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
«___» _____2017г.

БАКАЛАВРСКАЯ РАБОТА

Направление 02.03.01 Математика и компьютерные науки

РАЗРАБОТКА ГЕНЕТИЧЕСКОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ

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

Выпускник _____/ А.В. Брестер

Красноярск 2017

ВВЕДЕНИЕ

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

Задача кластеризации данных [14] является одной из наиболее актуальных и сложных задач анализа данных. Кластерный анализ [15] представляет собой раздел статистического анализа, объединяющий методы разбиения совокупности объектов на однородные группы, называемые кластерами. Его методы имеют широкий спектр применений практически во всех областях человеческой деятельности, связанных с изучением объектов и процессов: медицине, биологии, психологии, социологии, менеджменте, маркетинге, банковском деле, актуарной математике и других.

В первой главе работы приводятся основные понятия и постановка задачи кластеризации данных. Перечисляются виды наиболее часто выделяемых кластерных структур и критерии качества кластеризации. Во второй главе проводится обзор существующих методов кластеризации данных. Особое внимание в работе уделяется наиболее популярному методу кластеризации данных: методу k-средних [6]. Дается подробное описание данного алгоритма, и указываются особенности его работы.

В третьей главе 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 общий вид схемы этапов работы генетического алгоритма.

В работе предлагается генетический алгоритм, позволяющий решать задачу кластеризации многомерных данных с заранее заданным числом кластеров. Приводится подробное описание вида операторов инициализации,

скрещивания, мутации и селекции, а также последовательность этапов работы предложенного генетического алгоритма кластеризации.

Было разработано программное приложение, реализующее работу предложенного генетического алгоритма кластеризации и классического алгоритма k-средних. Выполняется серия численных экспериментов, позволяющих оценить работу разработанного алгоритма на ряде тестовых примеров.

В работе проводится сравнение предложенного генетического алгоритма и классического алгоритма k-средних по их вычислительной сложности.

Проводится исследование влияния параметров операторов скрещивания и мутации на скорость работы генетического алгоритма.

В работе решается практический пример кластеризации данных– кластеризация 57 муниципальных районов Красноярского края по четырём экономическим показателям. Данная задача решается с помощью разработанных генетических алгоритмов и классического алгоритма k-средних. Исследование основывается на реальных статистических данных. Проводится сравнение результатов, полученных обоими методами.

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 близки относительно выбранной метрики. Объекты из разных кластеров должны существенно отличаться. Данная задача называется ***задачей кластеризации данных***. Также ее принято называть таксономией, группировкой объектов или задачей обучения без учителя.

Как уже было сказано во введении, кластерный анализ представляет собой раздел статистического анализа данных, объединяющий методы разбиения (группировки) множества объектов на кластеры.

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.

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.

Большая часть литературы по кластерному анализу появилась в течение последних трех десятилетий, хотя первые работы, в которых упоминались кластерные методы, появились достаточно давно. Польский антрополог К.Чекановский выдвинул идею «структурной классификации», содержавшую основную идею кластерного анализа – выделение компактных групп объектов.

В 1925 г. советский гидробиолог П.В. Терентьев разработал так называемый «метод корреляционных плеяд», предназначенный для группировки коррелирующих признаков. Этот метод дал толчок развитию методов группировки с помощью графов.

Впервые термин «кластерный анализ» был введен Робертом Трайоном в 1939 году [14]. Слово «cluster» переводится с английского языка как «гроздь, кисть, пучок, группа». По этой причине первоначальное время этот вид анализа называли «гроздевым анализом».

В начале 1950–х годов появились публикации Р. Люиса, Е.Фикса и Дж. Ходжеса по иерархическим алгоритмам кластерного анализа. Заметный толчок развитие работ по кластерному анализу дали работы Р. Розенблатта по

распознающему устройству (персептрону), положившие начало развитию теории «распознавания образов без учителя».

Толчком к разработке методов кластеризации явилась книга «Принципы численной таксономии»[5], опубликованная в 1963г. двумя биологами – Робертом Сокэлом и Питером Снитом.

В эти же годы было предложено множество алгоритмов таких авторов, как Дж. Мак–Кин, Г. Болл и Д. Холл по методам k–средних; Г. Ланса и У. Уильямса, Н. Джардайна по иерархическим методам.

Заметный вклад в развитие методов кластерного анализа внесли и отечественные ученые Э. М. Браверман, А. А. Дорофеюк, И. Б. Мучник, Л. А.Растригин, Ю. И. Журавлев, И. И. Елисеева и другие. В частности, в 1960–1970 гг. большой популярностью пользовались многочисленные алгоритмы, разработанные новосибирскими математиками Н. Г. Загоруйко, В. Н. Елкиной и Г. С. Лбовым.

Кластерный анализ имеет ряд интересных исследовательских результатов и может применяться при решении ряда важных практико-ориентированных задач: извлечении и поиске информации; задачах принятия решений; задачах управления; визуализации многомерных данных; абстракции данных; классификации данных и других.

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

1.2 Постановка задачи кластеризации

Пусть $X = \{x_1, x_2, \dots, x_m\}$ – множество объектов, заданных значениями впространстве признаков $P = \{p_1, p_2, \dots, p_m\}$ – (т.е. каждый объект $x_i = \left(x_i^1, x_i^2, \dots, x_i^m\right)$) и задана функция расстояния (метрика) между объектами $\rho\left(x_i, x_j\right),\ x_i, x_j \epsilon X$.

Определение 1.2 *Функцией кластеризации* называется функция$f: X \to Y$, которая любому объекту $x \epsilon X$ ставит в однозначноесоответствие номер $y \epsilon Y = \{1, \dots, k\}, \ k \leq m.$

Определение 1.3. *Множество кластеров* $C = \{C_1, C_2, \dots, C_k\}$, $k \leq m$, представляет собой разбиение множества объектов X такое, что кластер$C_i = \{x \epsilon X, f(x) = i\}$, $C_i \cap C_j = \emptyset$. Причем для C справедливо следующее: если $x_i, x_j \in C_i$, то $\rho(x_i, x_j) \to \min$. Если $x_i \in C_i, x_j \in C_j$, то $\rho(x_i, x_j) \to \max$.

Тогда постановку задачи кластеризации данных можно сформулировать следующим образом:

Требуется найти такую функцию кластеризации f^*, чтобы

$$Q(f^*, C, \rho) = \min_f Q(f, C, \rho), \qquad (1.1)$$

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: $P_i = \mathfrak{R} -$множество действительных чисел.

Самой распространенной ситуацией является кластеризация объектов, у которых все признаки являются количественными, т.е. когда $P = \{P_1, P_2, \dots, P_n\} = \mathfrak{R}^n$ (каждый объект $x_i = (x_i^1, x_i^2, \dots, x_i^1) \epsilon \mathfrak{R}^n$).

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 $C_i \epsilon C$;

3. Заданы ограничения на пространственные характеристики кластеров $C_i \epsilon C$;

4. Нет информации о количестве и характеристиках кластеров $C_i \epsilon C$.

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 типа.

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 декартовыхкоординатах, если $x = (x_1, x_2, \dots, x_n)$ и $y = (y_1, y_2, \dots, y_n)$ − две точки в евклидовом пространстве, длина отрезка $\overline{xy}$ равна:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n} \left(x_p - y_p\right)^2} \tag{1.2}$$

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

$$\rho(x, y) = \sum_{p=1}^{n} \left(x_p - y_p\right)^2 \tag{1.3}$$

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:

$$\rho(x, y) = \sum_{p=1}^{n} \left|x_p - y_p\right| \tag{1.4}$$

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:

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

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:

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

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 $C = \{C_1, C_2, \dots, C_k\}$, 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 натри 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:

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

Среднее межкластерное расстояние должно быть как можно больше:

$$Q_1 = \sum_{i<j} \sum_{x \epsilon C_i, y \epsilon C_j} \rho(x,y) \to max \tag{1.8}$$

Отношение пары функционалов:

$$\frac{Q_0}{Q_1} \to min.$$

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

Сумма средних внутрикластерных расстояний должна быть как можно меньше:

$$\Phi_0 = \sum_{y \epsilon X} \frac{1}{|K_y|} \sum_{y_i = y} \rho^2(x_i, \mu_i) \to min \quad (1.9)$$

где$K_y = \{x_i \epsilon X^l | \, y_i = y\} -$ 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 ρ 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_yточек одинаковой массы.

Сумма межкластерных расстояний должна быть как можно больше:

$$\Phi_1 = \sum_{y_i = y} \rho^2(x_i, \mu_i) \to max \quad (1.10)$$

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 все они могут давать разные результаты;
- число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием;
- 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.

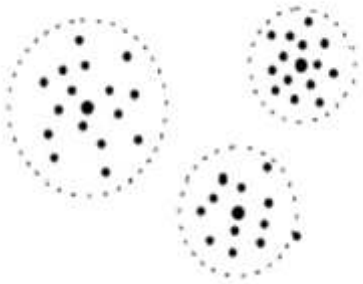


Рисунок 1.1 – Множество центроидов

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.

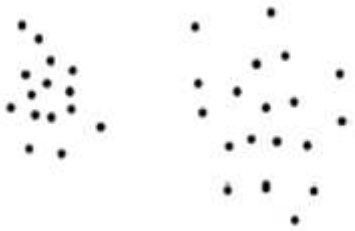


Рисунок 1.2 – Разбиения

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.

2 Классические методы решения задачи кластеризации данных

2.1 Обзор методов кластеризации

На данный момент существует свыше 100 разных алгоритмов кластеризации. Все 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 после-

дующих 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.

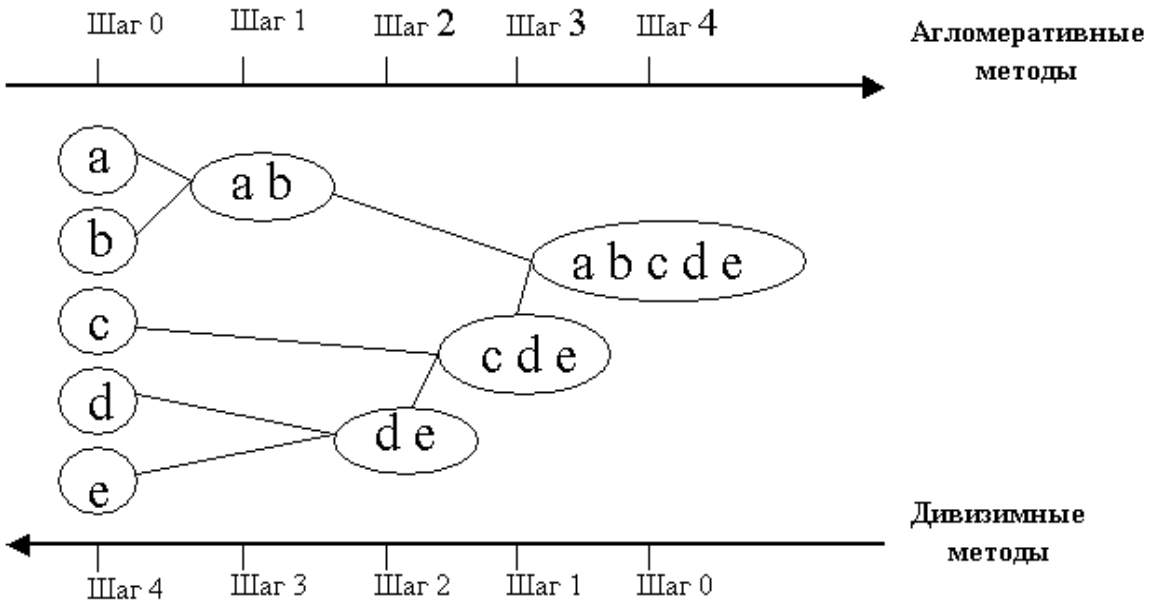


Рисунок1.6 –Принцип работы агломеративных и дивизимных методов

Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос — как определить расстояния между кластерами? Суще-

ствуют 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 действи-

тельно 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.

2.2 Метод k-средних

В данной работе подробно рассмотрим один из самых 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 $\mu_j = (\mu_j^1, \mu_j^2, \dots, \mu_j^n)$, представляющая собой центр масс точек кластера, т.е. покоординатное среднее точек из кластеров:

$$\mu_j = \sum_{x_j \epsilon C} x_j^i, j = \overline{1,n}, i = \overline{1,n}$$

Приведем 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 $S_j = \{\emptyset\}, j = 1, \dots, k.$

1. Производится распределение объектов по кластерам. Точка x_i, $i = 1, \dots, n$ относится к ближайшему кластеру, т.е. $x_i \epsilon S_i$, где $\rho(x_i, \mu_{j*}) = \min_{j=1,\dots,k} \rho(x_i, \mu_j)$

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

В результате каждый объект назначен определенному кластеру.

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

2. Продолжается 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=50*, 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.

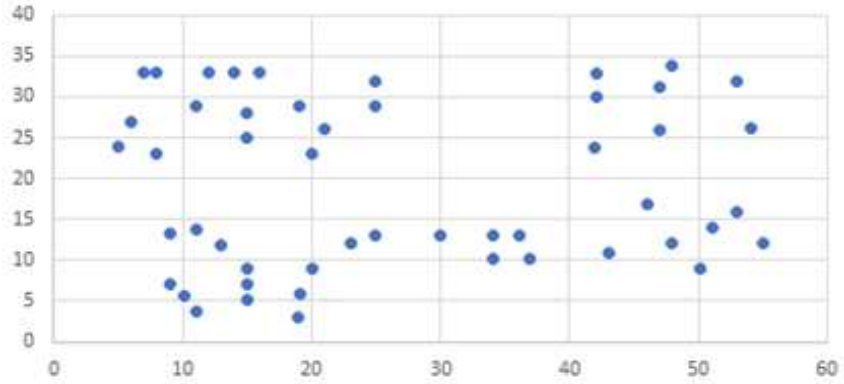


Рисунок 2.1 – Обучающая выборка (n=50)

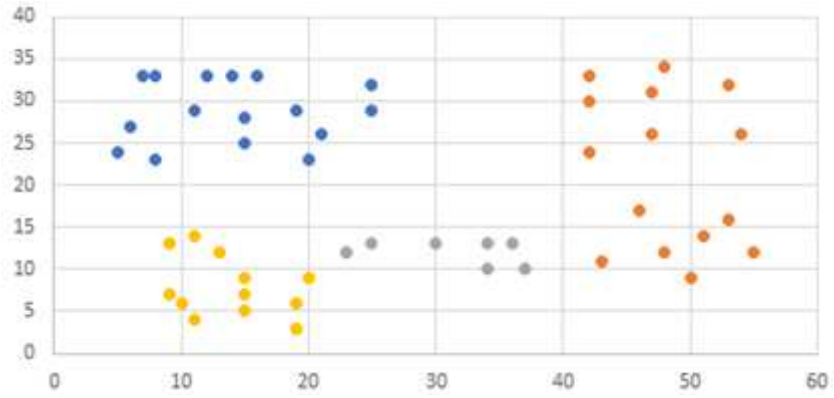


Рисунок 2.2 – Результат работы метода k-средних (n=50)

После получения 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;

3 Генетические алгоритмы

3.1 Основные понятия генетических алгоритмов

Генетические алгоритмы – эвристические алгоритмы поиска, используемые для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации параметров с использованием механизмов, аналогичных естественному отбору в природе.

Генетические алгоритмы (ГА) 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 *Популяция* – совокупность индивидуумов.

Определение3.6 *Кроссинговер* (скрещивание) — операция, при которой две хромосомы обмениваются своими частями. Например, 11|00&10|10 —> 1110&1000.

Определение 3.7 *Мутация* — случайное изменение одной или нескольких позиций в хромосоме. Например, 1010011 —> 1010001.

Определение 3.8 *Функция приспособленности* (*fitnessfunction*) - мера приспособленности данной особи в популяции.

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

Последовательность основных этапов работы классического генетического алгоритма выглядит следующим образом:

1) Генерируется или выбирается начальная популяция хромосом;
2) Вычисляется и оценивается приспособленность хромосом в популяции;
3) Проверка условия остановки алгоритма;
4) Селекция хромосом-родителей;
5) Применение генетических операторов;
6) Формирование новой популяции
7) Повторяются шаги 2-6, пока не будет достигнут критерий окончания процесса;

Схема работы классического генетического алгоритма представлена на рисунке 3.1.

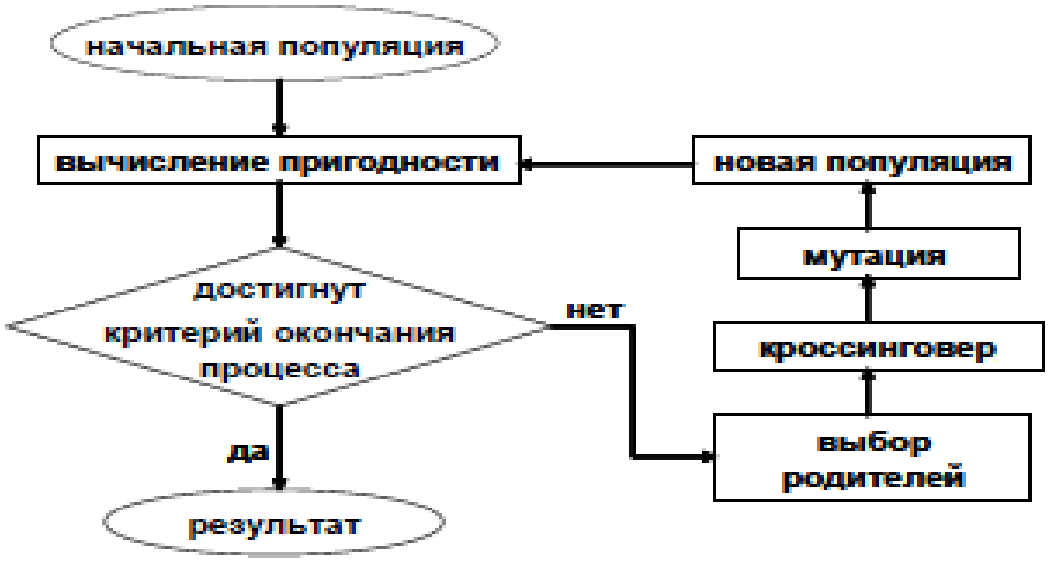


Рисунок 3.1– Схема работы генетического алгоритма

Рассмотрим каждый из этапов алгоритма более подробно:

1) ***Генерация популяции***: создание случайным образом заданного количества хромосом длины N, в виде требующимся для решаемой задачи.
2) ***Приспособленность хромосом в популяции*:** расчет функции приспособленности для каждой хромосомы популяции, чем больше значение этой функции, тем более «жизнеспособна» хромосома. В зависимости от характер выбранной задачи, будет меняться и форма функции приспособленности.
3) ***Проверка условий остановки алгоритма*:** критерием окончания процесса может служить схождение популяции или заданное количество итераций алгоритма

4) ***Селекция хромосом-родителей*:** по рассчитанным значениям функции приспособленности выполняется отбор тех хромосом, которые будут участвовать в создании потомков для следующей популяции.

Перечислим наиболее распространённые типы селекции:

4.1) Панмиксия.

Самый простой оператор отбора. Каждому члену популяции ставится в соответствие случайное целое число из отрезка [1; n], где n-кол-во особей в популяции. Будем рассматривать эти числа как номера особей, которые примут участие в скрещивании. Таким образом, с одной стороны, некоторые члены популяции не будут принимать участия в размножении, так как образуют пару с самим собой, с другой стороны, некоторые особи примут участие в процессе размножения неоднократно, с различными особями. Такой подход универсален для решения различных классов задач, но критичен к численности популяции.

4.2) Метод рулетки.

Особи отбираются с помощью N «запусков» рулетки, где N- размер популяции. Колесо рулетки содержит по одному сектору для каждой хромосомы популяции. Размер i-го сектора пропорционален вероятности попадания в новую популяцию $P(i)$, вычисляемой по формуле:

$$P(i) = \frac{f(i)}{\sum_{i=1}^{N} f(i)}$$

где$f(i)$- приспособленность i-й особи. Ожидаемое число копий i-й хромосомы после оператора рулетки определяются по формуле $N_i = P(i)N$. При таком отборе члены популяции с более высокой приспособленностью будут чаще выбираться, чем особи с низкой приспособленностью

Популяция из 5 особей	Пригодность	Вероятность выбора
C_1	52	$52/200 = 0{,}26$
C_2	85	$85/200 = 0{,}425$
C_3	37	$37/200 = 0{,}185$
C_4	3	$3/200 = 0{,}015$
C_5	23	$23/200 = 0{,}115$

Рисунок 3.2 –Метод рулетки. Суммарная пригодность = 200. Суммарная вероятность = 1.

4.3) Турнирный отбор.

При турнирном отборе из популяции, состоящей из N особей, случайным образом выбираются k (k≥2) особей, и лучшая из них записывается в промежуточный массив, такая операция повторяется N раз. Особи в полученном массиве зачем подвергаются скрещиванию (случайным образом). K называют численностью турнира. Преимущество способа в том, что он не требует никаких дополнительных вычислений.

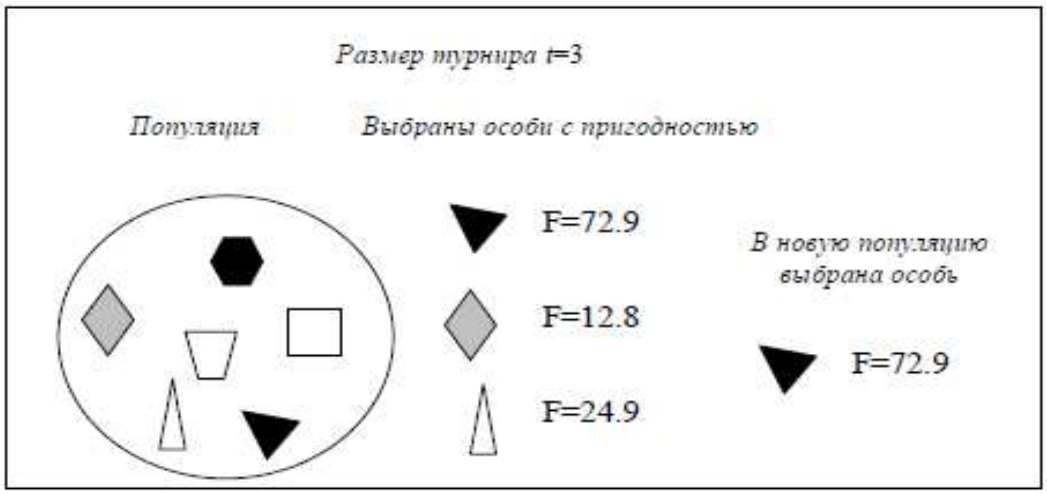


Рисунок 3.3 –Пример турнирного отбора. T=3.

4.4) Инбридинг и аутбридинг

При инбридинге первый родитель выбирается случайным образом, а вторым родителем является член популяции, ближайший к пер-

вому. «Ближайший» может пониматься в смысле минимального расстояния Хемминга (для бинарных строк) или евклидова расстояния между двумя вещественными числами.

При аутбридинге первый родитель, также, выбирается случайным образом, однако, пары для скрещивания формируют из максимально далеких особей

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

5) ***Применение генетических операторов*** к хромосомам-родителям влечет за собой формирование новой популяции хромосом-потомков от отобранных на предыдущем шаге хромосом-родителей

В классическом генетическом алгоритме применяются два основных генетических оператора: *оператор скрещивания (crossover)* и *оператор мутации (mutation)*. Следует отметить, 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 животных организмов, где мутации происходят чрезвычайно редко.

5.1) Оператор скрещивания:

При скрещивании на основе преобразования (скрещивания) хромосом родителей (или их частей) создавать хромосомы потомков.

Наиболее распространены следующие виды скрещивания:

5.1.а) Одноточечный оператор скрещивания

Перед началом работы оператора определяется разрезающая точка оператора, которая обычно определяется случайно. Эта точка определяет место в двух хромосомах, где они должны быть разрезаны, после чего хромосомы обмениваются частями, стоящими правее точки разреза.

5.1.б) Двухточечный оператор скрещивания

Хромосомы рассматриваются как кольца, затем выбираются две точки разрыва и родители обмениваются частями

5.1.в) Оператор скрещивания с уменьшением замены

Оператор уменьшения замены ограничивает скрещивание, чтобы всегда, когда возможно, создавать новые особи. Это осуществляется за счет ограничения на выбор точки разреза: точки разреза должны появляться только там, где гены различаются.

5.2) Оператор мутации

Данный оператор необходим для «выбивания» популяции из локального экстремума и препятствует преждевременной сходимости. Мутация обычно происходит с вероятностью p_m для каждого гена.

Хорошим эмпирическим правилом считается выбор вероятности мутацииравным $p_m = \frac{1}{n}$, где n - число генов в хромосоме (в среднем хотя бы одинген будет подвержен мутации).

6) ***Формирование новой популяции.*** Хромосомы, полученные в результате воздействия генетических операторов на хромосомы родителей, становятся текущей популяцией на данной итерации генетического алгоритма. На каждой итерации рассчитываются значения функции приспособленности для каждой хромосомы популяции.
7) ***Проверка условия остановки.*** Если достигнуто условие остановки алгоритма, то из имеющейся популяции выбирается хромосома с самым лучшим (самым большим) значением функции приспособленности, она и будет являться искомым решением задачи.

Страница изъята

Страница изъята

3.4 Сравнение предложенного генетического алгоритма с методом k-средних

В бакалаврской работе был проведен сравнительный анализ стандартного алгоритма кластеризации данных k-средних с разработанным генетическим алгоритмом кластеризации. 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)*, где *k* – число кластеров, *l* – число итераций.
Генетический алгоритм	Произвольная	Число кластеров	$O(n^2)$.

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

4 Исследование влияния параметров генетического алгоритма на скорость его работы

В работе был произведен анализ зависимости времени работы алгоритма от выбора параметров его операторов, а именно: оператора скрещивания и оператора мутации. В качестве параметра для оператора скрещивания бралось количество точек, которыми разделяется хромосома (одна, две), для оператора мутации параметром являлась вероятность мутации (p=0.05,p=0,1)

Все сравнения проводились на выборке n=20 и количестве кластеров k=3. В таблице приведено среднее время работы за 20 тестовых запусков.

Таблица 2 − Сравнение времени работы генетического алгоритма с различными параметрами операторов

Скрещивание	Вероятность мутации	Время работы
Одноточечное	P=0.05	1.997с
	P=0.1	2.051с
Двухточечное	P=0.05	1.935с
	P=0.1	1.842с

На рис 4.1 приведен сводный график сравнения изменения значений функции пригодности для каждого вида параметров. Как видно из графика, с ростом количества итераций значения функции пригодности для каждого вида параметров уменьшается. И, как можно увидеть на графике, с наименьшим значением свою работу закончил алгоритм с двухточечным скрещиванием и вероятностью мутации p=0.1.

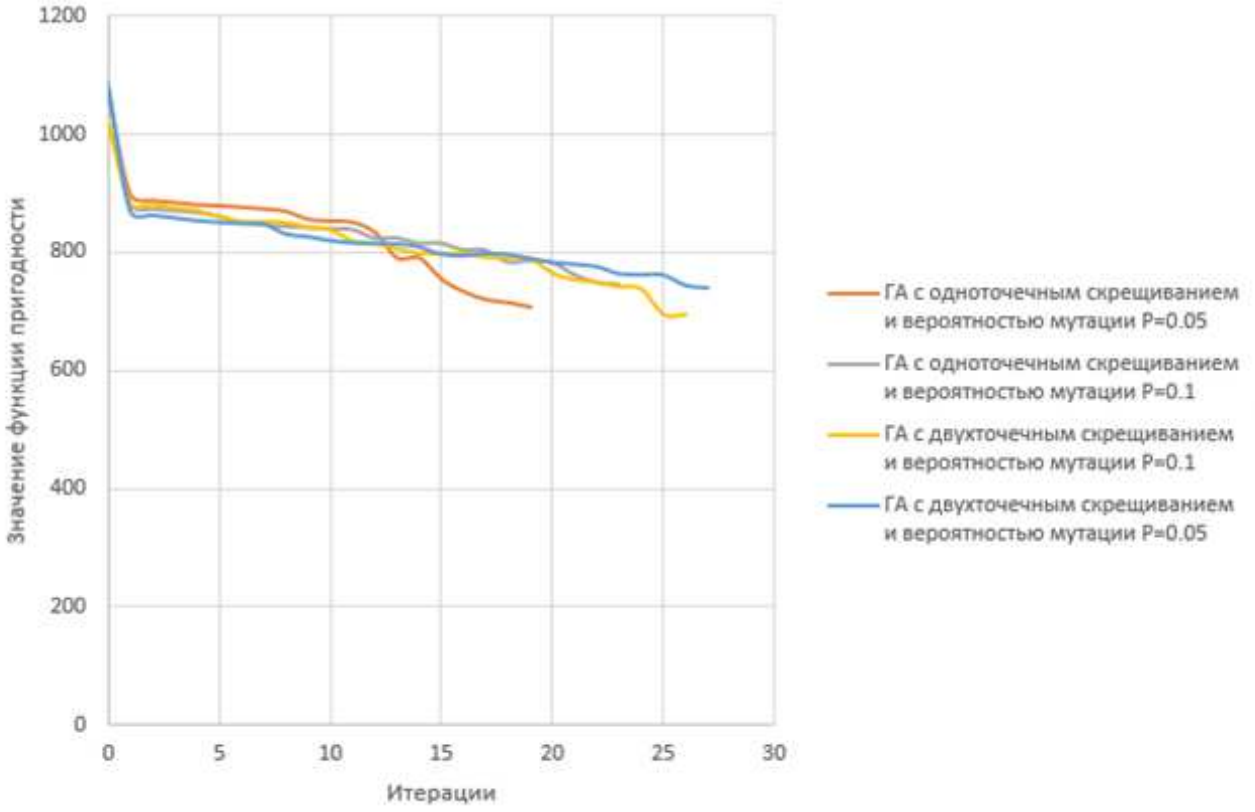
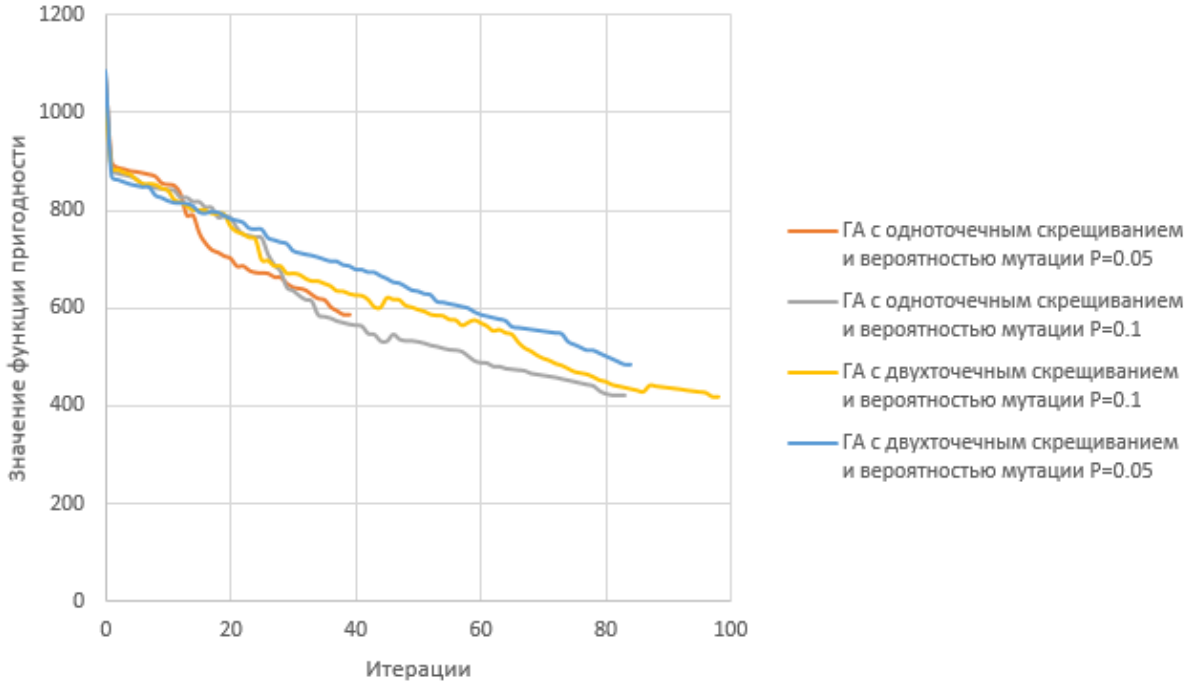


Рис 4.1 – График сравнения скорости уменьшения значения функции пригодности для каждого вида параметров

На рисунке 4.2 представлен график сравнения количества итераций, за которые алгоритм находит решение, и с каким значением функции пригодности он заканчивает работу. Как видно на графике, наиболее точное решение показал алгоритм с двухточечным скрещиванием и вероятностью мутации p=0.1.

Рис 4.2– График сходимости алгоритмов с различными параметрами



5 Решение практической задачи кластеризации

В работе решается практический пример кластеризации данных – кластеризация 57 муниципальных районов Красноярского края по четырём экономическим показателям. Данная задача решается с помощью разработанных генетических алгоритмов и классического алгоритма k-средних. Исследование основывается на реальных статистических данных.

5.1 Описание статистики

Для решения задачи использовалась официальная статистика основных показателей социально-экономического положения муниципальных образований Красноярского края за 2016 год, собранная Управлением Федеральной службы государственной статистики по Красноярскому краю, республике Хакасия и республике Тыва. Полный список всех муниципальных районов приведен в Приложении А. Кластеризация районов Красноярского края выполнялась по следующим четырём экономическим показателям: общая площадь жилых помещений, приходящаяся в среднем на одного жителя – всего; метр квадратный, доходы местного бюджета, фактически исполненные; профицит, дефицит (-) бюджета муниципального образования; среднемесячная заработная плата работников.

Кластеризация проводилась с помощью генетического алгоритма и метода k-средних.

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 схеме работы, приведен-

ной 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

CENTROID :0	Абанский, Ачинский, Балахтинский, Березовский, Богучанский, Емельяновский, Енисейский, Ермаковский, Иланский,Ирбейский, Канский, Каратузский, Кежемский, Краснотуранский, Курагинский, Минусинский, Мотыгинский, Назаровский, Нижнеингашский, Новоселовский, Партизанский, Пировский, Рыбинский, Северо-Енисейский, Сухобузимский, Туруханский, Тюхтетский, Ужурский, Уярский, Шушенский, Эвенкийский, Красноярск, Ачинск, Дивногорск, Енисейск, Канск, Лесосибирск, Минусинск, Назарово, Норильск, Сосновоборск, Шарыпово, Кедровый
CENTROID :1	Боготольский, Большеулуйский, Дзержинский, Казачинский, Тасеевский
CENTROID :2	Бирилюсский, Большемуртинский, Идринский, Козульский, Саянский, Боготол, Бородино

Затем задача кластеризации на три кластера была решена с помощью генетического алгоритма с детерминированным числом кластеров, предложенного в главе 3 данной работы. Результаты, полученные данным алгоритмом, представлены в таблице 3.

Далее этими же двумя алгоритмами на основании той же статистики была решена задача кластеризации районов края на четыре кластера.

Результаты, полученные методом k–средних, показаны в таблице 4. В таблице 5 демонстрируются результаты кластеризации с помощью генетического алгоритма с детерминированным числом кластеров.

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

CENTROID :0	Абанский, Ачинский, Богучанский, Большемуртинский, Дзержинский, Енисейский, Иланский, Ирбейский, Казачинский, Канский, Кежемский, Краснотуранский,Курагинский, Нижнеингашский, Рыбинский, Северо-Енисейский, Сухобузимский, Тасеевский, Шушенский, Лесосибирск, Назарово, Норильск, Кедровый
CENTROID :1	Балахтинский, Березовский, Боготольский, Большеулуйский, Емельяновский, Ермаковский, Идринский, Каратузский, Козульский, Манский, Минусинский, Мотыгинский, Назаровский, Новоселовский, Партизанский, Пировский, Саянский, Туруханский, Тюхтетский, Ужурский, Уярский, Шарыповский, Эвенкийский, Ачинск, Боготол, Бородино, Енисейск, Канск, Минусинск, Шарыпово
CENTROID :2	Бирилюсский, Красноярск, Дивногорск, Сосновоборск

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

CENTROID:0	Абанский, Ачинский, Балахтинский, Березовский, Богучанский, Емельяновский, Енисейский, Ермаковский, Иланский, Ирбейский, Канский, Каратузский, Кежемский, Краснотуранский, Курагинский, Минусинский, Мотыгинский, Назаровский, Нижнеингашский, Новоселовский, Партизанский, Пировский, Рыбинский, Северо-Енисейский, Сухобузимский, Туруханский, Тюхтетский, Ужурский, Уяр-

Продолжение таблицы 4

	ский, Шушенский, Эвенкийский, Красноярск, Ачинск, Дивногорск, Енисейск, Канск, Лесосибирск, Минусинск, Назарово, Норильск, Сосновоборск, Шарыпово, Кедровый
CENTROID:1	Боготольский, Большеулуйский, Дзержинский, Казачинский, Тасеевский
CENTROID:2	Бирилюсский, Большемуртинский, Идринский, Козульский, Саянский, Боготол, Бородино
CENTROID:3	Манский, Шарыповский

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

CENTROID:0	Бирилюсский, Большеулуйский, Ирбейский, Курагинский, Рыбинский, Красноярск, Минусинск, Норильск
CENTROID:1	Абанский, Березовский, Боготольский,Богучанский, Большемуртинский, Идринский, Каратузский, Кежемский, Краснотуранский, Минусинский, Мотыгинский, Нижнеингашский, Новоселовский, Партизанский, Ужурский, Уярский, Шушенский, Эвенкийский, Ачинск, Бородино, Дивногорск, Енисейск, Канск, Назарово, Кедровый
CENTROID:2	Ачинский, Дзержинский, Емельяновский, Иланский, Казачинский, Канский, Козульский, Манский, Назаровский, Пировский, Саянский, Северо-Енисейский, Тасеевский, Тюхтетский, Шарыповский, Сосновоборск
CENTROID:3	Балахтинский, Енисейский, Ермаковский, Сухобузимский, Туруханский, Боготол, Лесосибирск, Шарыпово

5.3 Сравнение полученных результатов

На основе реальной статистики была проведена кластеризация 57 муниципальных районов Красноярского края по четырём экономическим пока-

зателям. 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 6.

Алгоритм кластеризации	**Число кластеров**	**Результат статистических наблюдений**
Метод k-средних	3	0,093
Генетический алгоритм с детерминированным числом кластеров	3	0,091
Метод k-средних	4	0,1
Генетический алгоритм с детерминированным числом кластеров	4	0,094

Таблица 6 – Значение функционалов качества кластеризации

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. Для генетического алгоритма с числом кластеров k=3 в 1 кластер попали районы с самой большой площадью, с самым высоким бюджетом, со средним дефицитом бюджета и средней зара-

ботной платой, во 2 кластер попали районы со средней площадью, высоким бюджетом, самым большим дефицитом бюджета и самой высокой заработной платой, в 3 кластер попали районы с наименьшей площадью, наименьшим бюджетом, средним дефицитом бюджета и наименьшей заработной платой.

Из полученных результатов можно сделать следующее заключение: решение задачи генетическим алгоритмом с детерминированным числом кластеров дает наилучший результат при разбиении данных на 3 кластера.

5.4 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 рисунке 5.1 представлена карта Красноярского края, с нанесенными на неё кластерами, полученными в результате кластеризации генетическим алгоритмом с количеством кластеров k=3.

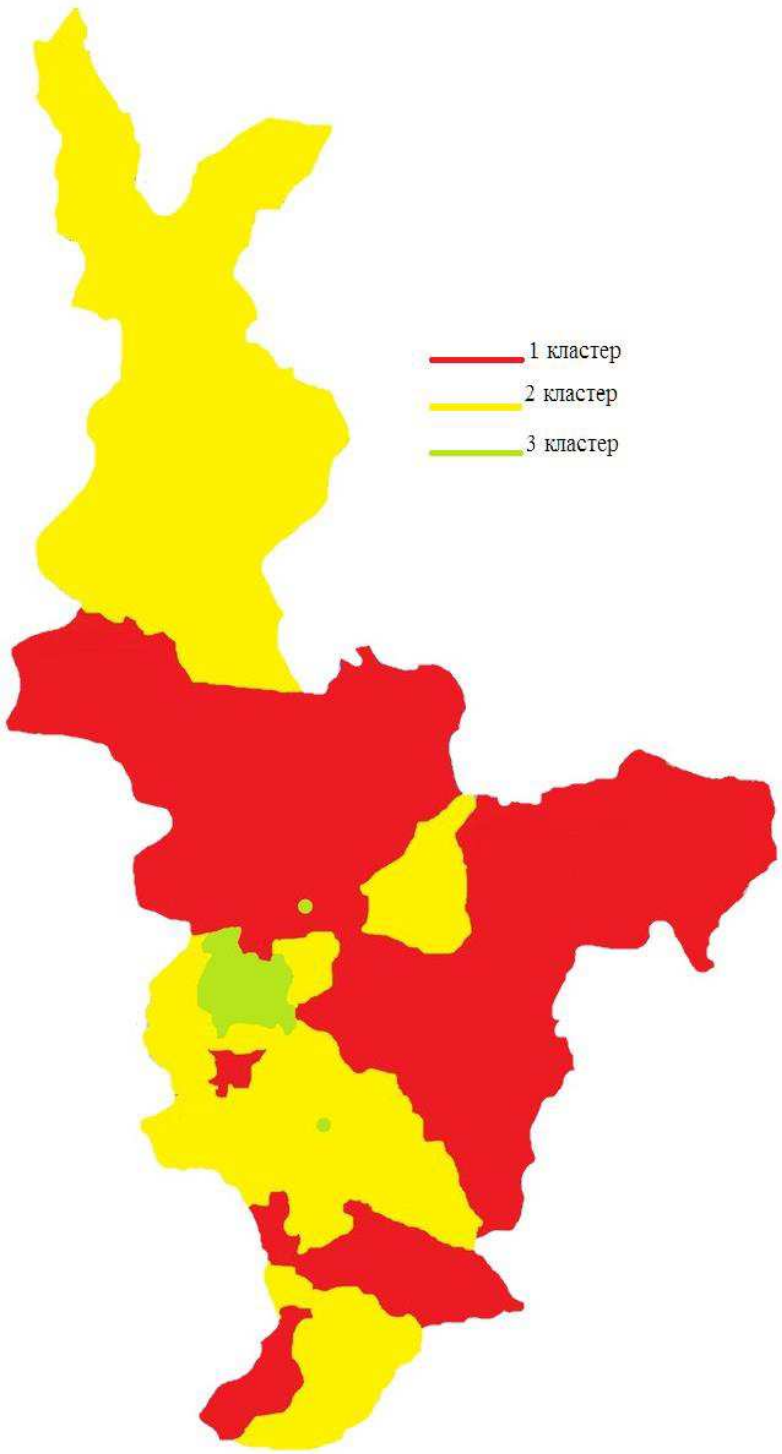


Рис 5.1 –Результат разбиения генетическим алгоритмом с детерминированным числом кластеров k=3.

ЗАКЛЮЧЕНИЕ

В работе получены следующие результаты:

- Изучена задача кластеризации данных типа и основные алгоритмы кластеризации многомерных данных.
- Разработан генетический алгоритм для решения задачи кластеризации многомерных данных с заданным количеством кластеров.
- Создано программное приложение, реализующее работу предложенного алгоритма, а также алгоритма кластеризации k-средних.
- Проведено сравнение изученного и предложенного методов по их вычислительной сложности и результатам работы.
- Проведено исследование влияния параметров генетического алгоритма на скорость его работы.
- Решена практическая задача кластеризации 57 муниципальных районов Красноярского края по четырём экономическим показателям
- Проведено сравнение результатов, полученных в результате работы каждого метода.

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 / И. З. Батыршин;подред. Н.Г. Ярушкиной. – М.: 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. –М.: 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. – М.: Финансы и статистика, 1988. – 176 с.

8. Панченко, Т. В. Генетические алгоритмы: учебно-методическое пособие / Т.В. Панченко; под ред. Ю. Ю. Тарасевича. – Астрахань: Издательский дом «Астраханский университет», 2007. – 87 с.

9. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д.Рутковская;пер. 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: Б.Г. Миркин. — М.: Изд. ДомНационального 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. 269–301с.
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. – 238 с.
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.

Приложение А.

Таблица 1 – Список городов

№	Наименование
1	Абанский
2	Ачинский
3	Балахтинский
4	Березовский
5	Бирилюсский
6	Боготольский
7	Богучанский
8	Большемуртинский
9	Большеулуйский
10	Дзержинский
11	Емельяновский
12	Енисейский
13	Ермаковский
14	Идринский
15	Иланский
16	Ирбейский
17	Казачинский
18	Канский
19	Каратузский
20	Кежемский
21	Козульский
22	Краснотуранский
23	Курагинский
24	Манский
25	Минусинский
26	Мотыгинский
27	Назаровский
28	Нижнеингашский
29	Новоселовский
30	Партизанский
31	Пировский
32	Рыбинский
33	Саянский
34	Северо-Енисейский
35	Сухобузимский
36	Тасеевский

Продолжение таблицы 1

37	Туруханский
38	Тюхтетский
39	Ужурский
40	Уярский
41	Шарыповский
42	Шушенский
43	Эвенкийский
44	Красноярск
45	Ачинск
46	Боготол
47	Бородино
48	Дивногорск
49	Енисейск
50	Канск
51	Лесосибирск
52	Минусинск
53	Назарово
54	Норильск
55	Сосновоборск
56	Шарыпово
57	Кедровый

Таблица 2 – Статистика

Город	Пло-щадь	Доходы Бюждета	Профицит, дефицит(-)	Средняя ЗП
Абанский муниципальный район	23,7	768715	-2511	25110.1
Ачинский муниципальный район	27,4	621560	-15963	28618,5
Балахтинский муниципальный район	28,9	879224	-13978	24428.2
Березовский муниципальный район	21,2	783618	11283	32616
Бирилюсский муниципальный район	23	516469	-206	26537.1
Боготольский муниципальный район	19	463525	1660	24529.9
Богучанский муниципальный район	23,6	1932182	-174931	42327.2
Большемуртинский муниципальный район	21,3	554431	-2534	26212.5
Большеулуйский муниципальный район	30	490898	7639	44139.3
Дзержинский муниципальный район	25,9	495359	698	23278.1
Емельяновский муниципальный район	30,1	1443442	-12616	41508.3
Енисейский муниципальный район	26,8	1779732	-39282	29879.5
Ермаковский муниципальный район	25,5	764886	790	24899.6

Продолжение таблицы 2

Идринский муниципальный район	26,2	538078	-2296	24190.8
Иланский муниципальный район	23,8	995762	2179	33842.2
Ирбейский муниципальный район	23,9	697692	3428	25184.1
Казачинский муниципальный район	27,3	504746	-1374	23846.2
Канский муниципальный район	20,3	858871	-8372	20582.8
Каратузский муниципальный район	23,9	699983	4488	23304.3
Кежемский муниципальный район	22,5	1183162	99144	40891.4
Козульский муниципальный район	22	521494	-7658	30235.5
Краснотуранский муниципальный район	25,1	667166	-12843	21292.6
Курагинский муниципальный район	23,2	1417102	2313	24720.6
Манский муниципальный район	27,8	588764	-23438	24376.2
Минусинский муниципальный район	22	929401	-52780	23100.4
Мотыгинский муниципальный район	25,5	1005894	91177	43547.1
Назаровский муниципальный район	19	833901	-6086	20707.9
Нижнеингашский муниципальный район	19,3	884439	-10637	25080.2
Новоселовский муниципальный район	23,1	638130	-3602	23925.5
Партизанский муниципальный район	28,6	435848	2866	27188.3
Пировский муниципальный район	26,8	411719	-1580	25123.6
Рыбинский муниципальный район	29,1	1195908	9916	31326.9
Саянский муниципальный район	27,9	538518	-741	23786.8
Северо-Енисейский муниципальный район	20,4	2096538	-356007	82776.7
Сухобузимский муниципальный район	24,4	803722	-5405	22684
Тасеевский муниципальный район	27,7	487779	1513	23792.5
Туруханский муниципальный район	28	3337126	-35649	66524.1
Тюхтетский муниципальный район	25,3	366920	-5411	24063.3
Ужурский муниципальный район	21,3	908852	-740	27182.2
Уярский муниципальный район	24,3	865400	1114	28284.7
Шарыповский муниципальный район	27,8	582873	-31150	40803.8
Шушенский муниципальный район	26,4	1316396	-30766	23951.3
Эвенкийский муниципальный район с 2012 г.	21,3	5827329	-74261	53696.6
город Красноярск	23,7	26121962	-1456094	41715.3
город Ачинск	24,1	3023427	70256	32718.6
город Боготол	25,5	546078	-15463	35395
город Бородино	26,8	523740	18490	33230.6
город Дивногорск	26,7	1098724	-43148	31998.5
город Енисейск	29,2	729801	-5456	33214.8
город Канск	22,8	2547580	-6024	27339
городЛесосибирск	24,1	2369166	161226	32536.5
город Минусинск	29,7	1715240	-9427	28099.5

Продолжение таблицы 2

город Назарово	25,1	1263010	61252	28922.5
город Норильск	24,3	16815796	-165049	82991.9
Город Сосновоборск	23,6	870717	-6170	28397.7
город Шарыпово	25,6	1002520	-269	28767.4
Поселок Кедровый	19	142467	2313	22568.6

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 2017г.

БАКАЛАВРСКАЯ РАБОТА

Направление 02.03.01 Математика и компьютерные науки

РАЗРАБОТКА ГЕНЕТИЧЕСКОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ

Научный руководитель
кандидат физико-математических наук,
доцент _____ / И.В. Баранова
16.06.17

Выпускник _____ / А.В. Брестер
16.06.17

Красноярск 2017