МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

КЛАСТЕРНЫЙ АНАЛИЗ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Нагаевой Любови Владимировны

Научный руководитель
доцент, к. ф.-м. н доцент _____ М. Г. Плешаков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2020

ВВЕДЕНИЕ

Актуальность темы. Процессы и явления, которые изучаются человеком в русле его практической хозяйственной и теоретической исследовательской деятельности, зависят, как правило, от значительного числа параметров, которые их характеризуют, что и обусловливает трудности, связанные с определением структуры взаимосвязей и зависимостей этих параметров. В таких ситуациях, т.е. когда решения принимаются на основании анализа стохастической, неполной или неточной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым. Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать такую, которая наилучшим образом соответствует исходным статистическим данным, описывающим истинное поведение изучаемой совокупности объектов, дать оценку надежности и точности выводов, сделанных на основании ограниченного статистического материала. К области приложения математической статистики может быть отнесен широкий круг задач, связанных с исследованием поведения индивидуума, единицы, как представителя большой совокупности объектов. Многомерный статистический анализ опирается на широкий спектр методов. Упомянем некоторые из наиболее используемых методов, а именно: факторный, кластерный и дискриминантный анализы. Методы многомерной классификации, которые предназначены разделять рассматриваемые совокупности объектов, субъектов или явлений на группы в определенном смысле однородные. Необходимо учитывать, что каждый из рассматриваемых объектов характеризуется большим количеством разных и стохастически связанных признаков. Для решения таких сложных задач классификации применяют кластерный и дискриминантный анализ. Наличие множества исходных признаков, характеризующих процесс функционирования объектов, которое может оказаться избыточным, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Такое решение может быть обеспечено методами снижения размерности, к которым относится

факторный анализ. Этот метод позволяет учитывать эффект существенной многомерности данных, дает возможность лаконичного и более простого объяснения многомерных структур. Вскрывает объективно существующие, непосредственно не наблюдаемые закономерности при помощи полученных факторов или главных компонент. Это дает возможность достаточно просто и точно описать наблюдаемые исходные данные, структуру и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число факторов или главных компонент – новых единиц измерения – используется значительно меньше, чем исходных признаков. Все перечисленные методы наиболее эффективны при активном применении статистических пакетов прикладных программ. При помощи этих пакетов представляется возможным также восстанавливать пропущенные данные и др.

Специализированные статистические пакеты, позволяющие применять современные методы математической статистики для обработки данных, обладают самыми широкими возможностями. По официальным данным Международного статистического института, число статистических программных продуктов приближается к тысяче. Среди них есть профессиональные статистические пакеты, предназначенные для пользователей, хорошо знакомых с методами математической статистики, и есть пакеты, с которыми могут работать специалисты, не имеющие глубокой математической подготовки; есть пакеты отечественного производства и созданные зарубежными программистами; различны такие программные продукты и по цене.

Среди программных средств данного типа можно выделить узкоспециализированные пакеты, в первую очередь статистические - STATISTICA, SPSS, STADIA, STATGRAPHICS, которые имеют большой набор статистических функций: факторный анализ, регрессионный анализ, кластерный анализ, многомерный анализ, критерии согласия и т. д.

Необходимость принятия проектных, управленческих, маркетинговых и других решений в условиях неопределенности и неполноты информации, которые вызваны неполнотой и (или) расплывчатостью данных, неточностью измерения или регистрации наблюдений, субъективной оценкой требует дополнительных средств для раскрытия такой неопределенности. Наряду с интеллектом экспертов сегодня для этой цели используются технологии под-

держки искусственного интеллекта, среди которых нейронные сети, генетические алгоритмы, нечеткие множества, кластеризация. Используя сочетания этих технологий, удается достичь большего совокупного эффекта.

В информационной поддержке принимаемых решений успешно используется технология кластерного анализа в таких областях, как машиностроение , экономика (сегментирование рынка , стратегическое планирование, анализ кредитных рисков, государственное и муниципальное управление (региональная экономическая политика, налоговое администрирование, избирательные кампании, медицина , образование и др. Еще более широкое распространение получила кластеризация в составе различных информационных технологий, например, распознавания изображений , анализа текстов, отражения кибератак . В настоящее время использование традиционных алгоритмов кластеризации, например, итеративных ("k-means", SOM и т. д.), иерархических дивизимных (BIRCH, MST и т.д.) и иерархических агломеративных (CURE, ROCK и др.) сменилось гибридным подходом, предполагающим использование при кластеризации вышеназванных технологий «мягких вычислений». Данное обстоятельство вызвано растущим количеством и сложностью задач, использующих кластеризацию, значительным увеличением объема анализируемых данных, зачастую в режиме онлайн. При этом, рассматривая масштабируемые алгоритмы для работы с большим объемом данных, следует отметить их работу с категорийными данными. В настоящее время возможности кластерного анализа расширяются за счет использования наряду с традиционными новых методов (например, метод волновых преобразований в алгоритме WaveCluster, объединения нескольких методов в один, выявления и учета при кластеризации взаимозависимости в данных, а также за счет гибридного подхода.

Для профессионального исследователя пакеты программ кластерного анализа обладают максимальной гибкостью и заметными удобствами для пользователя. Они сочетают в себе преимущества общих пакетов статистической обработки с чертами, представляющими особый интерес для пользователя кластерного анализа.

Хорошо известны такие из пакетов программ кластерного анализа, как существующая в различных редакциях программа CLUSTAN с одиннадца-

тью процедурами, которые содержат все семейства методов кластеризации; BCTRY (Tryon and Bayley), CLUS (Friedman and Rubin), NTSYS (Rohlf et al.). Последний пакет включает реализацию различных методов кластерного анализа и численной таксономии, а также ряд многомерных статистических процедур, а также многомерное шкалирование и факторный анализ.

Наконец, особое внимание стоит уделить R. R – это среда вычислений, разработанная для обработки данных, математического моделирования и работы с графикой. R можно использовать как простой калькулятор, можно редактировать в нем таблицы с данными, можно проводить простые статистические анализы (например, t-тест, ANOVA или регрессионный анализ) и более сложные длительные вычисления, проверять гипотезы, строить векторные графики и карты. При этом R - это свободный и кроссплатформенный продукт.

Кроме того, R – это язык программирования, благодаря чему можно писать собственные программы (скрипты) при помощи управляющих конструкций, а также использовать и создавать специализированные расширения (пакеты). Пакет - это набор R функций, файлов со справочной информацией и примерами, собранных вместе в одном архиве. R пакеты играют важную роль, так как они используются как дополнительные расширения на базе R. Каждый пакет, как правило, посвящен конкретной теме, например: пакет 'ggplot2' используется для построения векторных графиков определенного вида, а пакет 'qtl' идеально подходит для генетического картирования. Таких пакетов в библиотеке R насчитывается на данный момент более 7000.

Тема данной работы является актуальной, так как кластерный анализ занимает устойчивое и прочное место в наборе методов решения задач интеллектуального анализа данных, и роль его не становится меньше.

Целью исследования является демонстрация эффективности эволюционного подхода как сравнительного нового в кластерном анализе.

Объектом исследования является кластеризация как один из инструментов Data Mining.

Предметом исследования является возможность осуществить решение задачи кластеризации с помощью методов эволюционных вычислений.

Для достижения поставленной цели в работе необходимо решить сле-

дующие задачи:

— Имея исчерпывающие знание алгоритма кластеризации,основанного на эволюционном алгоритме, написать программный код интерфейс которого обеспечивал бы удобное использование программного продукта и допускал бы легко осуществимую модификацию.

— Знакомство с методами построения эволюционных алгоритмов и принципами их работы.

— Применить полученный программный продукт для исследования конкретного массива данных. Предусмотреть возможность последующего использования программного продукта в учебном процессе или исследовательской деятельности.

Практическая значимость данного исследования носит междисциплинарный характер.Методы кластеризации поставлены на службу археологическим исследованиям. Продемонстрирована эффективность этого подхода.

Стуктура и содержание магистерской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников содержащего 32 наименования, и четырех приложений. Общий объем работы составляет 74 страницы включая приложения.

Основное содержание работы

В **первом** разделе дается изложение традиционных методов кластеризации, основные понятия и алгоритмы.

Второй раздел посвящен пока еще сравнительно редко применяемому методу кластеризации, основанному на использовании эволюционных алгоритмов. Идея реализации такого подхода кажется естественной, ибо основана на способности генетических (эволюционных) алгоритмов к эффективной оптимизации любой выбранной в качестве целевой функции. И даже заметная неопределенность в процедуре определения характера неоднородности и установления структуры исследуемого массива, связанная с необходимостью априорного задания числа кластеров, которое как правило корректируется в ходе экспериментов, и присущая большинству известных методов кластеризации данных, вполне согласуется с духом эволюционных вычислений.

На рисунке 1 приведена принципиальная блок-схема простейшего генетического алгоритма.

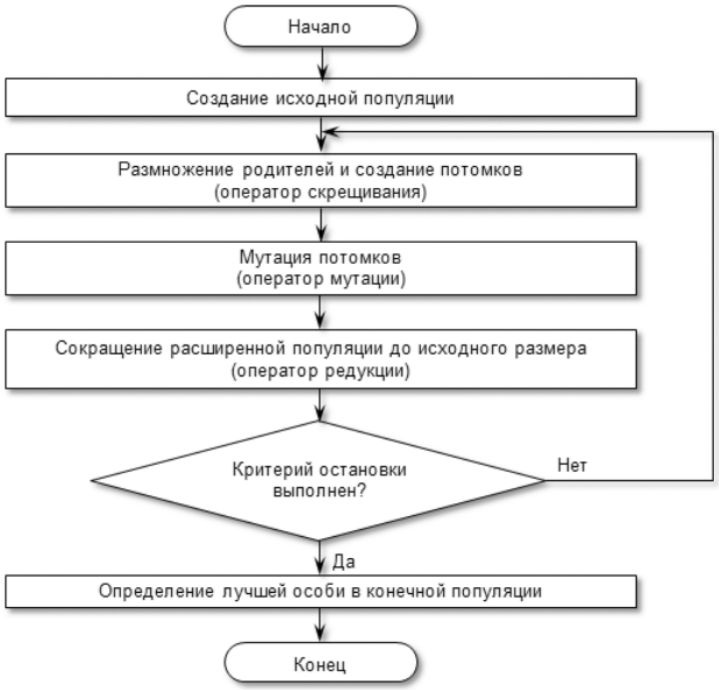

Рисунок 1

В этом разделе рассмотрим построенный алгоритм кластеризации, использующий эволюционный подход. На этом основании такой алгоритм кластеризации можно отнести к категории гибридных, так как эволюционный подход характерен для генетических алгоритмов.

Основанием для его разработки послужила необходимость использования в системах поддержки принятия решений универсального алгоритма по отношению к предпочтениям аналитика в различных ситуациях, то есть формирование кластеров не по одному, а по совокупности нескольких критериев.

В представленном алгоритме обеспечивается универсальность использования в системах поддержки принятия решений путем изменением расчета фитнес-функции (функции приспособленности) используемых в алгоритме хромосом без изменения структуры самого алгоритма. В рассматриваемом случае при кластеризации, кроме близости объектов к центру кластера в метрическом пространстве признаков (критерий формирования более плот-

ных кластеров), учитывается расстояние между кластерами (критерий максимального различия кластеров между собой).

Исходные данные представляют собой набор векторов в пространстве из n признаков $x^p = (x_1^p, x_2^p, \ldots, x_n^p)$, $p = 1, \ldots, k$. Каждый p-й вектор относится к соответствующему группируемому объекту. В составе исходных данных присутствует также требуемое количество кластеров – m.

Генетический алгоритм традиционно работает с популяцией хромосом, модифицируемых в итерационном процессе эволюции. В нашем случае в популяции размером w каждая хромосома отображает один из возможных вариантов кластеризации, в качестве которой используется вектор $g_l^t = (g_{l_1}^t, g_{l_2}^t, \ldots, g_{l_k}^t)$, $t = 1, \ldots w$ (t – номер хромосомы внутри популяции). Количество компонентов вектора-хромосомы является фиксированным и равным количеству кластеризуемых объектов, то есть k. В отличие от традиционного подхода, при кодировании значений компонентов хромосомы (значений «генов») используется десятичная система счисления. Целое значение компонента (r-го «гена» в хромосоме) кодирует номер кластера, к которому следует отнести r-й объект, и находится в интервале $[1, m]$.

Критерий качества кластеризации определяется с учетом среднего размера $\bar{R}_t$ формируемых кластеров

$$\bar{R}_t = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{k_j} \sum_{p=1}^{k_j} d^2(x_j^p, x_{0_j}) \tag{1}$$

и оценки удаленности кластеров друг от друга

$$\bar{S}_t = \frac{1}{C_m^2} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d^2(x_{0_i}, x_{0_j}), \tag{2}$$

где k_j – число входных образов (объектов), попавших в j-й кластер; x_{0_j} – точка центра j-го кластера в пространстве признаков; $d^2(x_j^p, x_{0_j})$ – квадрат евклидова расстояния от входного образа до центра своего j-го кластера; C_m^2 – число сочетаний из m по 2; m – количество кластеров; $d^2(x_{0_i}, x_{0_j})$ – квадрат расстояния между центрами i-го и j-го кластеров.

Данный критерий качества исполняет в генетическом алгоритме роль функции приспособленности потенциальных решений-хромосом и рассчиты-

вается по следующей формуле:

$$\mu(g_l^t) = \frac{\bar{S}_t}{\bar{R}_t}. \tag{3}$$

Как легко видеть из формулы (3), из двух кластеризаций более плотной группировкой образов кластеризуемых объектов вокруг центра своего кластера в пространстве признаков и более удаленным расположением центров этих кластеров друг от друга характеризуется та, у которой большее значение фитнес-функции μ.

В алгоритме использованы генетические операторы селекции, кроссовера, мутации, а также специальный оператор фильтрации, исключающий из рассмотрения хромосомы, отображающие недопустимую кластеризацию, то есть не соответствующую заданному числу кластеров. Вероятность такого события достаточно высока ввиду случайного характера выбора вышеназванных значений компонентов при формировании начальной популяции хромосом, а также вследствие работы других генетических операторов. Для учета такого события используется функция

$$\phi(g_l^t) = \begin{cases} 1, & \text{если } \exists r \in [1,k] \text{ такое, что } g_{l_r}^t = j \text{ для } \forall j \in [1,m]; \\ 0, & \text{если } \exists j \in [1,m] \text{ такое, что} \bar{\exists}\ r \in [1,k], \text{для которого } g_{i_r}^t = j. \end{cases}$$

В результате выполнения данного оператора рассматриваемая хромосома либо продолжает участвовать в дальнейшей работе алгоритма при $\phi(g_l^t) = 1$, либо она удаляется при $\phi(g_l^t) = 0$. Другими словами, оператор фильтрации – Fil ставит в соответствие хромосоме, над которой он выполняется, либо ее саму, либо пустое множество:

$$Fil(g_l^t) = \begin{cases} g_l^t, & \text{если } \phi(g_l^t) = 1; \\ \emptyset, & \text{если } \phi(g_l^t) = 0. \end{cases}$$

Основное описание предлагаемого алгоритма представлено блок-схемой, которая показана ниже. Те шаги алгоритма, которые не описаны непосредственно в блок-схеме, сопроводим пояснениями.

Шаг 1. Инициализация алгоритма. Исходные данные для работы алго-

ритма состоят из множества векторов $\{x^p\}$, (см. выше), количество искомых кластеров – m, количество группируемых объектов – k, количество хромосом в популяции – w, количество поколений в процессе эволюции – L, вероятность выполнения генетического оператора кроссовера – P_c, вероятность выполнения генетического оператора мутации – P_m.

Шаг 4. Необходимую последовательность равномерно распределенных целых чисел из диапазона $[1, m]$ можно получить, воспользовавшись соответствующей вычислимой функцией $R(a, b)$, использующей в свою очередь вычислимую функцию Rnd – генератор равномерно распределенных случайных действительных чисел из диапазона $[0, 1]$. Для генерации равномерно распределенных случайных чисел можно воспользоваться одним их известных методов, например, линейным конгруэнтным методом.

Один из возможных вариантов построения алгоритма функции $R(a, b)$ приведен ниже, где a и b – границы диапазона $[a, b]$ генерируемых целых чисел.

A1. Задать исходные данные – целые числа a и b $(a < b)$.

A2. Рассчитать приращение $\Delta = 1/(b - a + 1)$.

A3. Получить равномерно распределенное в интервале $[0, 1]$ вещественное случайное число $\lambda = Rnd$.

A4. Положить $i := 1$.

A5. Проверить: логическое выражение $(i - 1) \cdot \Delta \leq \lambda \leq i \cdot \Delta$ истинно? Если да, то закончить. Присвоить функции $R(a, b)$ искомое значение: $a + (i - 1)$.

A6. Положить $i := i + 1$ и перейти к шагу A5.

Шаг 5. Функция $\phi(g_l^t)$ вычисляется согласно (4).

Шаг 11. На данном шаге используется вычислимый оператор селекции Sel, отбирающий хромосому для дальнейшего воспроизводства по принципу «выживает лучший». При отборе участвует процедура «виртуальная рулетка», позволяющая хромосомам с большим значением фитнес-функции быть отобранными с большей вероятностью. Ниже приведен алгоритм оператора селекции.

Алгоритм B. Выполнить селекцию – отобрать хромосому

$$g_{l-1}^{\alpha} := Sel(g_{l-1}^{1}, g_{l-1}^{2}, ..., g_{l-1}^{w}) :$$

B1. Задать исходные данные – множество хромосом $\{g_{l-1}^{t}\}$ и множество значений соответствующих им фитнес-функций $\{\mu(g_{l-1}^{t})\}$, $t = 1, \ldots, w$.

B2. Рассчитать суммарное значение фитнес-функции по всей популяции

$$S = \sum_{t=1}^{w} \mu(g_{l-1}^{t}).$$

B3. Получить равномерно распределенное в интервале $[0, 1]$ вещественное случайное число $\lambda = Rnd$.

B4. Положить $\gamma := 1$ и $t := w$.

B5. Проверить: логическое выражение $(\gamma - \frac{\mu(g_{l-1}^{t})}{S}) \leq \lambda \leq \gamma$ истинно? Если да, то закончить: $g_{l-1}^{\alpha} := g_{l-1}^{t}$.

B6. Положить $\gamma := \gamma - \frac{\mu(g_{l-1}^{t})}{S}$ и $t := t - 1$. Перейти к пункту B5.

Шаг 13. Используется вычислимая функция $Rul(P_c)$, которая возвращает значение либо 1 в случае наступления случайного события с вероятностью P_c , либо 0 в противном случае. Алгоритм вычисления приведен ниже. Блок-схема разбита для удобства на три части, представленные на рисунках 2, 3, 4.

Алгоритм C. Вычислить значение функции $Rul(P_c)$:

C1. Задать исходные данные – значение вероятности P_c.

C2. Получить равномерно распределенное в отрезке $[0, 1]$ вещественное случайное число $\gamma = Rnd$.

C3. Проверить: логическое выражение $\gamma \leq P_c$ истинно? Если да, то присвоить функции $Rul(P_c)$ искомое значение 1, если нет – значение 0.

C4. Закончить.

Шаг 14. На данном шаге используется вычислимый оператор кроссовера Cr, скрещивающий две хромосомы и конструирующий в результате новую хромосому для помещения ее в новую популяцию. Для этого равновероятным случайным образом определяется точка кроссовера $r_c \in [2, k-1]$, т. е. номер «гена», с которого будет производиться обмен значениями рассматриваемых хромосом. Для этого можно воспользоваться рассмотренной выше вычислимой функцией $R(2, k-1)$. Ниже приведен алгоритм оператора кроссовера.

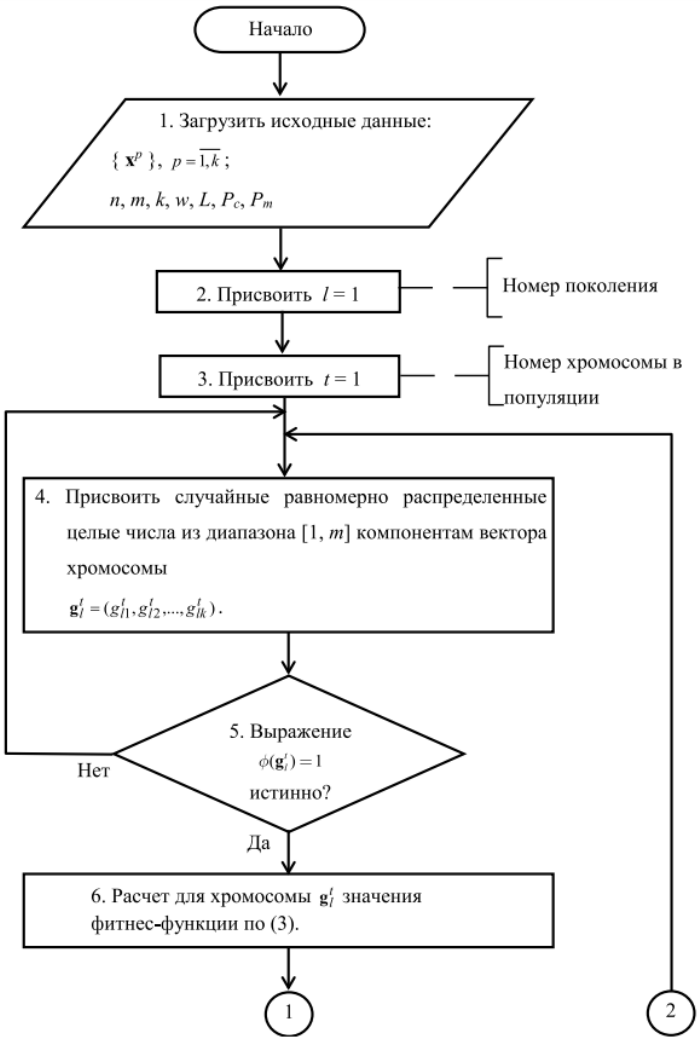

Рисунок 2

Алгоритм D. Выполнить кроссовер – отобрать хромосому $g_{l-1}^c = Cr(g_{l-1}^\alpha, g_{l-1}^\beta)$:

D1. Задать исходные данные – хромосомы g_{l-1}^α, g_{l-1}^β.

D2. Определить точку кроссовера $r_c = R(2, k-1)$.

D3. Построить хромосомы $\tilde{g}_{l-1}^\alpha$, $\tilde{g}_{l-1}^\beta$:

$\tilde{g}_{l-1}^\alpha = (g_{(l-1)1}^\alpha, g_{(l-1)2}^\alpha, ..., g_{(l-1)(r_c-1)}^\alpha, g_{(l-1)r_c}^\alpha, ..., g_{(l-1)k}^\alpha)$,
$\tilde{g}_{l-1}^\beta = (g_{(l-1)1}^\beta, g_{(l-1)2}^\beta, ..., g_{(l-1)(r_c-1)}^\beta, g_{(l-1)r_c}^\beta, ..., g_{(l-1)k}^\beta)$.

D4. Проверить: логическое выражение $Rul(P_c) = 1$ истинно? Если да, то $g_{l-1}^c = g_{l-1}^\alpha$, если нет – $g_{l-1}^c = g_{l-1}^\beta$.

D4. Закончить.

Шаг 20. На данном шаге используется вычислимый оператор мутации Mu, заменяющий значение случайно определенного «гена» с номером $r_m[1, k]$

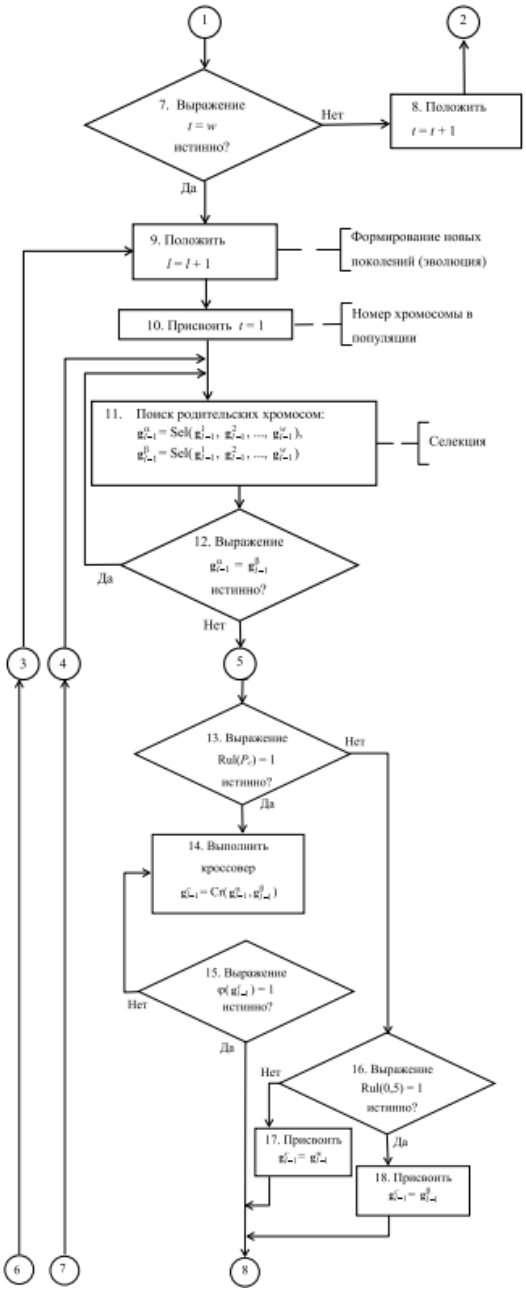

Рисунок 3

на новое случайно определенное значение с равной вероятностью из интервала $[1, m]$. Поиск значения r_m осуществляется также с равной вероятностью, для чего можно воспользоваться рассмотренной выше вычислимой функцией $R(1, k)$. Ниже приведен алгоритм оператора мутации.

Алгоритм Е. Выполнить мутацию хромосомы :

Е1. Задать исходные данные – хромосому .

Е2. Определить номер мутирующего «гена»: $r_m = R(1, k)$.

Е3. Присвоить новое значение мутирующему «гену»: $g^c_{(l-1)r_m} = R(1, m)$.

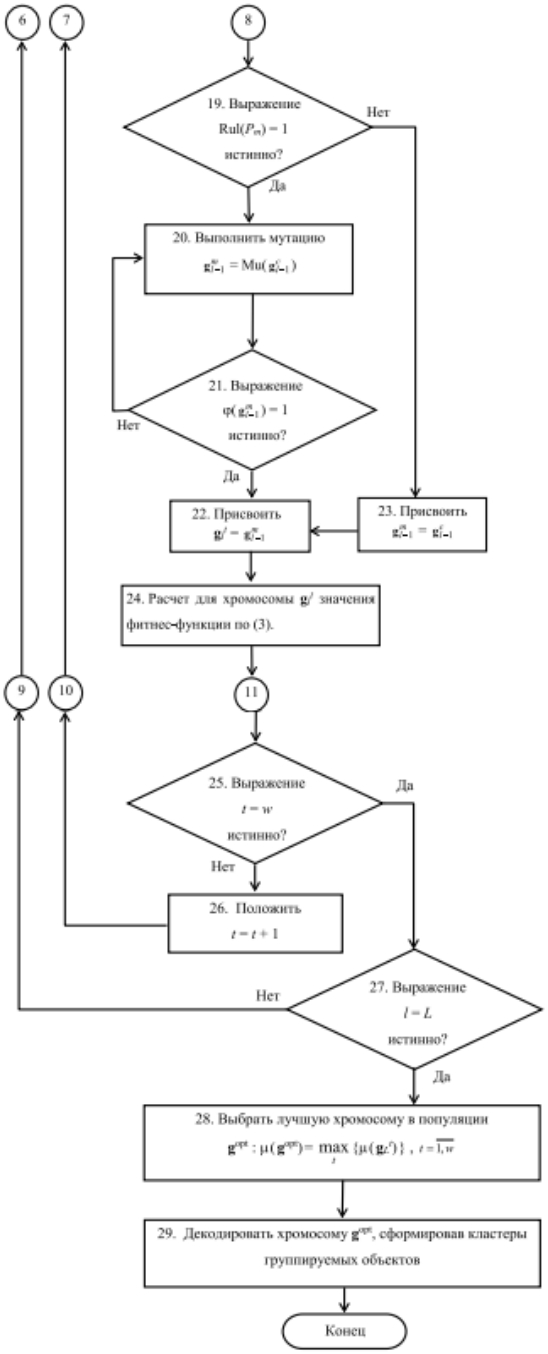

Рисунок 4

E4. Закончить.

Шаг 28. Поиск лучшей хромосомы g^{opt} в популяции последнего поколения (L – номер последнего поколения) осуществляется обычным перебором значений фитнес-функций всех хромосом с поиском хромосомы, имеющей максимальное значение данной функции.

Шаг 29. Декодирование найденной хромосомы предполагает сортировку ее компонентов («генов») по их значению, поскольку каждый компонент

вектора хромосомы соответствует группируемому объекту, а его значение – номеру кластера.

Поскольку работа алгоритма представляет собой стохастический процесс, то его сходимость является также категорией вероятностной, требующей своей оценки. Для выполнения данного анализа в работе были проведены исследования эффективности работы алгоритма в несколько этапов.

Первый этап состоял в подготовке исходных данных в виде множества многомерных векторов значений признаков абстрактных кластеризуемых объектов. Значения подбирались случайным образом с помощью несложной программы. При этом разброс координат объектов по измерениям внутри кластеров поддерживался в рамках заданной величины дисперсии. Данные готовились для двух серий вычислительных экспериментов – с сильно и слабо выраженным удалением кластеров. Во втором блоке данных расстояние между кластерами, оцениваемое методом «ближнего соседа», было соизмеримо с дисперсией значений координат внутри кластеров.

На втором этапе были проведены серии вычислительных экспериментов с использованием специально разработанной для этой цели программы. Работа алгоритма в рамках данной программы выполнялась в различных режимах, отличающихся варьированием размера популяции хромосом w и количества отработанных алгоритмом поколений эволюции L.

Наконец, третий этап состоял из вычислительного сравнительного эксперимента на реальных данных.

Для оценки приведенного алгоритма его автором был использован SOM-алгоритм , который по характеру является также стохастическим и гибридным, а также один из популярных алгоритмов – “k-means”. Вычисления с использованием нейронных сетей Кохонена и алгоритма “k-means” выполнялись в программе Deductor Studio Basegroup Labs. Результат оказался ожидаемо успешным, так как алгоритмы SOM и “k-means” учитывают только один критерий кластеризации, тогда как в задаче лучшее решение оценивается двумя критериями, такими же, как и в предлагаемом алгоритме. Чтобы поставить сравниваемые алгоритмы в равное положение, не связанное с приоритетами принятия решений в , в предлагаемом алгоритме была изменена фитнес-функция (смотреть функцию (3)) на значение, учитывающее только

один критерий кластеризации – близость образов кластеризуемых объектов к центру кластера в пространстве измерений. Именно этот критерий используется в алгоритмах SOM и "k-means".

Таким образом, в обоих случаях качество кластеризации оказалось лучше в описанном алгоритме. Это объясняется тем, что в данном алгоритме пространство поиска покрывает все допустимое множество решений (генетический оператор мутации обеспечивает появление в рассматриваемом множестве таких решений, которых не было заложено в генетическом материале начальной стартовой популяции), в то время как в сравниваемых алгоритмах пространство решений зависит от их инициализации (стартовых значений координат центров кластеров в "k-means" и координат нейронов в SOM). Сравнительная оценка алгоритмов по производительности учитывала то обстоятельство, что время работы алгоритма SOM напрямую зависит от количества эпох обучения нейронной сети входными данными, а исследуемого алгоритма – от размеров популяции хромосом и числа реализованных поколений эволюции. В процессе повторных вычислений было выявлено, что стабилизация алгоритма SOM по точности кластеризации наступает в диапазоне от 1500 до 2000 эпох. Для предпочтительной для данного алгоритма оценки было определено время его работы при реализации 1500 эпох обучения нейронной сети на 32-разрядном микропроцессоре с тактовой частотой 1.7 ГГц. Среднее значение времени его работы составило 4.5 с. Среднее время работы предлагаемого алгоритма при $w = 200$ и $L = 200$ на том же микропроцессоре составило 0.89 с, а алгоритма "k-means" – 0.64 с. Это говорит о соизмеримости производительности сравниваемых алгоритмов.

Резюмируя все сказанное выше о представленном эволюционном алгоритме, можно утверждать, что выбранный для реализации генетический алгоритм кластеризации является гибким по отношению к предпочтениям аналитика в процессе принятия решений, так как позволяет выполнять кластеризацию, исходя из различных критериев, например, максимального взаимного удаления кластеров, близости геометрических образов группируемых объектов к центру кластера, других критериев. Это достигается изменением расчетной формулы фитнес-функции, учитывающей необходимое сочетание критериев кластеризации без изменения структуры алгоритма. Как и лю-

бой генетический алгоритм, данный алгоритм является простым в его программной реализации, что делает его надежным в использовании. Высокая надежность алгоритма была подтверждена проведенными вычислительными экспериментами на данных, отражающих различные кластерные структуры. Эксперименты показали быструю сходимость алгоритма, оцениваемую по небольшому числу поколений эволюции в процессе нахождения искомого решения. При этом необходимы небольшие вычислительные ресурсы компьютера в виде его оперативной памяти, так как надежная работа алгоритма не требует большой популяции генетических хромосом.

Алгоритм отличает нечувствительность к инициализации, так как в процессе эволюции хромосом за счет использования генетических операторов алгоритм полностью покрывает все множество допустимых решений, что, в итоге, обеспечивает устойчивость алгоритма и высокое качество кластеризации. Проведенный вычислительный эксперимент на реальных данных показал не только хорошее качество кластеризации, но и вполне приемлемую производительность представленного алгоритма, соизмеримую с производительностью общепризнанных алгоритмов SOM и "k-means".

Для реализации описанного алгоритма нами был выбран язык Python. Листинг программы представлен в приложении А. Многократный запуск алгоритма продемонстрировал уверенное разбиение массива на три кластера. Ввиду современной обстановки, нам не представилось возможности провести полноценные консультации с археологами и дать научную интерпретацию этого факта, однако в будущем обсуждения непременно состоятся.

Третий раздел призван сделать общую картину методов кластеризации более полной. В ней приведен обзор современных методов и алгоритмов кластеризации в контексте их реализации в языке R.

По мнению большинства специалистов свободно распространяемая система R, является наиболее полной, надежной и динамично развивающейся статистической средой, объединяющей язык программирования высокого уровня и мощные библиотеки программных модулей для вычислительной и графической обработки данных.

Сегодня статистическая среда R (http://www.r-project.org) является признанным лидером среди некоммерческих систем статистического анализа и

неуклонно становится незаменимой при проведении научно-технических расчетов в университетских центрах и многих ведущих фирмах. Расширение библиотек программных модулей за счет усилий множества разработчиков привело к возникновению распределенной системы хранения и распространения пакетов R, то есть "CRAN" (от «Comprehensive R Archive Network»; http://cran.r-project.org), которая обладает также развитой системой информационной поддержки.

Значительный раздел среды R посвящен машинному обучению и собственно статистической обработке массивов данных. Хорошим путеводителем к поиску того, какие ресурсы, пакеты и библиотеки могут быть использованы для решения тех или иных задач пользователя, является страница CRAN Task Views (http://cran.r-project.org/web/views). Список только признанных фундаментальных монографий типа "Используем R!" (Use R!) насчитывает уже несколько сотен томов. В многочисленных фундаментальных руководствах детально описаны функции и пакеты R, использующие широкий набор статистических методов, таких как классические линейные и нелинейные регрессионные модели, различные формы дисперсионного анализа, алгоритмы иерархической кластеризации, анализ временных рядов, деревья классификации и регрессии. Отдельно можно отметить библиографические источники, в названии которых имеется термин "Data Mining", в частности, на www.kdnuggets.com).

В **заключении** приведены результаты магистерской работы.

Основные результаты

В настоящей работе сделана попытка дать обзор как традиционным методам и алгоримам кластеризации, так и современным, в том числе гибридным. Из представленного материала видно, как велико их разнообразие. Основное внимание было сосредоточено на сравнительно новом гибридном методе кластерного анализа – кластерном анализе с использованием эволюционных вычислений. Мотивом для этого послужили преимущества, которыми располагает этот алгоритм. Перечислим их:

— возможность работы с категорийными данными;

— универсальность, т.е. возможность формирования кластеров не по одному, а по совокупности нескольких критериев;
— возможность учитывать предпочтения аналитика в различных ситуациях;
— надежность в использовании применительно к различным кластерным структрурам;
— нечувствительность к инициализации;
— производительность, сопоставимую с алгоритмами SOM и "k-means".

Программа на основе указанного алгоритма была использована для исследования конкретного массива данных (см. Приложение Г) археологических раскопок, а именно амфор. Выявлено наличие трех кластеров, интерпретация не была произведена ввиду отсутствия необходимых знаний. Можно предполагать либо наличие в течение некоторого протяженного исторического периода последовательной смены стандартов внешнего вида амфор, либо наличие трех основных центров производства – законодателей стандартов внешнего вида. Заметим, что датировка при маркировке образцов нами не учитывалась.