластеров……46
5.3 Решение задачи кластеризации методами FOREL и генетическим алгоритмом кластеризации с заданным радиусом кластеров……...49
5.4 Сравнение полученных результатов…….....51
5.5 Визуализация результатов……..52
Заключение ……..…….55
Список использованных источников…….……..56
Приложение А……...59

ВВЕДЕНИЕ

Целью бакалаврской работы является разработка генетического алгоритма, решающего задачу кластеризации многомерных статистических данных.

Задача кластеризации данных [14] (также называемая таксономией, автоматической классификацией или группировкой объектов) является одной из наиболее важных и сложных задач анализа данных.

Кластерный анализ [15] представляет собой раздел статистического анализа данных, объединяющий методы разбиения (группировки) множества наблюдаемых объектов на сравнительно однородные группы, называемые кластерами. Однородность кластеров означает, что объекты, отнесенные к одному кластеру, должны быть близки относительно выбранной метрики. Объекты из разных кластеров должны существенно отличаться. Кластерный анализ является востребованной и успешно развивающейся дисциплиной современной теоретической информатики. Его методы имеют широкий спектр применений практически во всех областях человеческой деятельности, связанных с изучением объектов и процессов: медицине, биологии, химии, маркетингу, психологии, социологии, менеджменту, филологии, археологии и другим.

В первой части работы приводятся основные понятия и постановка задачи кластеризации данных. Перечисляются виды наиболее часто выделяемых кластерных структур и критерии качества кластеризации. Рассматриваются основные методы кластеризации данных. Особое внимание в работе уделяется двум наиболее популярным методам кластеризации данных: методу k-средних и FOREL [6]. Дается подробное описание данных алгоритмов, и указываются особенности их работы.

Во второй части работы излагаются принципы работы генетических алгоритмов [7], являющихся эвристическими алгоритмами поиска, широко используемыми для решения задач оптимизации и моделирования путём

случайного подбора и вариации параметров с использованием механизмов, аналогичных естественному отбору в природе. В работе рассматриваются назначение и виды основных операторов генетических алгоритмов, в том числе операторы скрещивания, мутации и селекции. Приводится схема этапов работы генетического алгоритма в общем виде.

В работе предлагаются два генетических алгоритма, позволяющих решать задачу кластеризации данных. Первый алгоритм решает задачу кластеризации данных с заранее заданным числом кластеров. Второй алгоритм разбивает объекты на кластеры заданного радиуса. Приводится подробное описание видов операторов инициализации, скрещивания, мутации и селекции и схема работы для обоих предложенных алгоритмов кластеризации данных.

Было разработано программное приложение, реализующее работу предложенных алгоритмов кластеризации, а также классических алгоритмов k-средних и FOREL.

В работе проводится сравнение предложенных генетических алгоритмов и классических алгоритмов (k-средних и FOREL) по их вычислительной сложности. Выполняется серия численных экспериментов, позволяющих оценить работу разработанных алгоритмов на ряде тестовых примеров. Полученные результаты кластеризации сравниваются с результатами вышеперечисленных классических алгоритмов кластеризации для тех же тестовых примеров.

В работе решается практический пример кластеризации данных – кластеризация 50 ведущих российских банков по основным показателям их финансовой деятельности. Данная задача решается с помощью разработанных генетических алгоритмов и классических алгоритмов (k-средних и FOREL). Исследование основывается на реальных статистических данных. Проводится сравнение результатов работы методов и анализ полученных результатов.

1 Задача кластеризации данных

1.1 Понятие кластеризации данных

В настоящее время практически во всех областях человеческой деятельности существует настоятельная потребность в изучении статистических данных, описывающих поведение наблюдаемых объектов, событий, процессов или явлений. Одной из наиболее актуальных и практически востребованных задач анализа данных является задача разбиения объектов на сравнительно однородные группы (подмножества), называемые кластерами.

Определение 1.1 *Кластер* — группа однородных элементов, характеризуемых общим свойством.

Однородность кластеров означает, что объекты, отнесенные к одному кластеру, должны быть схожи (близки) относительно выбранной метрики. Объекты из разных кластеров должны существенно отличаться. Данная задача называется задачей кластеризации данных. Также ее принято называть таксономией, автоматической классификацией, группировкой объектов или задачей обучения без учителя.

Кластерный анализ представляет собой раздел статистического анализа данных, объединяющий методы разбиения (группировки) множества объектов на группы, называемые кластерами. Термин «кластерный анализ» был впервые введен Робертом Трайоном в 1939 году [14].

К числу основных задач, выполняемых кластерным анализом, относятся:

- разработка типологии или классификации;
- создание полезных концептуальных схем группирования объектов;
- порождение гипотез на основе исследования данных;
- проверка гипотез или проведение исследования для определения, действительно ли выделенные группы присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

1. отбор выборки для кластеризации;

2. определение множества переменных, по которым будут оцениваться объекты в выборке;

3. вычисление значений той или иной меры сходства между объектами;

4. применение метода кластерного анализа для создания групп сходных объектов;

5. проверка достоверности результатов кластеризации.

Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Несомненным достоинством кластерного анализа является то, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных произвольной природы.

Приведем формальную постановку задачи кластерного анализа в общем виде, а также необходимые определения из кластерного анализа.

1.2 Постановка задачи кластеризации

Пусть $X = \{x_1, x_2, \ldots, x_m\}$ – множество объектов, заданных значениями в пространстве признаков $P = \{P_1, P_2, \ldots, P_n\}$ (т.е. каждый объект $x_i = (x_i^1, x_i^2, \ldots, x_i^m)$) и задана функция расстояния (метрика) между объектами $\rho(x_i, x_j)$, $x_i, x_j \in X$.

Определение 1.2. Функцией кластеризации называется функция $f : X \to Y$, которая любому объекту $x \in X$ ставит в однозначное соответствие номер $y \in Y = \{1, \ldots, k\}, k \le m.$

Определение 1.3. Множество кластеров $C=\{C_1,C_2,\ldots,C_k\}$, $k \le m$, представляет собой разбиение множества объектов X такое, что кластер $C_i=\{x \in X,\ f(x)=i\}$, $C_i \cap C_j = \varnothing$. Причем для C справедливо следующее: если $x_i, x_j \in C_i$, то $\rho(x_i,x_j)$ минимально. Если $x_i \in C_i, x_j \in C_j$, то $\rho(x_i,x_j)$ максимально.

Тогда постановку задачи кластеризации данных можно сформулировать следующим образом:

Требуется найти такую функцию кластеризации f^*, чтобы

$$Q(f^*,C,\rho)=\min_f Q(f,C,\rho), \tag{1.1}$$

где $Q(f,C,\rho)$ – выбранный критерий качества кластеризации.

Как уже было сказано выше, каждый объект описывается набором своих характеристик, называемых *признаками*. Признаки могут быть следующих типов:

- *бинарный* признак: $P_i=\{0,1\}$,
- *номинальный (качественный)* признак: P_i — конечное множество;
- *порядковый* признак: P_i — конечное упорядоченное множество;
- *количественный* признак: $P_i=\Re$ — множество действительных чисел.

Самой распространенной ситуацией является кластеризация объектов, у которых все признаки являются количественными, т.е. когда $P=\{P_1,P_2,\ldots,P_n\}=\Re^n$ (каждый объект $x_i=(x_i^1,x_i^2,\ldots,x_i^n)\in\Re^n$). В работе мы будем рассматривать именно такую ситуацию.

Иногда при формулировке задачи кластеризации кроме множеств *X* и *P* могут быть заданы дополнительные априорные данные о характеристиках

множества кластеров K. Таким образом, исходя из состава входных данных, можно выделить четыре основных типа задачи кластеризации:

1. Задано необходимое количество кластеров k;
2. Заданы ограничения на число объектов для всех кластеров $C_i \in C$;
3. Заданы ограничения на пространственные характеристики кластеров $C_i \in C$;
4. Нет информации о количестве и характеристиках кластеров $C_i \in C$.

Наиболее простой задачей из всех перечисленных является задача первого типа, которую можно назвать ***задачей кластеризации с заданным числом кластеров***. Это связано с тем, что в этих задачах уже практически задан критерий качества. Достаточно выбрать меру, в соответствии с которой будет вычисляться расстояние между объектами, и начать объединять наиболее близкие из них.

Очень похожими на задачи первого типа являются задачи, в которых заранее неизвестно количество классов, однако заданы ограничения на число объектов в кластере. Похожи и методы решения этих задач, за исключением критерия, который используется в этом случае. На первом шаге также следует выбрать меру расстояния между объектами. Далее объединяются наиболее близкие объекты. Если число объектов в каком-либо кластере достигает заданной величины, другие объекты, которые можно было бы отнести к этому таксону, образуют новый кластер.

Наиболее распространенным и наиболее сложным является последний тип задач кластеризации – задачи, в которых известны только значения признаков объектов выборки и нет никаких заданных требований к результатам решения. Этот тип задач сегодня можно решить только путем выдвижения некоторых эвристических гипотез, касающихся законов распределения объектов выборки. Данная работа посвящена решению задач кластеризации первого и второго типа.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что при классификации обязательно задается обучающая выборка объектов $\tilde{X} \subset X$, для которых известно, к каким классам они относятся.

1.3 Метрики для задания кластеров

Как было сказано выше, для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками [15] или функциями расстояний:

1) Наиболее популярной является *евклидова метрика*. Евклидова метрика между точками x и y это длина отрезка $\overline{xy}$. В декартовых координатах, если $x = \{x_1, x_2, ..., x_n\}$ и $y = \{y_1, y_2, ..., y_n\}$ две точки в евклидовом пространстве, длина отрезка $\overline{xy}$ равна:

$$p(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n} (x_p - y_p)^2}. \quad (1.2)$$

2) Для придания большего значения более отдаленным друг от друга объектам, можно использовать *квадрат евклидова расстояния.* Это расстояние вычисляется следующим образом:

$$p(x, y) = \sum_{p=1}^{n} (x_p - y_p)^2. \quad (1.3)$$

3) *Расстояние городских кварталов (манхэттенское расстояние)*. Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к. они не возводятся в квадрат). Формула для расчета манхэттенского расстояния:

$$p(x, y) = \sum_{p=1}^{n} | x_p - y_p |. \quad (1.4)$$

4) *Расстояние Чебышева*. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате. Расстояние Чебышева вычисляется по формуле:

$$p(x, y) = \max | x_p - y_p |. \quad (1.5)$$

5) *Степенное расстояние.* Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

$$p(x, y) = \sqrt[u]{\sum_{p=1}^{n} (x_p - y_p)^v}, \quad (1.6)$$

где u и v – параметры, определяемые пользователем. Параметр u ответственен за постепенное взвешивание разностей по отдельным координатам, параметр v ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – u и v равны двум, то это расстояние совпадает с расстоянием Евклида.

Выбор метрики зависит от конкретной задачи, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер.

1.4 Критерии качества кластеризации

Поскольку существует большое число различных алгоритмов, разбивающих один и тот же набор данных на разное множество кластеров, т.е. получающих разный набор $C = \{C_1, C_2, \ldots, C_k\}$, то возникает проблема

сравнения алгоритмов и качества получаемых ими решений. Как уже было сказано выше, для этого используются критерии качества кластеризации.

Оптимизационные критерии кластер-анализа могут быть разделены на три типа:

(а) эвристические; в таких критериях формализуется интуитивная идея, что объекты внутри кластеров должны быть близки друг к другу, а в разных кластерах – далеки друг от друга;

(б) аппроксимационные; такие критерии основаны на представлении искомой кластерной структуры математическими объектами того же типа, что и данные, обычно в виде матриц, так что в качестве критерия выступает степень близости между матрицей исходных данных и матрицей формируемой кластер-структуры.

(в) статистического оценивания; обычно это критерий максимального правдоподобия какой-либо статистической модели, такой, как смесь распределений.

В настоящее время основное значение имеют эвристические критерии, которые, по мере их использования в анализе данных, постоянно модифицируются и уточняются, в том числе на основе аппроксимационных или статистических соображений.

Для сравнения качества разбиения на классы [12] используется ряд функционалов качества. Наиболее распространенные:

Среднее внутрикластерное расстояние должно быть как можно меньше:

$$Q_0 = \frac{\sum_i \sum_{x,y \in C_i} p(x,y)}{k} \to \min. \tag{1.7}$$

Среднее межкластерное расстояние должно быть как можно больше:

$$Q_1 = \sum_{i<j} \sum_{x \in C_i, y \in Cj} p(x,y) \to \max. \tag{1.8}$$

Отношение пары функционалов: $Q_0 / Q_1 \to \min$.

Если алгоритм кластеризации вычисляет центры кластеров , $y \in Y$, то можно определить функционалы, вычислительно более эффективные.

Сумма средних внутрикластерных расстояний должна быть как можно меньше:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{y_i = y} p^2(x_i, \mu_y) \to \min, \tag{1.9}$$

где $K_y = \{x_j \in X^l \mid y_i = y\}$ – кластер с номером y, μ_y – центр масс кластера y.

В этой формуле можно было бы взять не квадраты расстояний, а сами расстояния. Однако, если р евклидова метрика, то внутренняя сумма в Φ_0 приобретает физический смысл момента инерции кластера K_y относительно его центра масс, если рассматривать кластер как материальное тело, состоящее из $|K_y|$ точек одинаковой массы.

Сумма межкластерных расстояний должна быть как можно больше:

$$\Phi_1 = \sum_{y \in Y} p^2(\mu_y, \mu) \to \max, \tag{1.10}$$

где μ – центр масс всей выборки.

Отношение пары функционалов: $\Phi_0 / \Phi_1 \to \min$.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

- не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих четко выраженного критерия, но осуществляющих достаточно разумную кластеризацию по построению.
- результаты кластеризации существенно зависят от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.

Определение 1.3. *Центр кластера* — это среднее геометрическое место точек в пространстве переменных.

Определение 1.4. *Радиус кластера* — максимальное расстояние точек от центра кластера. Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров.

Определение 1.5. *Спорный объект* — это объект, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

1.5 Типы кластерных структур

В процессе развития кластерного анализа было замечено, что методы кластеризации работают успешно с одними типами кластерных структур, и показывают плохие результаты с другими. Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов. В связи с этим, появился раздел кластерного анализа, в котором были систематизированы наиболее часто встречающиеся типы кластеров. В литературе [6] принято выделять следующие типы кластерных структур:

1) Множества центроидов;
2) Разбиения;
3) Разбиения с центроидами;
4) Отдельные кластеры;
5) Аддитивные кластеры;

Далее эти виды структур будут кратко охарактеризованы, прежде всего, с точки зрения оснований.

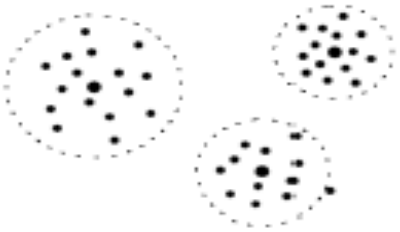

Рисунок 1.1 – Множество центроидов

Как только задано конечное множество центроидов в пространстве, каждая точка пространства приписывается одному из центроидов согласно так называемому принципу минимального расстояния – ближайшему в рассматриваемой метрике, обычно Евклидовой. При этом совокупность гиперплоскостей, разделяющих области притяжения каждого из центроидов (рис.1.1). Очевидно, эти области притяжения образуют разбиение пространства, определяемое данной системой центроидов.

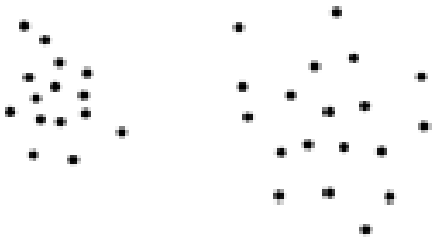

Рисунок 1.2 – Разбиения

Разбиение (рис.1.2) – совокупность непустых непересекающихся классов – одна из самых популярных кластерных структур, особенно часто применяемая при анализе данных о сходстве между объектами. Типичная проблема, возникающая при этом – интерпретация классов получаемого разбиения. Поэтому по возможности кластеры сопровождаются их «представителями» - объектами или усредненными характеристиками, представляющими основные тенденции кластера.

Рисунок 1.3 – Разбиения с перемычками

Разбиения с перемычками – отличие данного типа кластера от вышеописанного в том, что кластеры дополнительно могут соединяться перемычками. Пример данного типа можно увидеть на рис. 1.3.

Рисунок 1.4 – Отдельные кластеры

Отдельные кластеры (рис.1.4) - это структура, которая оправдывает себя в частных случаях, когда часть объектов «не ложится» в кластеры, будучи либо уникальными, включая выбросы и ошибки, либо частями бесформенной массы.

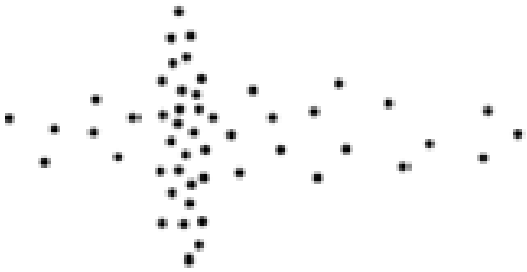

Рисунок 1.5 – Аддитивные кластеры

Аддитивные кластеры (рис.1.5) - это совокупность отдельных кластеров, обычно пересекающихся, в которой каждый кластер ассоциирован с положительной величиной – интенсивностью кластера. Предполагается, что сходства между любыми двумя объектами равно сумме интенсивностей тех кластеров, которым принадлежат оба объекта.

1.6 Методы кластеризации

Методы кластерного анализа [15] можно разделить на две группы: *иерархические и неиерархические.* Каждая из групп включает множество подходов и алгоритмов.

Большинство известных методов, направленных на решение задачи кластеризации, способны решать задачи первого, второго и третьего типа. К этим методам относятся:

- алгоритмы метода динамических сгущений, в которых вводится понятие центров кластеров, быстродействующие, разработанные для формирования первых поверхностных представлений о структуре данных в пространстве признаков [6];
- алгоритмы, основанные на теории нечетких множеств, которые допускают, что один объект может быть одновременно отнесен к нескольким классам с заданной количественной мерой принадлежности [9], [10];
- алгоритмы, использующие нейронные сети для разделения множества объектов на классы, такие, как нейронная сеть Кохонена или Хебба [7].

Иерархические методы

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие, или разделении больших кластеров на меньшие (рис. 1.6). Следовательно, эти методы можно разделить на две группы:

а) Иерархические агломеративные методы.

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров. В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

б) Иерархические дивизимные (делимые) методы.

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Иерархические методы кластеризации различаются правилами построения кластеров [19]. В качестве правил выступают критерии, которые используются при решении вопроса о «схожести» объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы). Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

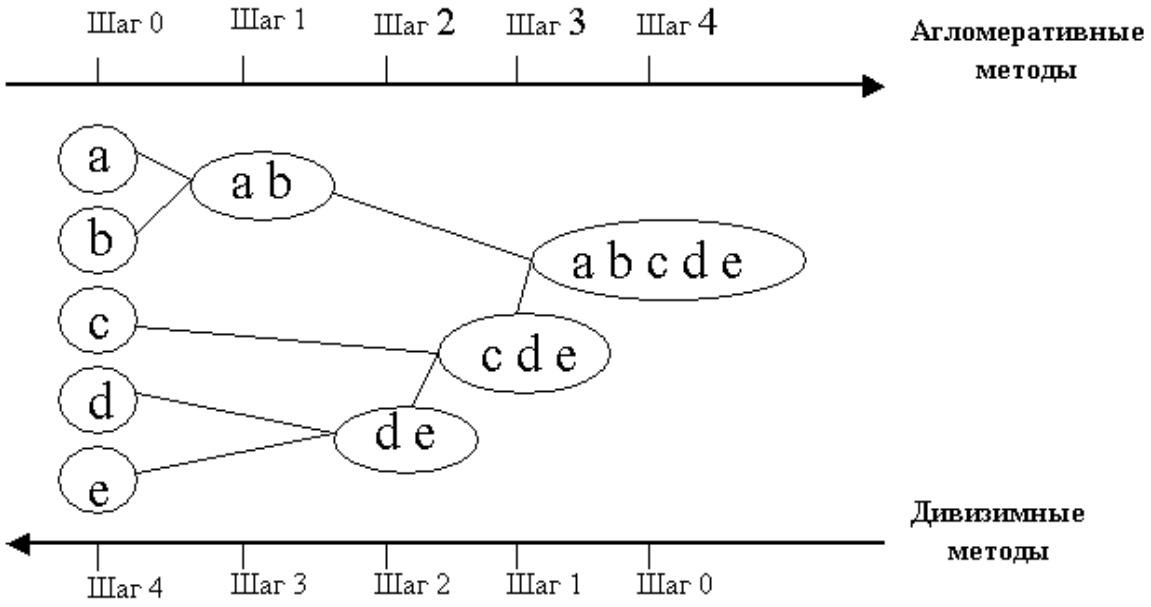

Рисунок 1.6 Принцип работы агломеративных и дивизимных методов

Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос — как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

- **Метод ближайшего соседа.** Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными «цепочками», «сцепленными вместе» только отдельными элементами, которые случайно оказались ближе остальных друг к другу [17].
- **Метод наиболее удаленных соседей.** Здесь расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. «наиболее удаленными соседями»). Метод хорошо использовать, когда объекты действительно происходят из различных «рощ». Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является «цепочечным», то этот метод не следует использовать.
- **Метод Варда.** В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и «стремится» создавать кластеры малого размера.

- **Метод невзвешенного попарного среднего.** В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты действительно происходят из различных «рощ», в случаях присутствия кластеров «цепочного» типа, при предположении неравных размеров кластеров.

Неиерархические методы

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют *неиерархические методы,* основанные на разделении, которые представляют собой *итеративные методы дробления исходной совокупности.* В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки [18].

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое «сгущение точек». Второй подход заключается в минимизации меры различия объектов.

- **Метод k-средних.** Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров. Главная идея — минимизация разницы между элементами кластера и максимизация расстояния между кластерами.

- **Метод k-медиан.** Модификация метода k-средних. В качестве центров кластеров выбираются медианы. Алгоритм менее чувствителен к

шумам и выбросам данных, чем алгоритм k-средних, поскольку медиана меньше подвержена влияниям выбросов. Также этот алгоритм применяется, когда нет возможности определить центроиду.

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации. Ценой, которую приходится платить за эти достоинства метода, является слово «априори». Аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации. Это особенно сложно начинающим специалистам.

Если нет предположений относительно числа кластеров, рекомендуют использовать иерархические алгоритмы. Однако если объем выборки не позволяет это сделать, возможный путь — проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты. За счет такого «варьирования» результатов достигается достаточно большая гибкость кластеризации.

2 Классические методы решения задачи кластеризации данных

2.1 Метод k-средних

В работе рассмотрен один из популярных алгоритмов кластеризации – алгоритм k-средних [12], относящийся к неиерархическому подходу. Также этот метод называют быстрым кластерным анализом. Данный алгоритм основан на минимизации функционала суммарной выборочной дисперсии разброса элементов относительно центров тяжести кластеров $Q = Q^{(3)}$. Этот алгоритм представляет собой итерационное нахождение центров тяжести

кластеров и разбиение обучающей выборки на кластеры до тех пор, пока функционал Q не перестанет меняться.

В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Число «k» в названии метода означает количество кластеров, на которое производится разбиение данных. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции. Слово «средние» в названии метода относится к центроидам кластеров.

Определение 2.1. *Центроид* — точка данных $\mu_j = (\mu_j^1, \mu_j^2, ..., \mu_j^n)$, представляющая собой центр масс точек кластера, т.е. покоординатное среднее точек из кластеров: $\mu_j = \sum_{x_j \in C} x_j^i$, $j = \overline{1,n}$, $i = \overline{1,k}$.

Описание алгоритма:

Пусть имеется множество точек данных $X = \{x_1, ..., x_m\}$, где $x_i = (x_i^1, x_i^2, ..., x_i^m) \in \Re^n$.

Задается количество кластеров k, и на первом шаге производится задание центроидов μ_j, $l = 1, ..., n$ – «центров масс» кластеров S_j, $j = 1, ..., k$.

Каждому кластеру соответствует один центр. Выбор начальных центроидов может осуществляться следующим образом:

- выбор k – наблюдений для максимизации начального расстояния;
- случайный выбор k – наблюдений;
- выбор первых k - наблюдений.

Кластеры $S_j = \{\varnothing\}$, $j = 1, ..., k$.

1. Производится распределение объектов по кластерам. Точка x_i, $i = 1, ..., n$ относится к ближайшему кластеру, т.е. $x_i \in S_j$, где

$$p(x_i, \mu_{j*}) = \min_{j=1,\dots k} p(x_i, \mu_j).$$

В качестве метрики используется одна из приведенных выше метрик, чаще всего евклидова.

В результате каждый объект назначен определенному кластеру.

2. Вычисляются новые центры кластеров μ_j, $j = 1,\dots,k$, как центры масс новых кластеров S_j, $j = 1,\dots,k$, полученных на предыдущем этапе.

3. Продолжать итерационный процесс вычисления центров и перераспределения объектов до тех пор, пока не выполнится одно из условий:

- кластерные центры μ_j стабилизировались (перестали изменяться);
- число итераций равно максимальному числу итераций (ограничение на число итераций).

Алгоритм k-средних минимизирует функционал суммарной выборочной дисперсии Φ_0 и сходится за конечное число шагов.

В работе было реализовано программное приложение, проводящее кластеризацию тестовых данных методом k-средних.

Пример 2.1. Имеются исходные образы (точки на плоскости) $n=500$, представленные в виде множества точек с координатами x и y (рисунок 2.1).

Найдем кластеризацию этих образов по k классам ($k=4$). Для этого выполним последовательно шаги рассмотренного алгоритма:

1. Пусть случайным образом выбираются начальные центры кластеров. Разбивать выборку будем на 4 кластера.

2. Для получения решения методом k-средних, вычисляется расстояние от текущей точки до 4 начальных центров, и точка относится в кластер, с наименьшим расстоянием до центра.

3. После того как все точки распределены по кластерам, пересчитываются центры кластеров, как среднее арифметическое всех координат. Таким образом, получаем новые центры.

4. Алгоритм повторяется, пока не буден достигнут критерий остановки.

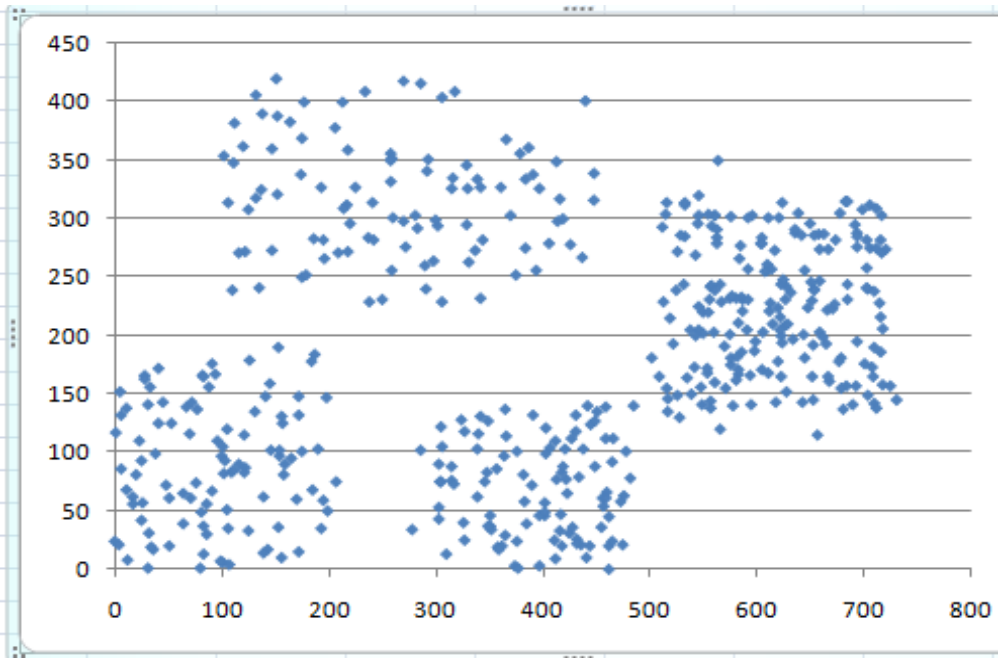

Рисунок 2.1 – Обучающая выборка (n=500)

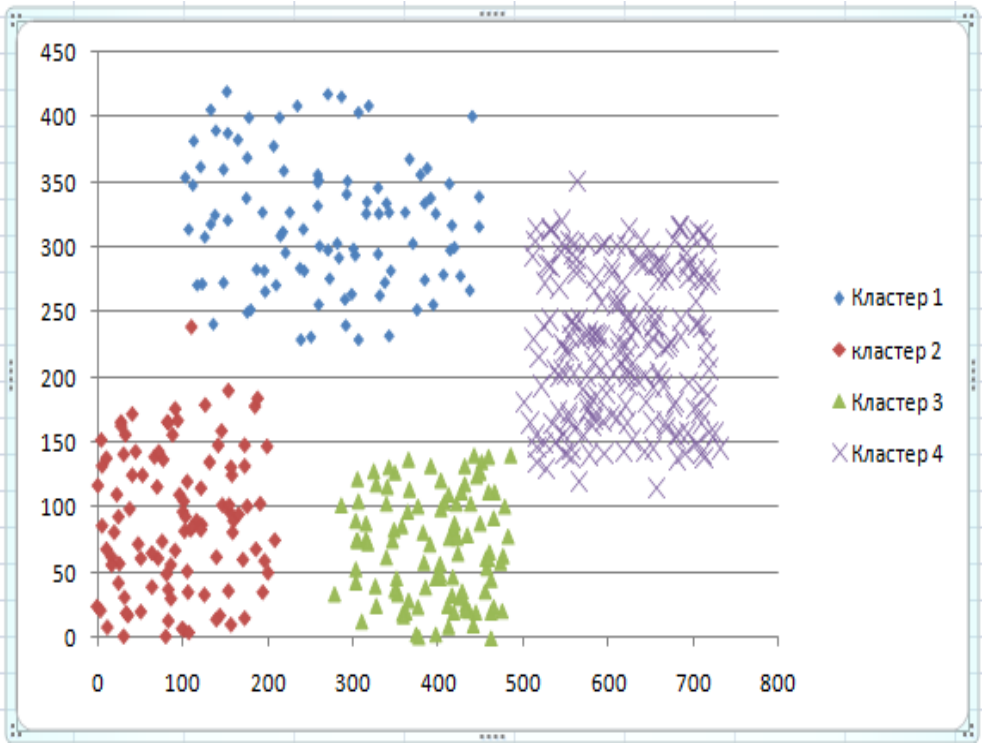

Рисунок 2.2 – Результат работы метода k-средних (n=500)

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства алгоритма k-средних:

- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

При практической реализации алгоритма k-средних, возникают следующие проблемы:

1) Алгоритм k-средних осуществляет локальную, но не глобальную минимизацию функционала Q. Поэтому гарантии «хорошей» кластеризации этот алгоритм не дает;

2) Алгоритм чувствителен к шумам. Качество кластеризации зависит от начальной расстановки центров кластеров. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;

2.2 Алгоритм классификации FOREL

Кластеры, получаемые этим алгоритмом [6], имеют сферическую форму. Количество кластеров зависит от радиуса сфер: чем меньше радиус, тем больше получается кластеров. Вначале признаки объектов нормируются так, чтобы значения всех признаков находились в диапазоне от нуля до единицы. Затем строится гиперсфера минимального радиуса R, которая охватывает все n точек. Если бы нам был нужен один кластер, то он был бы представлен именно этой начальной сферой. Но такое огрубление экспериментального материала нас обычно не устраивает, и мы пытаемся получить большее количество кластеров.

Для этого мы постепенно уменьшаем радиус сфер. Берем радиус 0,9*R и помещаем центр сферы в любую из имеющихся точек. Находим точки, расстояние до которых меньше радиуса, и вычисляем координаты центра тяжести этих «внутренних» точек. Переносим центр сферы в этот центр тяжести и снова находим внутренние точки. Сфера как бы плывет в сторону локального сгущения точек. Такая процедура определения внутренних точек и переноса центра сферы продолжается до тех пор, пока сфера не остановится, т. е. пока на очередном шаге мы не обнаружим, что состав внутренних точек, а следовательно и их центр тяжести, не меняется. Это значит, что сфера остановилась в области локального максимума плотности точек в признаковом пространстве.

Точки, оказавшиеся внутри остановившейся сферы, мы объявляем принадлежащими кластеру номер 1 и исключаем их из дальнейшего рассмотрения. Для оставшихся точек описанная выше процедура повторяется до тех пор, пока все точки не окажутся включенными в кластеры. Доказана сходимость алгоритма за конечное число шагов, однако легко видеть, что решение может быть не единственным. Так, на рис. 2.3 видно, что результат таксономии зависит от того, с какой первой точки был начат процесс.

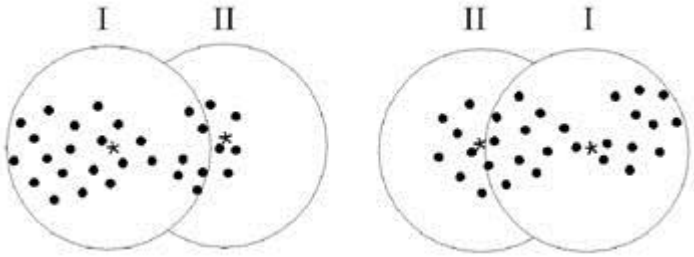

Рисунок 2.3 – Результаты кластеризации

Если начальную точку менять случайным образом, то может получиться несколько разных вариантов кластеризации, и тогда нужно останавливаться на таком варианте, который соответствует минимальному значению величины F. Функционал качества имеет вид:

$$F = \sum_{j=1}^{k} \sum_{x \in K_j} \rho(x, W_j),$$

где первое суммирование ведется по всем кластерам выборки, второе суммирование – по объектам x, принадлежащим текущему кластеру K_j, а W_j - центр текущего кластера, $\rho(x, y)$ - расстояние между объектами.

Описание алгоритма FOREL:

1. Задаем параметр R – радиус сферы.

2. Выбираем случайно из выборки точку x_0. Строим сферу $S(x_0, R)$ с центром в точке x_0 радиуса R.

3. Вычисляем центр тяжести $\mu = \frac{1}{|S(x_0, R)|} \sum_{x \in S(x_0, R)} x$. Перемещаем центр сферы x_0 в точку c.

4. Повторяем шаг 3 до тех пор, пока центр тяжести не перестанет изменяться.

5. Исключаем из выборки точки, принадлежащие $S(x_0, R)$.

6. Повторяем шаги 2-5, пока все точки не будут принадлежать построенным сферам.

В результате вычислительных экспериментов был выявлен ряд *достоинств* алгоритма FOREL:

1. Точность минимизации функционала качества (при удачном подборе параметра R).

2. Сходимость алгоритма.

3. Возможность операций над центрами кластеров — они известны в процессе работы алгоритма.

4. Возможность подсчета промежуточных функционалов качества, например, длины цепочки локальных сгущений.

5. Возможность проверки гипотез схожести и компактности в процессе работы алгоритма.

При практической реализации алгоритма FOREL, возникают следующие *проблемы:*

1. Плохая применимость алгоритма при плохой разделимости выборки на кластеры.

2. Неустойчивость алгоритма (кластеризация сильно зависит от выбора начального объекта).

3. Произвольное по количеству разбиение на кластеры.

4. Необходимость априорных знаний о ширине (диаметре) кластеров.

3 Генетические алгоритмы

3.1 Основные понятия генетических алгоритмов

Генетические алгоритмы (ГА) — это адаптивные методы поиска, в которых используются аналоги, как механизма генетического наследования, так и аналоги естественного отбора [8]. При этом сохраняется биологическая терминология в упрощенном виде и основные понятия линейной алгебры. Основной идеей генетических алгоритмов является организация «борьбы за существование» и «естественного отбора» среди этих пробных решений.

Генетические алгоритмы применяются для решения таких задач, как:

- поиск глобального экстремума многопараметрической функции,
- аппроксимация функций,
- задачи о кратчайшем пути,
- задачи размещения,
- настройка искусственной нейронной сети,
- игровые стратегии и т.д.

Введем основные понятия, применяемые в генетических алгоритмах:

Определение 3.1. *Вектор* — упорядоченный набор чисел, называемых *компонентами* вектора. Так как вектор можно представить в виде строки его

координат, то в дальнейшем понятия вектора и строки считаются идентичными.

Определение 3.2. *Булев вектор* — вектор, компоненты которого принимают значения из двух элементного (булева) множества, например, {0,1} или {-1,1}.

Определение 3.3. *Хромосома* — вектор (или строка) из каких-либо чисел. Если этот вектор представлен бинарной строкой из нулей и единиц, например, 1010011, то он получен либо с использованием *двоичного кодирования,* либо *кода Грея.* Каждая позиция (бит) хромосомы называется *геном*.

Определение 3.4. *Индивидуум* (генетический код, особь) — набор хромосом (вариант решения задачи). Обычно особь состоит из одной хромосомы, поэтому в дальнейшем особь и хромосома идентичные понятия.

Определение 3.5. *Кроссинговер* (скрещивание) — операция, при которой две хромосомы обмениваются своими частями. Например, 11|00&10|10 —> 1110&1000.

Определение 3.6. *Мутация* — случайное изменение одной или нескольких позиций в хромосоме. Например, 1010011 —> 1010001.

Определение 3.7. *Популяция* — совокупность индивидуумов.

Определение 3.8. *Функция приспособленности* (fitness function) представляет меру приспособленности данной особи в популяции. Эта функция играет важнейшую роль, поскольку позволяет оценить степень приспособленности конкретных особей в популяции и выбрать из них наиболее приспособленные (т.е. имеющие наибольшие значения функции приспособленности) в соответствии с эволюционным принципом выживания «сильнейших» (лучше всего приспособившихся).

Терминология ГА представляет собой синтез генетических и искусственных понятий. Так, для понятия, заимствованного из генетики, можно предъявить его искусственный (символический) аналог. Например, хромосома и строка.

3.2 Классический генетический алгоритм

Классический генетический алгоритм [1] состоит из следующих шагов:

1. Инициализация, или выбор исходной популяции хромосом;
2. Оценка приспособленности хромосом в популяции;
3. Проверка условия остановки алгоритма;
4. Селекция хромосом;
5. Применение генетических операторов;
6. Формирование новой популяции;
7. Выбор «наилучшей» хромосомы.

Схема работы классического генетического алгоритма представлена на рисунке 3.1.

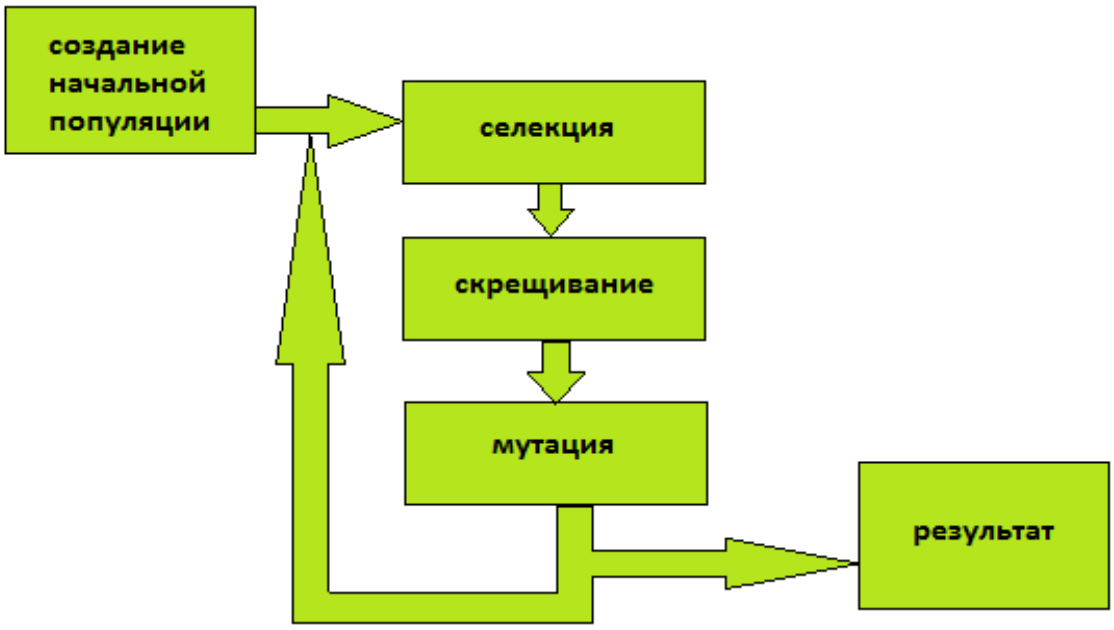

Рисунок 3.1 – Схема работы генетического алгоритма

Рассмотрим конкретные этапы этого алгоритма более подробно.

1) ***Инициализация***, т.е. формирование исходной популяции, заключается в случайном выборе заданного количества хромосом (особей), представленных числовыми последовательностями длины N, кодирующих номера кластеров. Если априорная информация о пространстве поиска

отсутствует, начальная популяция должна быть инициализирована равномерно в пространстве поиска (если это возможно).

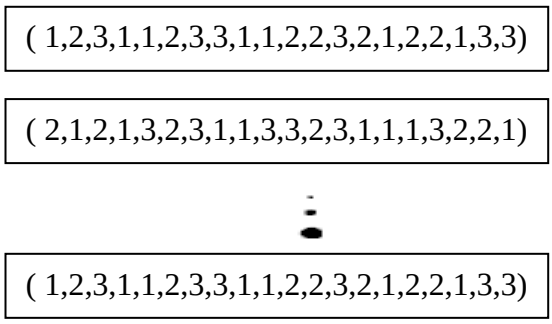

Рисунок 3.2 – Популяция хромосом: число классов k=3, число объектов N=20.

2) ***Оценивание приспособленности хромосом в популяции*** состоит в расчете функции приспособленности для каждой хромосомы этой популяции [2]. Чем больше значение этой функции, тем выше «качество» хромосомы. Форма функции приспособленности зависит от характера решаемой задачи. Обычно функция приспособленности принимает положительные значения.

3) ***Проверка условия остановки алгоритма***. Определение условия остановки генетического алгоритма зависит от его конкретного применения. Остановка алгоритма может произойти в случае, когда его выполнение не приводит к улучшению уже достигнутого значения. Алгоритм может быть остановлен после выполнения заданного максимального числа вычислений функции приспособленности, а так же после выполнения заданного максимального числа поколений без улучшения решения, либо после максимального числа поколений без улучшения средней приспособленности по популяции.

4) ***Селекция хромосом*** заключается в наборе (по рассчитанным на втором этапе значениям функции приспособленности) тех хромосом, которые будут участвовать в создании потомков для следующей популяции, т.е. для очередного поколения. Такой выбор производится согласно принципу естественного отбора, по которому наибольшие шансы на участие в создании новых особей имеют хромосомы с наибольшими значениями

функции приспособленности. Существуют различные методы селекции. Наиболее распространены следующие базовые типы селекции [7].

4.1) Пропорциональная селекция

В пропорциональной селекции вероятность индивида быть отобранным пропорциональна его пригодности. Вероятность вычисляется следующим образом (для задачи минимизации):

$$P(X^i) = \frac{-\mathrm{Q}(X^i) + C}{r * C - \sum_{j=1}^{r} \mathrm{Q}(X^j)}, \tag{3.1}$$

где r - размер популяции, $Q(X^i)$ - пригодность индивида.

$$C : P(X^i) \geq 0, \forall i, \ \sum_{j=1}^{r} P(X^j) = 1. \tag{3.2}$$

Пропорциональная селекция обладает следующими недостатками: преждевременная сходимость и стагнация.

Стагнация возникает, когда на определенном этапе поиска все индивиды получают относительно высокую и примерно равную пригодность, что приводит к очень низкому селективному давлению (наилучшее решение лишь немного предпочитается худшему).

Преждевременная сходимость (проблема супериндивида) возникает, когда на ранних этапах появляется индивид с пригодностью намного большей, чем у других индивидов в популяции, но очень плохой с точки зрения решаемой задачи. Вероятность супериндивида быть отобранным стремится к единице, в то время как вероятности других членов популяции – к нулю. В итоге он копирует себя в следующее поколение и вскоре «широкий» поиск прекращается.

4.2) Ранговая селекция

При применении ранговой селекции в задаче минимизации индивиды популяции ранжируются в соответствии с их пригодностью: $R_i < R_j$ если $f(X^i) \leq f(X^j)$. Тогда

$$P(X^i) = \frac{R_i}{\sum_{k=1}^{r} R_k} = \frac{i}{\sum_{k=1}^{r} i} = \frac{2i}{r(r+1)}, \text{ где } \sum_{j=1}^{r} P(X^j) = 1. \quad (3.3)$$

Ранговая селекция устраняет недостатки пропорциональной: нет стагнации, т.к. даже к концу работы алгоритма $P(X^1) \neq P(X^2) \neq \ldots$, нет преждевременной сходимости, т.к. нет индивидов с вероятностью отбора близкой к единице.

4.3) Турнирная селекция

В турнирной селекции для отбора индивида создается группа из t ($t \geq 2$) индивидов, выбранных случайным образом. Индивид с наибольшей пригодностью в группе отбирается, остальные – отбрасываются. Параметр t называется размером турнира. Наиболее популярным является бинарный турнир. Этот тип селекции не требует сортировки популяции и вычисления пригодности для всех индивидов. Недостатки: худший индивид никогда не выбирается.

4.4) Селекция с усечением

В процессе селекции с усечением с порогом τ, только доля τ из всех лучших индивидов может быть отобрана, причем в этой доле каждый имеет одинаковую вероятность отбора.

4.5) Элитарная селекция

Как минимум одна копия лучшего индивида всегда переходит в следующее поколение. Преимущества: гарантия сходимости. Недостатки: большой риск захвата локальным оптимумом.

4.6) Инбридинг и аутбридинг.

Метод инбридинга построен на формировании пары на основе близкого «родства». Под «родством» здесь понимается расстояние между членами популяции, как в смысле геометрического расстояния особей в пространстве параметров, так и Хеммингово расстояние между генотипами. Потому различают генотипный и фенотипный (или географический) инбридинг.

Первый член пары для скрещивания выбирается случайно, а вторым с большей вероятностью будет максимально близкая к нему особь.

При аутбридинге, пары формируются на основе дальнего «родства», для максимально далеких особей.

Однако два этих способа по-разному влияют на поведение генетического алгоритма. Так инбридинг можно охарактеризовать свойством концентрации поиска в локальных узлах, что фактически приводит к разбиению популяции на отдельные локальные группы вокруг подозрительных на экстремум участков ландшафта, напротив аутбридинг как раз направлен на предупреждение сходимости алгоритма к уже найденным решениям, заставляя алгоритм просматривать новые, неисследованные области.

В результате процесса селекции создается родительская популяция, с численностью, равной численности текущей популяции.

5) ***Применение генетических операторов*** к хромосомам, отобранным с помощью селекции, приводит к формированию новой популяции потомков [3] от созданной на предыдущем шаге родительской популяции.

В классическом генетическом алгоритме применяются два [4] основных генетических оператора: *оператор скрещивания* (crossover) и *оператор мутации* (mutation). Однако следует отметить, что оператор мутации играет явно второстепенную роль по сравнению с оператором скрещивания. Это означает, что скрещивание в классическом генетическом алгоритме производится практически всегда, тогда как мутация – достаточно редко. Вероятность скрещивания, как правило, достаточно велика (обычно $0{,}5 \le p_c \le 1$), тогда как вероятность мутации устанавливается весьма малой (чаще всего $0 \le p_m \le 0{,}1$). Это следует из аналогии с миром живых организмов, где мутации происходят чрезвычайно редко.

В генетическом алгоритме мутация хромосом может выполняться на популяции родителей перед скрещиванием либо на популяции потомков, образованных в результате скрещивания.

5.1) Оператор скрещивания:

При скрещивании отобранные индивиды (родители) по заданному правилу передают части своих хромосом. Потомок может унаследовать только те гены, которые есть у его родителей. Наиболее популярным типом скрещивания является одноточечное скрещивание – случайно выбирается точка разрыва, родительские хромосомы разрываются в этой точке и обмениваются правыми частями. Скрещивание осуществляется с вероятностью p_{xover}, иначе с вероятностью $(1-p_{xover})$ родители клонируются в следующее поколение.

Возможно использование одного или нескольких операторов скрещивания для выбранного представления. Необходимо учитывать следующие важные моменты:

- Потомок должен унаследовать гены от каждого родителя. В противном случае оператор скрещивания становится оператором мутации;
- Оператор скрещивания может быть разработан с учетом представления, чтобы скрещивание не было всегда катастрофичным (чтобы потомки не были всегда хуже родителей);
- Скрещивание должно производить допустимые решения.

Наиболее распространены следующие типы скрещивания для дискретного представления.

5.1.а) Одноточечное скрещивание

Наиболее популярным типом скрещивания является одноточечное скрещивание – случайно выбирается точка разрыва, родительские хромосомы разрываются в этой точке и обмениваются правыми частями.

5.1.б) Двухточечное скрещивание

При двухточечном скрещивании хромосома как бы замыкается в кольцо, выбираются 2 точки разрыва, родители обмениваются частями.

5.1.в) Равномерное скрещивание

При равномерном скрещивании потомок может унаследовать с равной вероятностью гены любого из родителей.

Равномерное скрещивание по всей популяции

5.1.г) Равномерное скрещивание по всей популяции (uniform gene pool recombination) получается применением равномерного скрещивания ко всем членам популяции, т.е. потомок может унаследовать любой ген, имеющийся в популяции в заданной позиции хромосомы.

5.2) Оператор мутации:

В генетическом алгоритме мутация рассматривается как метод восстановления потерянного генетического материала, а не как поиск лучшего решения. Обычно мутация применяется к генам с очень низкой вероятностью $p_m \in [0.001; 0.01]$. Возможно использование одного или нескольких операторов мутации в алгоритме для выбранного типа представления. Необходимо учитывать следующие важные моменты:

- По крайней мере, один оператор мутации должен быть в алгоритме;
- Оператор мутации должен позволять достигнуть любой части пространства поиска;
- Величина мутации важна и должна быть управляемой;
- Мутация должна производить допустимые решения.

Мутация обычно происходит с вероятностью p_m для каждого гена. Хорошим эмпирическим правилом считается выбор вероятности мутации равным $p_m = 1/n$, где n - число генов в хромосоме (в среднем хотя бы один ген будет подвержен мутации).

6) ***Формирование новой популяции***. Хромосомы, полученные в результате применения генетических операторов к хромосомам временной родительской популяции, включаются в состав новой популяции. Она становится так называемой текущей популяцией для данной итерации генетического алгоритма. На каждой очередной итерации рассчитываются значения функции приспособленности для всех хромосом этой популяции, после чего проверяется условие остановки алгоритма и либо фиксируется

результат в виде хромосомы с наибольшим значением функции приспособленности, либо осуществляется переход к следующему шагу генетического алгоритма, т.е. к селекции. В классическом генетическом алгоритме вся предшествующая популяция хромосом замещается новой популяцией потомков, имеющей ту же численность.

7) ***Выбор «наилучшей» хромосомы***. Если условие остановки алгоритма выполнено, то следует вывести результат работы, т.е. представить искомое решение задачи. Лучшим решением считается хромосома с наибольшим значением функции приспособленности [5].

3.3 Генетический алгоритм кластеризации с детерминированным числом кластеров

В бакалаврской работе предлагается генетический алгоритм для решения задачи кластеризации данных с детерминированным числом кластеров.

Главным достоинством генетических алгоритмов в данном применении является то, что они ищут глобальное оптимальное решение.

Большинство популярных алгоритмов кластеризации данных выбирают начальное решение, которое затем изменяется в ту или иную сторону. Таким образом, получается хорошее разбиение, но не всегда – самое оптимальное. Операторы рекомбинации и мутации позволяют получить решения, непохожие на исходные.

Разработанный генетический алгоритм основан на идее объединения в один кластер объектов в областях их наибольшего сгущения.

1. *Цель* генетического алгоритма для решения задачи кластеризации данных – разбить выборку на k кластеров, чтобы сумма расстояний от объектов кластеров до центров кластеров была минимальной по всем кластерам. Т.е. наша задача - выделить группы максимально близких друг к

Страница изъята

Страница изъята

На рис. 3.4 показан результат работы алгоритма на тестовом примере, состоящем из 100 объектов

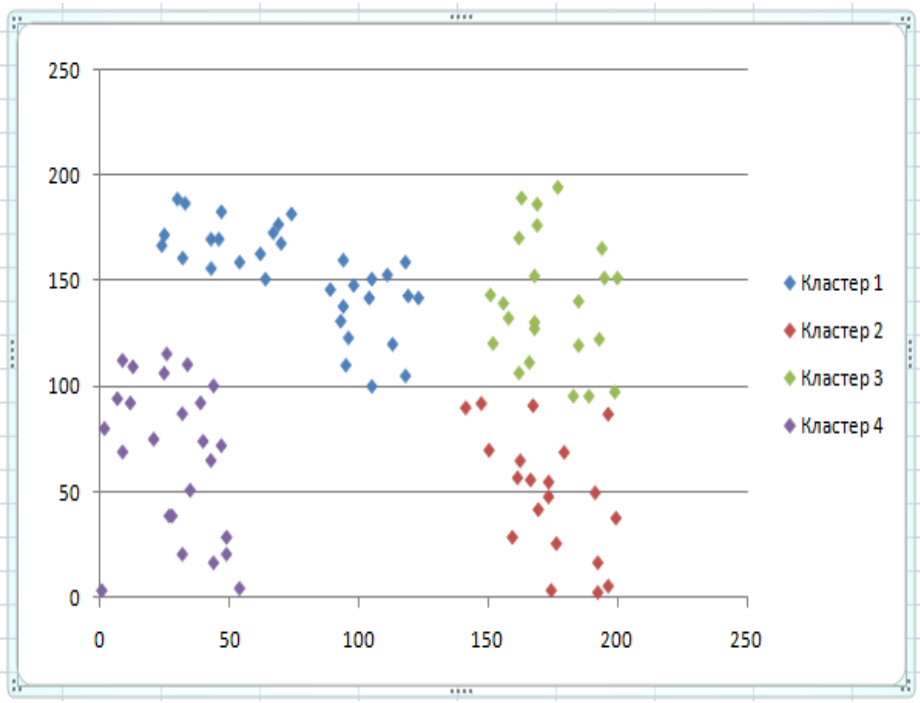

Рисунок 3.4 – Результат работы генетического алгоритма детерминированным числом кластеров (n=100)

3.4 Генетический алгоритм кластеризации с заданным радиусом кластеров

Также в данной работе предлагается генетический алгоритм для решения задачи кластеризации данных с заданным радиусом кластеров.

Разработанный генетический алгоритм основан на идее генетического алгоритма кластеризации с детерминированным числом кластеров, но решает другой тип задачи кластеризации, а именно – с заданным ограничением на пространственные характеристики кластера (глава 1).

Цель генетического алгоритма с заданным радиусом кластеров – разбить выборку на кластеры фиксированного радиуса.

Страница изъята

Страница изъята

На рисунке 3.6 показан результат работы генетического алгоритма заданным радиусом кластеров на тестовом примере, состоящим из 100 объектов, и заданным радиусом $R = 18$

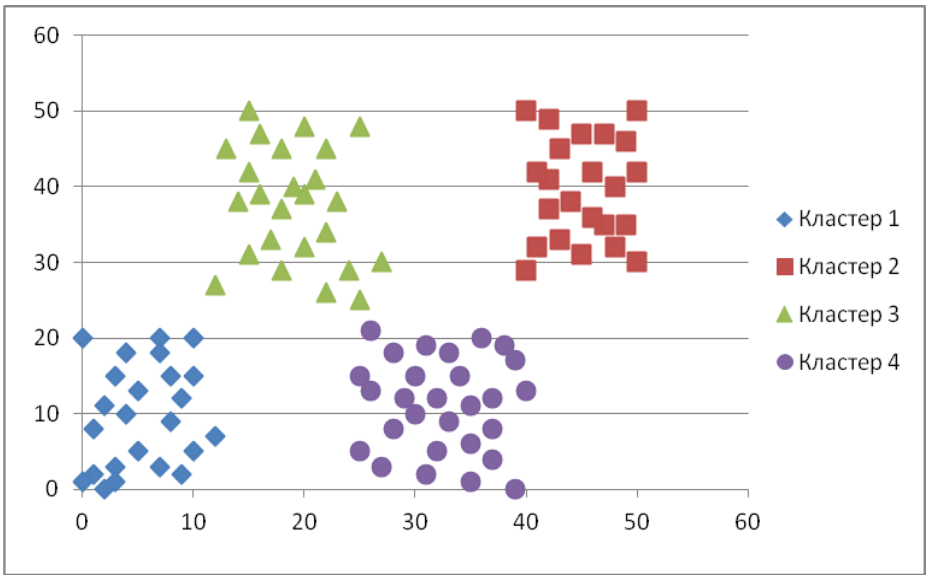

Рисунок 3.6 – Результат работы генетического алгоритма заданным радиусом кластеров (n=100, R = 18)

4 Сравнение предложенных алгоритмов со стандартными методами кластеризации

В бакалаврской работе был проведен обзор и сравнительный анализ стандартных алгоритмов кластеризации данных с разработанными генетическими алгоритмами. В таблице 1 представлены основные характеристики рассматриваемых методов.

Таблица 1 – Сравнительная таблица алгоритмов

Алгоритм кластеризации	Форма кластеров	Входные данные	Вычислительная сложность

K-средних	Центроид	Число кластеров, начальные центры	$O(nkl)$
FOREL	Произвольная	Радиус поиска	$O(n^2)$
Генетический алгоритм с детерминированным числом кластеров	Центроид	Число кластеров	$O(n^2k)$
Генетический алгоритм с заданным радиусом кластеров	Произвольная	Радиус поиска	$O(n^2k^2)$

Здесь n – количество точек в обучающей выборке, k – количество кластеров, l – количество итераций алгоритмов.

На рис. 3.8 показаны результаты сравнения работы алгоритмов на тестовом примере, состоящем из 50 точек, в первых двух введено k=4, во вторых двух R=10.

Проведем более детальное сравнение предложенных в работе генетических алгоритмов кластеризации и изученных классических методов (k-средних и FOREL).

Алгоритм k–средних крайне чувствителен к выбору начальных приближений центров, а также к выбросам, которые могут искажать среднее. Недостатком алгоритма также является необходимость задавать число кластеров, исследователь может при этом исходить из некоторых априорных знаний, соображений.

Результаты работы алгоритма FOREL зависят от выбранного радиуса R, а также от выбора начального объекта. Для алгоритма FOREL можно выделить следующие проблемы: плохая применимость алгоритма при плохой разделимости выборки на кластеры, произвольное по количеству разбиение на кластеры, необходимость априорных знаний о диаметре кластеров.

Большинство популярных алгоритмов кластеризации требуют введения дополнительных параметров, которые затем изменяется в ту или иную сторону. Таким образом, получается хорошее разбиение, но не всегда – самое

оптимальное. Главным достоинством генетических алгоритмов в данном применении является то, что они ищут глобальное оптимальное решение. Операторы скрещивания и мутации выбивают популяцию из локального экстремума и позволяют получить решения, непохожие на исходные.

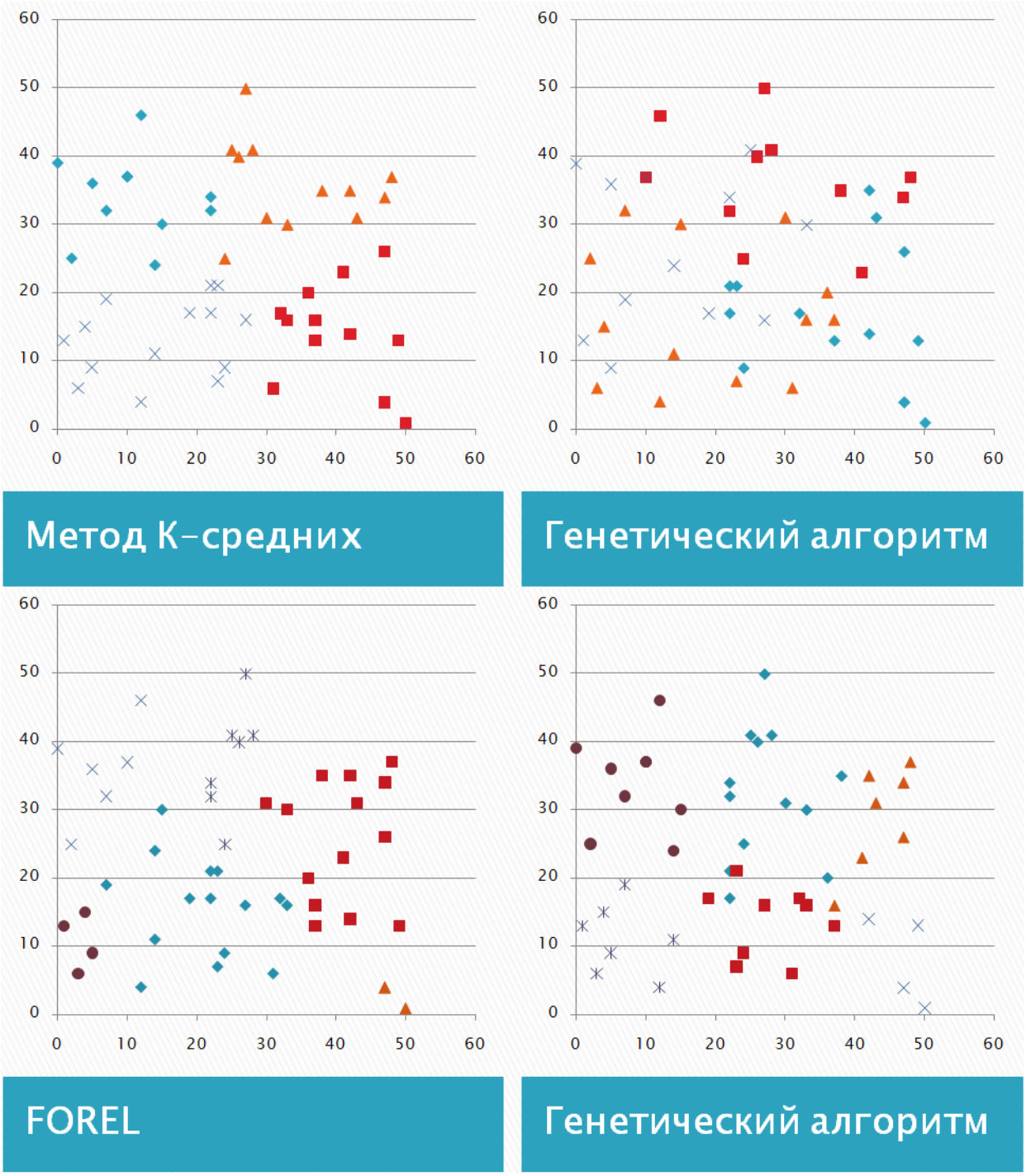

Рисунок 4.1 – Сравнение методов кластеризации

Предложенные генетические алгоритмы эффективно функционируют при обработке массивов большой размерности, поскольку в них оптимально

сочетаются целенаправленный поиск и элементы случайности, направленные на выбивание целевой функции из локальных минимумов. Никаких предварительных условий для их использования не требуется. Главным условием кластеризации данных является правильная алгоритмизация расчета значений функции приспособленности.

5.Решение практической задачи кластеризации

В работе решается практическая задача кластеризации многомерных данных – кластеризация ведущих российских банков по основным показателям их финансовой деятельности.

5.1 Описание статистики

В работе решается практическая задача кластеризации 50 ведущих российских банков по основным показателям их финансовой деятельности. Для решения задачи использовалась официальная статистика отчетности кредитных организаций РФ по показателям их деятельности за март 2015г., публикуемая на сайте Банка России (http://www.banki.ru/). В статистике для каждого из банков приводятся показатели его деятельности: активы нетто, чистая прибыль, капитал, кредитный портфель, просроченная задолженность в кредитном портфеле, вклады физических лиц. Всего рассматривалось шесть показателей, все они являются числовыми.

Кластеризация данных проводилась с помощью четырех методов:

- метода k–средних,
- алгоритма FOREL,
- генетического алгоритма кластеризации с детерминированным числом кластеров
- генетического алгоритма кластеризации с заданным радиусом кластеров.

5.2 Решение задачи кластеризации методом k-средних и генетическим алгоритмом кластеризации с детерминированным числом кластеров

Согласно алгоритму метода k–средних изначально было задано число кластеров равное 3. На первом этапе центры кластеров были заданы случайным образом. Затем алгоритм вычислил, согласно системе, приведенной в главе 2.1, новые центры кластеров. В результате было найдено разбиение по 3 кластерам. Время вычисления составило 6 итераций. Результаты решения задачи представлены в Таблице 2.

Таблица 2 – Результаты кластеризации методом k–средних для 3 кластеров

№ кластера	Наименование
1	Сбербанк России
2	НОМОС-Банк, ЮниКредит Банк, Росбанк, Промсвязьбанк, Райффайзенбанк, Национальный Клиринговый Центр, Банк «Санкт-Петербург», Московский Кредитный банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ак Барс, Ситибанк, МДМ Банк, Связб Банк, Нордеа Банк, Петрокоммерц, ИНГ Банк Глобэкс, Зенит, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение, Бинбанк, Уральский Банк Реконструкции и Развития, Московский Индустриальный Банк, Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Российский Капитал.
3	ВТБ, Газпромбанк, ВТБ 24, Россельхозбанк, Банк Москвы, Альфа-Банк

Затем задача кластеризация на три кластера была решена с помощью генетического алгоритма с детерминированным числом кластеров. Полученные результаты представлены в таблице 3.

Далее данная задача была решена теми же методами для 4 кластеров. Результаты решения задачи приведены в таблице 4 и таблице 5, для метода k–средних и генетического алгоритма, соответственно.

Таблица 3 – Результаты кластеризации генетическим алгоритмом с детерминированным числом кластеров для 3 кластеров

№ кластера	**Наименование**
1	Сбербанк России
2	НОМОС-Банк, ЮниКредит Банк, Райффайзенбанк, Национальный Клиринговый Центр, Банк «Санкт-Петербург», Московский Кредитный банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ак Барс, Ситибанк, МДМ Банк, Связб Банк, Нордеа Банк, Петрокоммерц, ИНГ Банк Глобэкс, Зенит, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение, Бинбанк, Уральский Банк Реконструкции и Развития, Московский Индустриальный Банк, Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Российский Капитал.
3	ВТБ, Газпромбанк, ВТБ 24, Россельхозбанк, Росбанк, Промсвязьбанк, Банк Москвы, Альфа-Банк

Таблица 4 – Результаты кластеризации методом k–средних для 4 кластеров

№ кластера	**Наименование**
1	Сбербанк России
2	ВТБ, Газпромбанк, ВТБ 24, Россельхозбанк, Банк Москвы
3	Альфа-Банк, НОМОС-Банк, ЮниКредит Банк, Росбанк, Промсвязьбанк, Райффайзенбанк, Национальный Клиринговый Центр, Банк «Санкт-Петербург», Московский Кредитный Банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ак Барс
4	Ситибанк, МДМ Банк, Связб Банк, Нордеа Банк, Петрокоммерц, ИНГ Банк, Глобэкс, Зенит, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение, Бинбанк, Уральский Банк Реконструкции и Развития, Моск.Индустр. Банк,

	Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Рос. Капитал.

Таблица 5 – Результаты кластеризации генетическим алгоритмом с детерминированным числом кластеров для 4 кластеров

№ кластера	**Наименование**
1	Сбербанк России
2	ВТБ, Газпромбанк, ВТБ 24, Росбанк, Альфа-Банк, Россельхозбанк, Промсвязьбанк, Банк Москвы
3	НОМОС-Банк, ЮниКредит Банк, Райффайзенбанк, Национальный Клиринговый Центр, Банк «Санкт-Петербург», Зенит, Московский Кредитный Банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ситибанк, МДМ Банк, Связь Банк, Нордеа Банк, Петрокоммерц, Ак Барс, ИНГ Банк, Глобэкс, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение
4	Бинбанк, Уральский Банк Реконструкции и Развития, Моск.Индустр. Банк, Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Рос. Капитал.

5.3 Решение задачи кластеризации методами FOREL и генетическим алгоритмом кластеризации с заданным радиусом кластеров

Теперь для той же статистики показателей банков найдем решение задачи кластеризации второго типа – количество классов заранее неизвестно, однако заданы ограничения на число объектов в кластере. Вначале решим задачу методом FOREL. Для этого алгоритма необходимо указать максимальный радиус кластера R. Результаты решения задачи кластеризации алгоритмом FOREL с радиусом R=19000000 приведены в таблице 6.

Далее решим эту же задачу с помощью генетического алгоритма с заданным радиусом кластеров (значение радиуса R = 19000000).

Результаты работы алгоритма представлены в таблице 7.

Таблица 6 – Результаты кластеризации алгоритмом FOREL

№ кластера	**Наименование**
1	Сбербанк России
2	ВТБ
3	Газпромбанк
4	ВТБ 24, Россельхозбанк, Альфа-Банк, Банк «Санкт-Петербург», Московский Кредитный банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ак Барс, Ситибанк, МДМ Банк, Связь Банк, Нордеа Банк, Петрокоммерц, ИНГ Банк, Глобэкс, Зенит, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение, Бинбанк, Уральский Банк Реконструкции и Развития, Московский Индустриальный Банк, Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Российский Капитал, НОМОС-Банк, ЮниКредит Банк, Росбанк, Промсвязьбанк, Райффайзенбанк, Национальный Клиринговый Центр.

Таблица 7 – Результаты кластеризации генетическим алгоритмом

№ кластера	**Наименование**
1	Сбербанк России
2	ВТБ
3	Газпромбанк
4	ВТБ 24, Россельхозбанк, Альфа-Банк, Банк «Санкт-Петербург», Московский Кредитный банк, Уралсиб, Ханты-Мансийский Банк, Хоум Кредит Банк, Ак Барс, Ситибанк, МДМ Банк, Связь Банк, Нордеа Банк, Петрокоммерц, ИНГ Банк, Глобэкс, Зенит, Восточный Экспресс Банк, Национальный Банк «Траст», Возрождение, Бинбанк, Уральский Банк Реконструкции и Развития, Московский Индустриальный Банк, Новикомбанк, СМП Банк, ОТП Банк, Внешпромбанк, Кредит Европа Банк, Совкомбанк, Транскапиталбанк, Дойче Банк, СКБ-Банк, МСП

	Банк, РосЕвроБанк, Татфондбанк, Ренессанс Кредит, Инвестторгбанк, Абсолют Банк, Росгосстрах Банк, Российский Капитал, НОМОС-Банк, ЮниКредит Банк, Росбанк, Промсвязьбанк, Райффайзенбанк, Национальный Клиринговый Центр.

5.4 Сравнение полученных результатов

На основе реальной статистики была проведена кластеризация 50 крупных российских банков по их финансовым показателям. Для сравнения эффективности разбиения объектов по кластерам была проведена оценка качества кластеризации с помощью отношения функционалов качества Φ_0 / Φ_1 из главы 1. Результаты качества кластеризации приведены в Таблице 8.

Таблица 8 – Значение функционалов качества кластеризации

Алгоритм кластеризации	**Число кластеров**	**Результат статистических наблюдений**
Метод k–средних	3	0, 074
Генетический алгоритм с детерминированным числом кластеров	3	0,073
Метод k–средних	4	0, 079
Генетический алгоритм с детерминированным числом кластеров	4	0, 075
FOREL	4	0, 014
Генетический алгоритм с заданным радиусом кластеров	4	0, 014

Результаты, представленные в таблице, показывают, что при решении задачи статистическим методом k–средних при увеличении числа кластеров качество кластеризации ухудшилось. Также можно отметить, что решение задачи генетическим алгоритмом с детерминированным числом кластеров дало не такое явное ухудшение функционала качества при увеличении числа кластеров.

Анализируя результаты разбиения объектов по кластерам можно сделать следующие выводы: объекты, имеющие между собой близкие по значению данные по каждому из показателей, попадают в один и тот же кластер. Таким образом, создается группа сходных объектов, что является одной из задач применения кластеризации.

Из полученных результатов можно сделать следующее заключение: решение задачи генетическим алгоритмом с детерминированным числом кластеров дает наилучший результат при разбиении данных на 3 кластера. Данный результат проиллюстрирован на рис. 5.1. Метод FOREL и генетический алгоритм с заданным числом кластеров показали одинаково хорошие результаты разбиения выборки на 4 кластера, результаты на рисунке 5.3.

Основываясь на полученных значениях функционалов качества кластеризации, лучший результат разбиения соответствует минимальному значению 0, 014 и принадлежит методу FOREL и генетическому алгоритму с заданным числом кластеров.

5.5 Визуализация результатов

Визуализация данных – задача, с которой сталкивается в своей работе любой исследователь. К задаче визуализации данных сводится проблема представления в наглядной форме данных эксперимента или результатов теоретического исследования.

Для визуализации могут быть использованы 1–, 2– и 3–мерные пространства отображений, но я в своем рассмотрении ограничусь способом визуализации с помощью 2–мерного пространства, поскольку именно в таком виде отношения между объектами выглядят наиболее наглядно. На рисунках 5.1–5.3 представлен срез разбиения данных для каждого из рассмотренных методов. Визуализируются кластеры, дающие наилучшее разбиение.

Генетический алгоритм кластеризации с детерминированным числом кластеров показывает наилучшее качество кластеризации при разбиении объектов на 3 кластера. Результат разбиения для двумерного пространства представлен на рисунке 5.1

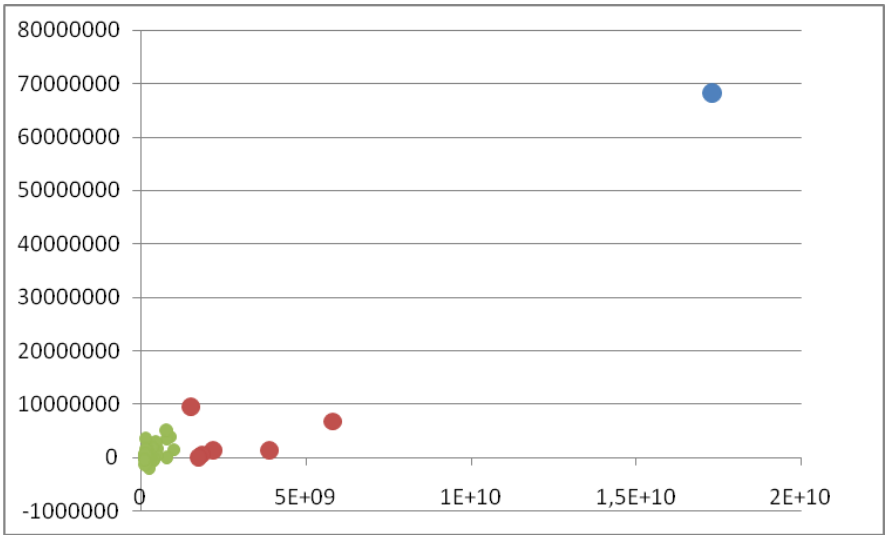

Рисунок 5.1 – Результат разбиения методом генетическим алгоритмом с детерминированным числом кластеров

Также генетический алгоритм с детерминированным числом кластеров дает наилучшее качество кластеризации при разбиении объектов на 4 кластера (среди алгоритмов решающих задачу с детерминированным числом кластеров). Результат разбиения для двумерного пространства представлен на рисунке 5.2

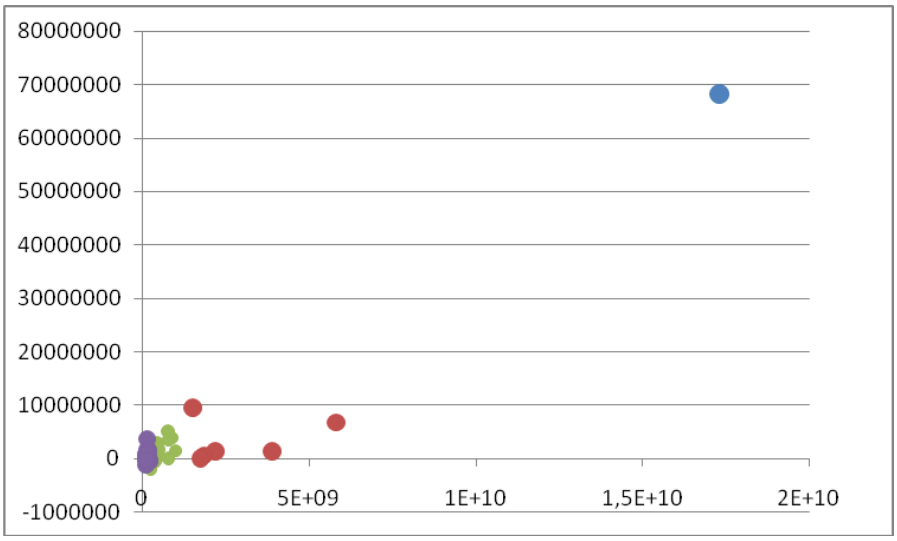

Рисунок 5.2 – Результат разбиения на 4 кластера

Алгоритм FOREL и генетический алгоритм с заданным радиусом кластеров показали наилучший результат разбиения выборки на 4 кластера. Результат разбиения представлен на рисунке 5.3.

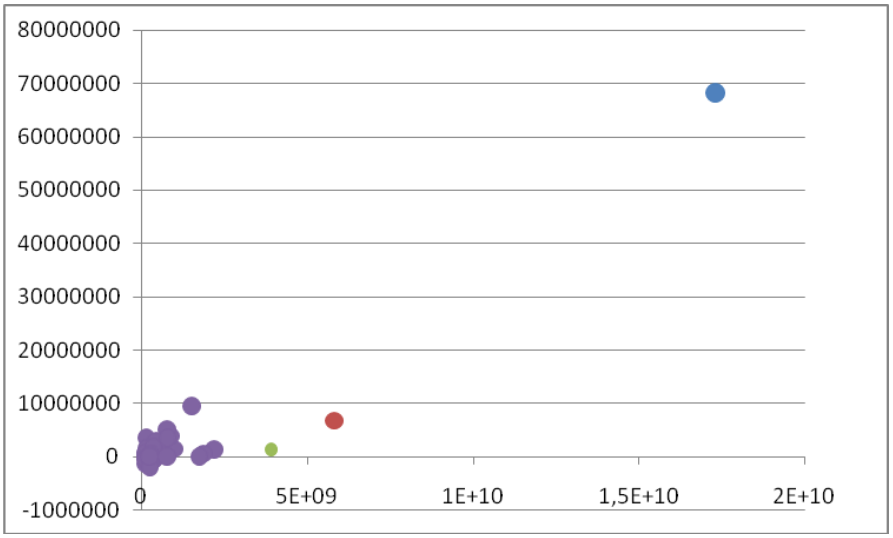

Рисунок 5.2 – Результат разбиения на 4 кластера методами FOREL и генетическим алгоритмом с заданным радиусом кластеров

Представленные рисунки показывают, что с увеличением числа кластеров расстояние между кластерами уменьшается, кластеры более плотно располагаются друг к другу. Таким образом, мы можем регулировать значение той или иной меры сходства между объектами.

ЗАКЛЮЧЕНИЕ

В работе получены следующие результаты:

- Изучены основные алгоритмы кластеризации многомерных данных.
- Реализованы методы кластеризации k-средних и FOREL.
- Разработан генетический алгоритм для решения задачи кластеризации многомерных данных с заданным количеством кластеров.
- Разработан генетический алгоритм для решения задачи кластеризации с заданным размером кластеров.
- Создано программное приложение, реализующее работу предложенных алгоритмов, а также изученных классических алгоритмов кластеризации.
- Проведено сравнение изученных и предложенных методов по их вычислительной сложности и результатам работы.
- Решена практическая задача кластеризации 50 российских банков по показателям их финансовой деятельности.
- Проведено сравнение результатов, полученных в результате работы каждого метода.

Результаты работы докладывались и опубликованы на международной научно-технической конференции студентов, аспирантов и молодых ученых «Проспект Свободный – 2015» (г. Красноярск, 2015), на международной научно-технической конференции студентов, аспирантов и молодых ученых «Проспект Свободный – 2016» (г. Красноярск, 2016).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Батищев, Д.И. Генетические алгоритмы решения экстремальных задач / Д.И. Батищев. – Воронеж : ВГТУ 1995. — 62с.
2. Батыршин, И.З. Нечеткие гибридные системы. Теория и практика / под ред. Н.Г. Ярушкиной. – Москва : ФИЗМАТЛИТ, 2007. – 208 с.
3. Вороновский, Г.К. Генетические алгоритмы, искусственные нейронные сети и проблемы виртуальной реальности / Г.К. Вороновский. – Харьков : ОСНОВА, 1997. – 112с.
4. Дарвин, Ч. О происхождении видов путём естественного отбора или сохранении благоприятствуемых пород в борьбе за жизнь / Ч. Дарвин. – Москва : АН СССР, 1939. – 322 с.
5. Еремеев, А.В. Генетические алгоритмы и оптимизация: учебное пособие/ А.В. Ермеев. – Омск : Издательство «Омский государственный университет», 2008. – 48 с.
6. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. 270 с.
7. Мандель, И.Д. Кластерный анализ/ И.Д. Мандель. – Москва : Финансы и статистика, 1988. – 176 с.
8. Панченко, Т. В. Генетические алгоритмы : учебно-методическое пособие / под ред. Ю. Ю. Тарасевича. — Астрахань : Издательский дом «Астраханский университет», 2007. — 87 с.
9. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы. / Перевод с польского И.Д. Рудинского. — М.: Горячая линия - Телеком, 2006. — 452с.
10. Воронцов, К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования / К. В. Воронцов. — М.: МГУ, 2007. — 18 с.
11. Миркин, Б. Г. Методы кластер – анализа для поддержки принятия решений: обзор: препринт WP7/2011/03/ Б.Г. Миркин. — М.: Изд. Дом

Национального исследовательского университета «Высшая школа экономики», 2011. — 88 с.

12. Дюран, Б. Кластерный анализ: пер. с англ. Е. З. Демиденко под ред. А.Я. Боярского / Б. Дюран, — П. Одел. М.: «Статистика», 1977. — 128 с.
13. Местецкий, Л. М. Математические методы распознавания образов / Л. М. Местецкий. — М.: МГУ, 2002. — 139 с.
14. Потапов, А. С. Распознавание образов и машинное восприятие / А. С. Потапов. — М.: "Политехника", 2007. — 552 с.
15. Аксенов, С.В. Организация и использование нейронных сетей (методы и технологии) / под общ. ред. В.Б. Новосельцева. – Томск : Изд-во НТЛ, 2006. – 128 с.
16. Гладков, Л.А. Генетические алгоритмы / Л.А. Гладков, В.В. Курейчик, В.М. Курейчик. – М. : ФИЗМАТЛИТ, 2006. – 320 с.
17. Лепский, А. Е. Математические методы распознавания образов / А.Е. Лепский, А.Г. Броневич. — Таганрог: Изд-во ТТИ ЮФУ, 2009. 155 с.
18. Мищенко, В.А. Использование генетических алгоритмов в обучении нейронных сетей // В.А Мищенко, А.А. Коробкин Современные проблемы науки и образования, 2011. – № 6;
19. Олдендерфер, М.К. Кластерный анализ / М.К. Олдендерфер, М.С. Блэшфилд – М.: Финансы и статистика, 1985г. – 227 с.
20. Уиллиамс, У. Т., Ланс Д. Н. Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. – М.: Наука, 1986.-С. 269–301.
21. Шуметов, В. Г., Кластерный анализ: подход с применением ЭВМ / В. Г. Шуметов, Л.В. Шуметов. – Орел : ОрелГТУ, 2000. – 119 с.
22. Muhlenbein H., Voigt H.-M. Gene Pool Recombination in Genetic Algorithms. In Proc. Of the Metaheuristics Inter. Conf., 1995.
23. Гладков, Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы / Под ред. В.М. Курейчика. — 2-е изд., испр. и доп. — М.: ФИЗМАТЛИТ, 2006. — 320 с.

24. Семенов, М.Г. Стохастические методы решения задачи о рюкзаке / М. Г. Семенов // Сборник материалов международной конференции студентов, аспирантов и молодых ученых «Проспект Свободный-2015», г. Красноярск, СФУ, 2015. с. 43-46.
25. Семенов, М.Г. Разработка генетического алгоритма для решения задачи кластеризации данных / М. Г. Семенов // Сборник материалов международной научно-технической конференции студентов, аспирантов и молодых ученых «Проспект Свободный-2016», г. Красноярск, СФУ, 2016.

ПРИЛОЖЕНИЕ А

Таблица 1 – Список банков

№	Название
1	Сбербанк России
2	ВТБ
3	Газпромбанк
4	ВТБ 24
5	Россельхозбанк
6	Банк Москвы
7	Альфа-Банк
8	НОМОС-Банк
9	ЮниКредит Банк
10	Росбанк
11	Промсвязьбанк
12	Райффайзенбанк
13	Национальный Клиринговый Центр
14	Банк «Санкт-Петербург»
15	Московский Кредитный Банк
16	Уралсиб
17	Ханты-Мансийский Банк
18	Хоум Кредит Банк
19	Ак Барс
20	Ситибанк
21	МДМ Банк
22	Связь-Банк
23	Нордеа Банк
24	Петрокоммерц
25	ИНГ Банк
26	Глобэкс
27	Зенит
28	Восточный Экспресс Банк
29	Национальный Банк «Траст»
30	Возрождение
31	Бинбанк
32	Уральский Банк Реконструкции и Развития
33	Московский индустриальный Банк
34	Новикомбанк

Продолжение Таблицы 1

35	СМП Банк
36	ОТП Банк
37	Внешпромбанк
38	Кредит Европа Банк
39	Совкомбанк
40	Транскапиталбанк
41	Дойче Банк
42	СКБ-Банк
43	МСП Банк
44	РосЕвроБанк
45	Татфондбанк
46	Ренессанс Кредит
47	Инвестторгбанк
48	Абсолют Банк
49	Росгосстрах Банк
50	Российский Капитал

Таблица 2 – Список показателей

№	Наименование
1	Активы нетто
2	Активы нетто/ кредиты предприятиям и организациям
3	Активы нетто/ кредиты физическим лицам
4	Активы нетто/ высоколиквидные активы
5	Активы нетто/ основные средства и нематериальне активы
6	Чистая прибыль

Таблица 3 – Статистика

Название банка	Активы нетто	Чистая прибыль	Капитал	Кредитный портфель	Просроченная задолженность в кредитном портфеле	Вклады физических лиц
Сбербанк России	17308935043	68210568	207300755 8	11737704871	288996934	7707257973
ВТБ	5818555489	6743528	629026832	2426467498	100455070	19051097
Газпромбанк	3905878925	1333996	434941652	2460386750	15743798	376225974
ВТБ 24	2202101644	1332721	221016700	1391729614	81150387	1384207567
Россельхозбанк	1864238034	640728	246391671	1240127210	113886997	252759716
Банк Москвы	1755314592	78075	182262411	936104947	227165744	233194649
Альфа-Банк	1514216523	9499741	210901396	1079913411	40088686	375397085
НОМОС-Банк	985682014	1540376	113841022	498708544	16764510	108584227
ЮниКредит Банк	883166179	3940040	133428679	540749399	18818453	75388402
Росбанк	785118189	3594433	88664514	443919948	36205992	153961161
Промсвязьбанк	774630701	145460	92816513	517659128	20468705	203406017
Райффайзенбанк	764394149	5158640	99237344	461424660	12629392	269905126
Национальный Клиринговый Центр	497058806	1729988	30563254	63627518	4118	11784
Банк «Санкт-Петербург»	473812878	423920	50333955	276358381	10184181	120197902
Московский Кредитный Банк	451526611	2934365	60898556	325656157	3543218	136821267
Уралсиб	385419198	42476	50068023	246286911	20355244	151405271
Ханты-Мансийский Банк	379365129	-433459	40819513	204902562	8924600	67709181

Продолжение Таблицы 3

Хоум Кредит Банк	375325073	-420699	59913011	306410998	49938337	193425794
Ак Барс	372947048	34824	49538191	227161416	9466015	71980986
Ситибанк	364731007	1961890	53392990	132299070	140222	75267826
МДМ Банк	339788707	-640374	33007771	161580316	23333663	106654184
Связь-Банк	318427179	173295	39458008	183357426	5104957	29377978
Нордеа Банк	290434975	538728	32559884	213987437	1238367	9667133
Петрокоммерц	270724998	53538	25216896	137087012	14601564	76497159
ИНГ Банк	265524585	-752338	28637815	31752868	170	1670252
Глобэкс	264076968	-272775	36052136	160076826	4190247	50357343
Зенит	255379901	10265	33784449	164130850	4021173	48017823
Восточный Экспресс Банк	254776529	-1888617	33591671	199242293	8069027	131999793
Национальный Банк «Траст»	226918248	365342	23882035	153364965	9953330	111555312
Возрождение	225221699	144099	24065623	166242418	12343091	104927716
Бинбанк	221584703	-594071	25174959	122617649	4039698	107647517
Уральский Банк Реконструкции и Развития	215337726	-293991	16599957	110570683	3253031	86582425
Московский Индустриальный Банк	194888068	181201	22648716	144809826	1638988	103044694
Новикомбанк	192634771	382122	20389188	95605285	3193151	21418867
СМП Банк	177448885	1600921	17255451	68178039	454004	80053484
ОТП Банк	171900448	-271841	25825244	128645217	16530112	54734635
Внешпромбанк	170459990	346840	21112217	95433297	632748	33495075
Кредит Европа Банк	155959338	233861	22349451	140679152	3635232	10345386
Совкомбанк	147243072	3640053	13017953	65834254	2642643	78812563
Транскапиталбанк	145878018	240654	19475058	103378960	5132190	34993596
Дойче Банк	140694093	1656105	16364269	6114016	0	2001561
СКБ-Банк	131839920	122801	15221451	93695042	4757869	70491399
МСП Банк	131804472	22147	30961441	15827893	1156264	0
РосЕвроБанк	130759215	634646	19431951	68587909	2031645	27219375
Татфондбанк	129473219	-133	15313601	85193883	2788918	51022497
Ренессанс Кредит	125674797	-1179676	13306182	88024338	10086619	61816273
Инвестторгбанк	125339404	251826	16659733	77324967	3324777	46077878

Продолжение Таблицы 3

Абсолют Банк	125086459	-23959	17018196	74130752	2209902	25135408
Росгосстрах Банк	120338548	755338	19536919	54606212	4867520	39554987
Российский Капитал	117605257	-267177	11748682	58651827	6986273	60647616