

На правах рукописи



**Аль Хашеди Адам Абдо Ахмед**

**Разработка математических моделей и алгоритмов классификации  
динамических объектов**

Специальность 05.13.18 –  
Математическое моделирование, численные методы  
и комплексы программ

Автореферат  
диссертации на соискание ученой степени  
кандидата технических наук

Казань - 2018

Работа выполнена на кафедре информатики и прикладной математики в федеральном государственном бюджетном образовательном учреждении высшего образования «Казанский национальный исследовательский технологический университет».

Научный руководитель: доктор педагогических наук, кандидат технических наук, профессор **Нуриев Наиль Кашапович**

официальные оппоненты: **Новикова Светлана Владимировна**  
доктор технических наук, доцент, федеральное государственное бюджетное образовательное учреждение высшего образования «Казанский национальный исследовательский технический университет им. А.Н. Туполева - КАИ», профессор кафедры прикладной математики и информатики;

**Ажмухамедов Искандар Маратович**  
доктор технических наук, доцент, федеральное государственное бюджетное образовательное учреждение высшего образования «Астраханский государственный университет», заведующий кафедрой информационной безопасности.

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный университет».

Защита состоится «07» декабря 2018 года в 14:00 часов на заседании диссертационного совета Д 212.080.13, созданного на базе ФГБОУ ВО «Казанский национальный исследовательский технологический университет» по адресу: 420015, г. Казань, ул. К. Маркса, 68, зал заседаний Ученого совета (А-330).

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Казанский национальный исследовательский технологический университет» и по адресу <http://www.kstu.ru/servlet/contentblob?id=252038>.

Автореферат разослан « \_\_\_\_ » октября 2018 года.

Ученый секретарь диссертационного совета Д 212.080.13, доктор технических наук, профессор



Клинов  
Александр  
Вячеславович

## ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИИ

### Актуальность темы

Круг областей человеческой деятельности, где встречаются задачи, связанные с проблемой распознавания образов, чрезвычайно обширен и продолжает неуклонно расширяться. Системы медицинской диагностики, образовательные SMART-системы, разного рода системы охраны и сигнализации, системы поиска и обработки информации, - вот далеко не полный список тех сфер, где востребованность этих задач не вызывает сомнений.

Одной из ключевых проблем, возникающих при обработке информационных потоков, является проблема классификации, т.е. отнесения каждого объекта, обнаруженного информационной системой, к соответствующему классу по наличию некоторых характерных признаков.

Явление информационной глобализации, характерное для современного мира, определяется прежде всего чрезвычайно большими объемами информации и значительным числом классифицирующих признаков. Это актуализировало разработки новых математических моделей и высокопроизводительных алгоритмов для описания и обработки информационных потоков.

### Цель и задачи исследования

Целью исследования является разработка математических моделей, алгоритмов и программных комплексов ориентированных на решение задач распознавания образов с произвольной размерностью пространства классифицирующих признаков при наличии динамических изменений.

Для достижения цели необходимо решить следующие **задачи**:

1. Разработать алгоритм с возможностями саморазвития и самоорганизации, который бы мог эффективно использоваться для решения задач распознавания независимо от размерности пространства признаков;
2. Построить математическую модель самоорганизующейся системы эволюционирующих кластеров и указать технику ее использования для решения задач эффективного управления и прогнозирования;
3. Разработать и апробировать алгоритмы кластеризации, действующие на заданном множестве, характеризующих признаков определенной предметной области;
4. Создать специализированный программный комплекс для практической реализации этих алгоритмов;
5. Разработать алгоритмы классификации объектов на основе Байесовских процессов;
6. Провести экспериментальные исследования эффективности разработанных методов;

**Объектом исследования** является классификация объектов в задачах распознавания образов.

**Предметом исследования** является математическое моделирование классификации динамических объектов в задачах распознавания образов.

### **Методы исследований**

При решении поставленных задач использовались методы вычислительной математики, методы математической статистики, аппарат анализа временных рядов, теория искусственных самоорганизующихся нейронных сетей для графических массивов базы данных с применением процесса распознавания с учителем, алгоритмы обратного распространения ошибки, теория кодирования/декодирования, математическая модель персептрона а также методы объектно-ориентированного программирования (ООП) на языке Java в среде Netbeans.

### **Научная новизна**

1. На основе аппарата кластерного анализа разработана математическая модель динамической кластеризации применительно к решению задач распознавания образов.
2. Разработан самоорганизующийся алгоритм для моделирования и анализа динамических процессов в задачах распознавания образов, учитывающий эволюционные изменения кластерных образований.
3. Показано, что кластеры следует рассматривать как динамические образования, изменяющиеся под воздействием потока распознаваемых образов.
4. Разработан алгоритм кластеризации, работоспособный в пространстве больших данных.

**Теоретическая значимость работы** состоит в следующем:

- Проведена экспертиза методов, используемых в различных задачах распознавания образов и получено статистическое подтверждение ее результатов.
- Предложена математическая модель для решения задачи распознавания образов с помощью самоорганизующейся кластерной структуры.
- С помощью аппарата анализа временных рядов исследован процесс эволюционных изменений кластерных образований.
- На основе вероятностного подхода построены модель и алгоритм вычислительной процедуры, позволяющий идентифицировать принадлежность объекта к определенному классу предметной области.

**Практической значимостью работы** является следующее:

- Разработан и апробирован алгоритм динамической кластеризации, действующий на заданном множестве характеризующих признаков определенной предметной области.
- Разработан специализированный программный комплекс для практической реализации алгоритмов .

- Разработано математическое и программное обеспечение на основе предложенной модели динамической кластеризации для классификации запросов-задач коммуникационных услуг, пригодное для практического использования в режиме онлайн.

### **Основные результаты, выносимые на защиту:**

- Результаты экспертизы методов, используемых в различных задачах распознавания образов.
- Математическая модель динамической кластеризации для решения задач распознавания образов.
- Алгоритм динамической кластеризации, действующий на заданном множестве характеризующих признаков определенной предметной области.
- Математическая модель эволюционных изменений саморазвивающихся кластерных образований.
- Комплекс программ для классификации запросов-задач коммуникационных услуг.

**Степень достоверности и апробация результатов.** Достоверность полученных результатов исследования подтверждается корректным применением математического аппарата и широким спектром публикаций и выступлений на **Международных конференциях:** Международная научно-практическая конференция «Электронная Казань 2016», (г. Казани – 2016г.), Международная научно-практическая конференция «Наукофера», (г. Смоленск – апреля 2016г.), Международная научно-практическая конференция «Электронная Казань 2017», (г. Казани – 2017г.), Международная научно-практическая конференция «Электронная Казань 2018», (г. Казани – 2018г.), Фундаментальные и прикладные научные исследования: актуальные вопросы, достижения и инновации XIV Международной научно-практической конференции «Наука и Просвещение» (г. Пенза – 2018г.).

### **Публикации**

По теме диссертации опубликовано 16 работ, в том числе 11 статей, 10 из которых в журналах, рецензируемых ВАК, 16 работ изложены в сборниках материалов международных и всероссийских научно-практических конференций. Автор лично принимал непосредственное участие в создании математических моделей, разработке программного обеспечения и внедрении результатов исследований.

### **Структура и объем работы**

Основное содержание диссертационной работы изложено на 166 страницах машинописного текста, содержит 45 рисунков, 24 таблицы и состоит из введения, четырёх глав, заключения, списка литературы, содержащего 140 наименований, и 3 приложений.

**Во введении** обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследований, научная новизна, теоретическая и практическая ценность полученных результатов, приведены сведения об использовании, реализации и апробации результатов работы, структуре диссертации.

**В первой главе** по данным литературных источников проведен обзор методов распознавания образов, их классификация и сравнительный анализ в соответствии с целью исследования, проанализированы известные целевые функции, применяемые в задачах распознавания образов, сформулированы основные задачи исследований диссертационной работы. Рассмотрены различные типы задач распознавания образов, и история их развития. Подчеркнуто чрезвычайно широкое распространение задач распознавания образов в самых различных сферах деятельности человека и их большое практическое значение.

Также в первой главе излагаются понятия искусственного нейрона, весов, активационных функций, алгоритм обратного распространения ошибки; описаны многослойные искусственные нейронные сети; архитектура и свойства сетей с обратным распространением ошибки (Error Backpropagation); обучение с учителем; рассмотрены алгоритмы классификации объектов в задачах распознавания образов: алгоритм Хо-Кашьяпа; метод Байеса; алгоритмы кластеризации: алгоритм k-средних (k-means); алгоритм FOREL; алгоритм с полным перебором.

**Во второй главе** приведены статистически обработанные результаты экспертных оценок методов и алгоритмов, используемых в задачах распознавания образов; построена математическая модель самоорганизующейся системы кластеров и показана возможность ее использования для решения задач распознавания образов; разработан и апробирован алгоритм динамической кластеризации, действующий на заданном множестве характеризующих признаков определенной предметной области и предложена техника анализа эволюционных изменений кластерных структур.

В процедуре экспертного оценивания были выделены две стадии. На первой стадии восемь независимых специалистов оценивали значимость некоторого набора свойств алгоритмов распознавания образов для четырех категорий задач, различающихся размерностью пространства классифицирующих признаков и требованиями к быстродействию систем, в которых они реализуются. В таблице 1 представлены сводные усреднённые оценки, приведённые к единичной шкале, а на рисунках 1-4 выполнена их визуализация с помощью диаграмм Кивиата.

Таблица 1. Сводные усреднённые оценки, приведённые к единичной шкале

№	Наименование свойства	Размерность пространства признаков			
		малая		большая	
		быстродействие		быстродействие	
		низкое	высокое	низкое	высокое
1	Простота реализации	0.90625	0.88750	0.67500	0.73125
2	Чувствительность к шумам	0.75625	0.81250	0.78750	0.86875
3	Способность к саморазвитию	0.91875	0.93125	0.93475	0.98125
4	Способность к самоорганизации	0.93125	0.93125	0.95625	0.98750
5	Интенсивность потока данных	0.63125	0.81875	0.80000	0.90000
6	Возможность работы в режиме реального времени	0.59375	0.89575	0.63750	0.96250

Данные, содержащиеся в табл. 1, показывают, что независимо от размерности пространства признаков и показателей быстродействия, эксперты единодушно признали способность к саморазвитию и самоорганизации важнейшими качествами современных алгоритмов распознавания образов. Это означает, что система распознавания образов высокого уровня должна быть самообучаемой, обладать возможностями масштабирования, и обработки динамических последовательностей.

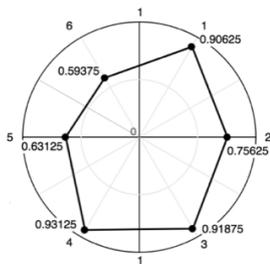


Рисунок 1. Малая размерность пространства признаков. низкое быстродействие

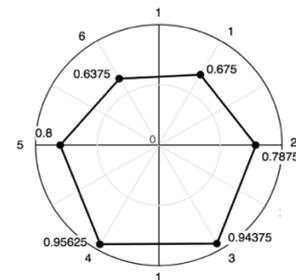


Рисунок 3. Большая размерность пространства признаков. Низкое быстродействие

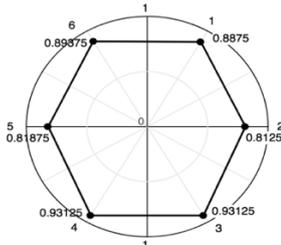


Рисунок 2. Малая размерность пространства признаков. высокое быстродействие

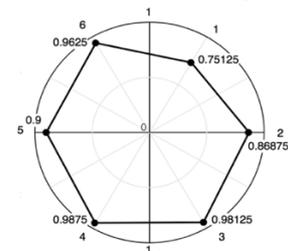


Рисунок 4. Большая размерность пространства признаков. Высокое быстродействие

На следующей стадии исследования той же группе экспертов было предложено расположить несколько хорошо им известных алгоритмов, используемых для решения задач распознавания, в порядке убывания эффективности для двух категорий задач, различающихся по размерности пространства классифицирующих признаков.

Итоговые результаты работы группы экспертов представлены в таблице 2.

Таблица 2. Суммарные ранги различных методов распознавания образов

№	Оцениваемые методы	Суммарные ранги экспертных оценок	
		$\sum_{i=1}^m r_{ij}$	
		Размерность пространства признаков	
		малая	большая
1	Метод опорных векторов	53.5	33.5
2	Метод ближайшего соседа	23	54.5
3	Метод Байеса	24.5	20
4	Метод ветвей и границ	41.5	26.5
5	Метод потенциальных функций	57	22.5
6	Метод внутригрупповых средних	54.5	22
7	Нейронные сети	18	53.5
8	Метод волновой кластеризации	16	55.5

Согласованность мнений экспертов проверялась с помощью коэффициента ранговой корреляции (конкордации) Кендалла-Смита

$$K = \frac{\sum_{i=1}^n (\sum_{j=1}^m r_{ij} - \bar{r})^2}{\frac{1}{12} (m^2(n^3 - n) - m \sum_{j=1}^m T_j)} , \quad (1)$$

где  $n$  – количество, оцениваемых показателей;  $m$  – численный состав экспертной группы;  $r_{ij}$  – ранг, присвоенный  $i$ -ому показателю  $j$ -ым экспертом. Величина  $\bar{r}$  вычисляется по формуле

$$\bar{r} = \frac{\sum_{i=1}^n \sum_{j=1}^m r_{ij}}{n} , \quad (2)$$

а

$$T_j = w_j^3 - w_j , \quad (3)$$

где  $w_j$  – число связанных рангов, выставленных в ходе оценки  $j$ -ым экспертом.

Проверка статистической значимости значений коэффициентов ранговой корреляции производилась по критерию  $\chi^2$  для числа степеней свободы  $f = n - 1$ , расчетное значение которого будет

$$\chi^2 = \frac{\sum_{i=1}^n (\sum_{j=1}^m r_{ij} - \bar{r})^2}{\frac{1}{12} (m * n (n + 1) - \frac{1}{n-1} \sum_{j=1}^m T_j)} . \quad (4)$$

Табличное значение этого критерия при уровне значимости 5%  $\chi_{0.05}^2 = 14.1$ .

Вычисленное по формуле (1) значение коэффициента ранговой корреляции для малой размерности пространства классифицирующих признаков  $K = 0.854$ , что, в совокупности с высоким значением критерия  $\chi_{расч}^2 = 47.8$ , подтверждает согласованность мнений экспертов и безусловную статистическую достоверность полученных результатов. Лидерами предпочтений экспертов для задач малой размерности являются алгоритмы метода волновой кластеризации и нейронных сетей. Общеизвестно, что они вполне эффективны для задач распознавания с небольшой

размерностью пространства признаков. Важным их достоинством является способность к самообучению и, в какой то мере, самоорганизуемость, что позволяет достаточно успешно использовать эти методы в динамических системах распознавания образов. Далее с небольшим отставанием идут методы ближайшего соседа и Байеса. Эти алгоритмы не представляют сложности в реализации, однако свойствами самообучаемости и самоорганизуемости не обладают.

Обратимся теперь к результатам экспертных оценок тех же методов применительно к задачам с большой размерностью пространства классифицирующих признаков.

Значение коэффициента ранговой корреляции для большой размерности пространства признаков также оказалось высоким  $K = 0.702$ . Согласованность экспертных оценок не вызывает сомнения, поскольку  $\chi^2_{расч} = 39.3$ . Интересно отметить, что это как раз те методы, которые для задач с малой размерностью пространства признаков имели наивысшие ранги: метод ближайшего соседа, нейронные сети, метод волновой кластеризации, для задач большой размерности признаны экспертным сообществом неперспективными. Действительно, попытки применить эти методы к задачам большой размерности наталкивались на значительные трудности различной природы и успеха не имели. Суммарные ранги остальных методов значительно выше и довольно близки друг к другу. Это ясно указывает на то, что однозначных рекомендаций по выбору метода решения задач распознавания образов с большим набором классов и признаков не существует. Выбор же наиболее подходящего алгоритма представляет самостоятельную проблему и зависит не только от факторов размерности, но и от конкретных особенностей и содержательного смысла задачи.

Таким образом, проблема разработки алгоритма с возможностями саморазвития и самоорганизации, который бы мог эффективно использоваться для решения задач распознавания любой размерности, является очень востребованной и актуальной.

В втором разделе этой главы аппарат кластерного анализа применен к решению задачи распознавания образов в предположении, что образы объектов могут быть интерпретированы как векторы пространства конечной размерности с евклидовой метрикой  $\|x_i - x_j\| = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$ . В отличие от других известных работ, посвященных аналогичной тематике, здесь кластеры сферической формы предлагается рассматривать как динамические образования, положение и размеры которых изменяются по мере изменения состава обучающей выборки.

Для построения кластерных сфероидов необходимо все классифицирующие признаки привести к единым безразмерным единицам. С этой целью, путем анализа априорной информации и материалов обучающей выборки, устанавливаются точная верхняя  $x_{jsup}$  и точная нижняя  $x_{jinf}$  грани по каждому из  $n$  признаков и выполняется переход к безразмерным переменным  $y_j$ ,  $j = \overline{1, n}$  по формуле

$$y_j = M \frac{x_j - x_{jinf}}{x_{jsup} - x_{jinf}}, \quad (5)$$

где  $M$  – масштабирующий множитель, выбираемый из соображений удобства представления данных.

На множестве элементов  $\{y_j\}$  выделяется подмножество образов, образующих более или менее тесную группировку, и строится первый кластерный сфероид  $B^1(r)$  с центром в точке  $e^1_0$ , которая расположена в непосредственной близости от геометрического центра группировки. Точку  $e^1_0$  назовем начальной базовой точкой; верхний индекс указывает номер кластера, а нижний индекс номер точки.

Радиус сфероида  $r$  определяется сообразно целям исследования и особенностям группировки элементов обучающей выборки. Таким образом, в состав кластера  $B^1(r, e^1_0)$  входят элементы множества  $\{y_j\}$ , удовлетворяющие условию  $\|e^1_0 - y_i\| \leq r$ .

Положение полученного кластера следует уточнить, поскольку геометрический центр области, занимаемый исследуемым подмножеством обучающей выборки, как правило, не совпадает с центром тяжести группировки. С этой целью находится центр тяжести совокупности элементов, находящихся в границах сфероида  $B^1(r, e^1_0)$  по формуле

$$e^1_1 = \frac{1}{|B^1(r, e^1_0)|} \sum_{y \in B^1(r, e^1_0)} y_i \quad (6)$$

где  $|B^1(r, e^1_0)|$  – мощность подмножества элементов в составе сфероида  $B^1(r, e^1_0)$ .

Точка  $e^1_1$  будет первой базовой точкой и геометрическим центром сфероида  $B^1(r, e^1_1)$ . В результате этих действий в  $B^1(r, e^1_1)$  появляются элементы, которых не было в  $B^1(r, e^1_0)$  а часть элементов, входивших в состав  $B^1(r, e^1_0)$ , будет потеряна. Очевидно, что и в сфероиде  $B^1(r, e^1_1)$  геометрический центр не будет совпадать с центром тяжести, поэтому в том же порядке находятся координаты второй базовой точки  $e^1_2$ , являющийся геометрическим центром сфероида  $B^1(r, e^1_2)$  и т.д. Последовательность точек  $\{e^1_h\}$  сходится, как ограниченная последовательность, определенная на компакте, а значит за конечное число шагов достигается выполнение условия  $\|e^1_h - e^1_{h-1}\| < \varepsilon$ , где  $\varepsilon$  – любое заданное наперед положительное число, определяющее желаемую точность итерационной процедуры.

В ходе практической реализации описанного алгоритма может выясниться, что радиус сферической оболочки кластера выбран неудачно и часть элементов группировки в состав кластера  $B^1(r, e^1_h)$  не вошло и это противоречит содержательному смыслу задачи. В этом случае следует увеличить размеры сфероида  $B^1(r, e^1_h)$  так, чтобы самый удаленный от базовой точки  $e^1_h$  объект, из тех, что должны стать элементами кластера, попал на границу нового расширенного сфероида.

Обозначим множество объектов, которые необходимо дополнительно внедрить в кластер  $B^1$ , через  $\{\hat{y}_j\}$ , и найдем величину

$$R = \max_{\{\hat{y}_j\}} \|e^1_h - \hat{y}_j\| \quad (7)$$

,которая и будет определять радиус искомого кластерного сфероида  $B^1(\hat{r}, e^1_h)$  с центром в базовой точке  $e^1_h$ . Приращение радиуса при этом составит

$$\Delta r = R - r.$$

Естественно, что в результате таких действий центр тяжести кластера сместится относительно геометрического центра-точки  $e^1_h$ . Это смещение, однако, легко устраняется с помощью описанной выше итеративной процедуры. Подобное явление будет иметь место и тогда, когда формирование кластера на базе исходной обучающей выборки завершится и система начнет работать в режиме распознавания образов, обрабатывая поток объектов, поступающих на ее вход. Поэтому, если систему предполагается использовать как динамическую самообучающуюся структуру, необходимо осуществлять коррекцию положения и размеров кластерных сфероидов постоянно в течение всего периода эксплуатации. Заметим, однако, что по мере наполнения кластеров, смещение центра тяжести при поступлении новых объектов становится все менее и менее значительным. И в тех случаях, когда оно не выходит за границы  $\varepsilon$ - окрестности, корректирующее воздействие перестает быть обязательным.

Действие описанного алгоритма иллюстрируется на примере достаточно представительной обучающей выборки, элементами которой являются города Российской Федерации, а в качестве их характеризующих признаков выступают численный состав населения этих городов  $X$ , и территория ими занимаемая  $S$ . В таблицах 3 и 4 содержатся отдельные результаты вычислений, выполненных с помощью этого алгоритма.

Таблица 3. Этапы преобразования кластеров

№	Центр кластера [население, площадь]	Радиус $r$
Кластер №1		
1	[220698, 102.5]	7.2
2	[227734, 108.1]	7.2
Кластер №2		
1	[567468, 334.4]	10.2
2	[461127, 316.1]	8.2
3	[442760, 305.5]	7.2
4	[439437, 301.9]	8.2
5	[443997, 293.5]	6.2
Кластер №3		
1	[1059716, 788.7]	29.7
2	[978274, 557.8]	13.7
3	[1024047, 514.5]	11.7
4	[1073611, 485.8]	10.7
5	[1117677, 480.0]	11.7
6	[1117677, 480.0]	10.7

В таблице 4 приведен результат реализации разработанного алгоритма динамической кластеризации в данном примере (Фрагмент списка городов с указанием их кластерной принадлежности).

Таблица 4. Фрагмент списка городов с указанием их кластерной принадлежности

№	Город	Население. Ч.	Площадь $km^2$	№ Кластер
1	Новосибирск	1 602 915	506.67	3
2	Уфа	1 115 560	707	4
3	Орск	230 414	621.33	5
4	Казань	1 231 878	614.16	3
5	Волжский	326 055	229.12	2
6	Омск	1 178 391	566.9	3
7	Самара	1 169 719	541.382	3
8	Ростов-на-Дону	1 125 299	348.5	3
..	...	...	...	
137	Дербент	123 162	69.63	1

На рисунках 5 и 6 представлена визуализация процесса вычислений.

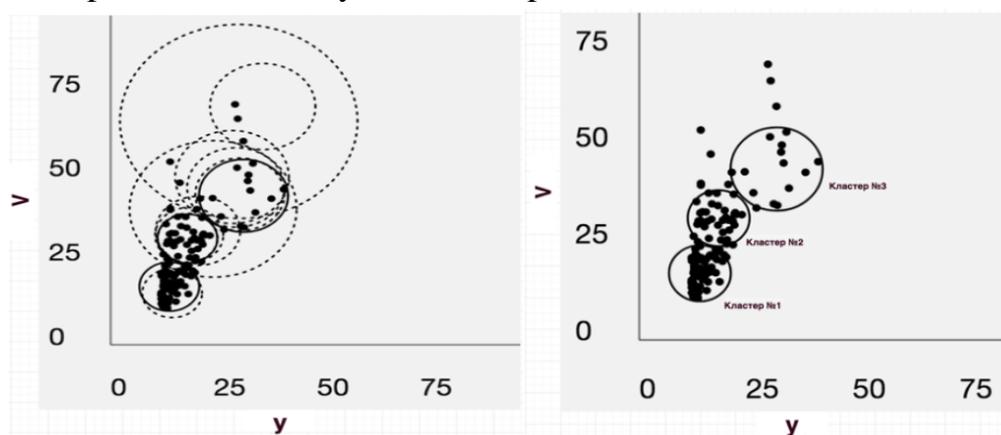


Рисунок 5. Этапы и результат расчетов, выполненных в соответствии с разработанным алгоритмом

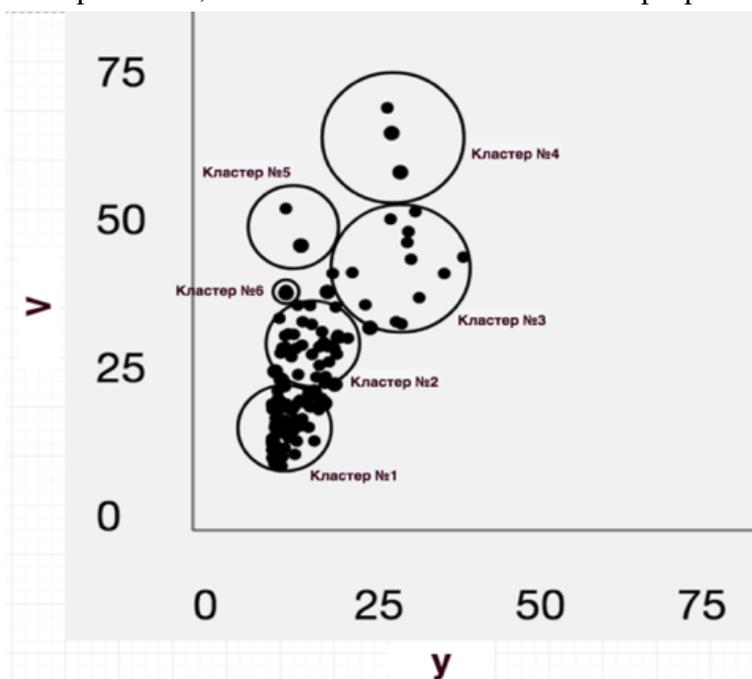


Рисунок 6. Этапы и результат повторного расчета, выполненного в соответствии с разработанным алгоритмом

Анализ рассмотренного примера и рисунков 5 и 6 указывают на наличие у предложенного алгоритма свойства самоорганизуемости и эволюционных изменений у порожденных им кластерных образований.

Также в данной главе рассмотрена модель эволюционных изменений кластерных образований и показана высокая эффективность аппарата анализа временных рядов для оценки текущих и прогнозирования будущих изменений. Состояние кластерного сфероида в любой момент времени полностью определяется координатами его центра в пространстве классифицирующих признаков, размерами и мощностью множества элементов, вошедших в его состав. Таким образом, поскольку, как было показано выше, положение кластерных центров может меняться не только вследствие корректировки, но и с течением времени, их можно рассматривать как траекторию в фазовом пространстве признаков. Вид и особенности траектории зависят от характеристик временных рядов  $x_i = x_i(t)$  и взаимовлияния классифицирующих признаков. Согласно теории временных рядов, их уровни формируются как результат совокупного действия трех составляющих: трендовой  $T$ , циклической  $S$  и случайной  $E$ . Оценка вклада каждой из них осуществляется путем построения коррелограмм, или таблиц коэффициентов автокорреляции.

На рисунке 7 представлен демонстрационный пример эволюции трех кластеров в фазовом пространстве двух признаков  $x$  и  $y$ , а в таблице 5 параметры эволюционирующих объектов.

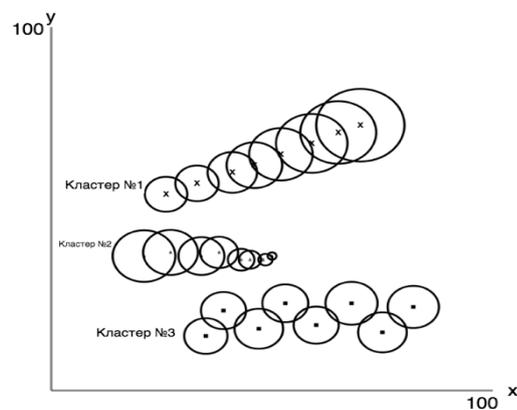


Рисунок 7. Пример расположения кластерных последовательностей

Таблица 5. Эволюционные изменения кластеров

время t	Кластер №1			Кластер №2			Кластер №3		
	Признак x	Признак y	Мощность N	Признак x	Признак y	Мощность N	Признак x	Признак y	Мощность N
1	26	54	440	21	37	310	35	15	390
2	33	57	457	24	38	292	42	26	415
3	41	60	460	26	37	267	47	17	403
4	46	62	474	27	38	231	53	31	394
5	52	65	470	29	36	206	58	15	410
6	59	68	490	30	36	181	62	29	398
7	65	71	492	32	36	168	67	16	406
8	70	73	505	35	37	153	74	28	412

Коэффициенты автокорреляции, вычисленные по формуле

$$r_z^k = \frac{\sum_{t=k+1}^n (z_t - \bar{z}_t)(z_{t-k} - \bar{z}_{t-k})}{\sqrt{\sum_{t=k+1}^n (z_t - \bar{z}_t)^2 \cdot \sum_{t=k+1}^n (z_{t-k} - \bar{z}_{t-k})^2}}, \quad (8)$$

где  $k$  – порядок коэффициента автокорреляции, приведены в таблице 6.

Таблица 6. Коэффициенты автокорреляции динамических рядов

Кластер №1				Кластер №2				Кластер №3			
$r_x^1$	$r_x^2$	$r_y^1$	$r_y^2$	$r_x^1$	$r_x^2$	$r_y^1$	$r_y^2$	$r_x^1$	$r_x^2$	$r_y^1$	$r_y^2$
0.998	0.995	0.997	0.991	0.976	0.951	0.176	0.091	0.991	0.990	-0.793	0.931

Данные таблицы 6 указывают на наличие сильных трендовых составляющих по обоим признакам, формирующих траекторию кластера №1, сильной трендовой составляющей по признаку  $x$  и слабой случайной составляющей по признаку  $y$  в кластере №2, сильной трендовой составляющей по признаку  $x$  и сильной циклической составляющей по признаку  $y$  в кластере №3.

Наличие совпадающих трендов по обоим признакам в первой кластерной последовательности требует выяснения, является ли это случайным, или результатом проявления сущностной связи между классифицирующими признаками. Для решения этой задачи используется метод отклонения от тренда, где значение коэффициента линейной корреляции рядов  $x_t$  и  $y_t$  сравнивается с аналогичным коэффициентом рядов остатков  $\Delta x_t$  и  $\Delta y_t$  регрессионных зависимостей вида

$$\hat{x}_t = a_0 + a_1 t \text{ и } \hat{y}_t = b_0 + b_1 t. \quad (9)$$

В рассматриваемом примере эти величины близки  $r_{xy} = 0.996$ , а  $r_{\Delta x \Delta y} = 0.768$  и статистически значимы. Это означает, что в ходе эволюционных изменений кластера №1 близость трендов классифицирующих признаков  $x$  и  $y$  имеет сущностный характер, и данный факт должен учитываться как при решении задач оперативного управления так и в прогнозировании.

Эволюционные изменения кластера №2 происходят под влиянием только трендовой составляющей признака  $x$ . Значительное снижение мощности кластерообразующих множеств свидетельствует о явной деградации этого кластера и делает любые прогнозные предсказания для него малодостоверными.

Эволюционное поведение кластера №3 определяется совокупным действием сильной трендовой составляющей признака  $x$  и выраженной циклической составляющей динамического ряда признака  $y$  с периодом равным двум лаговым единицам. Методом скользящего среднего производится сглаживание ряда признака  $y$ , что позволяет получить количественную оценку влияния циклической составляющей. В масштабе таблицы 5 она оказалась равной для четных лаговых единиц 6.459, а для нечетных – 6.459. Поскольку циклические изменения признака  $y$  носят устойчивый характер, необходимость их учета, как влиятельного фактора очевидна.

Результаты рассмотрения примера подтверждают эффективность аппарата анализа временных рядов для описания хода эволюционных изменений кластерных

образований и возможность его применения в процессе выработки и принятия управленческих решений.

В завершающем разделе главы выполнена сравнительная оценка эффективности метода динамической кластеризации с другими аналогичными методами классификации. Сравнение проводилось на материале обучающей выборки мощностью более 500000 единиц, при этом число классифицирующих признаков менялось от 1 до 150. Результаты эксперимента представлены на рисунке 8.

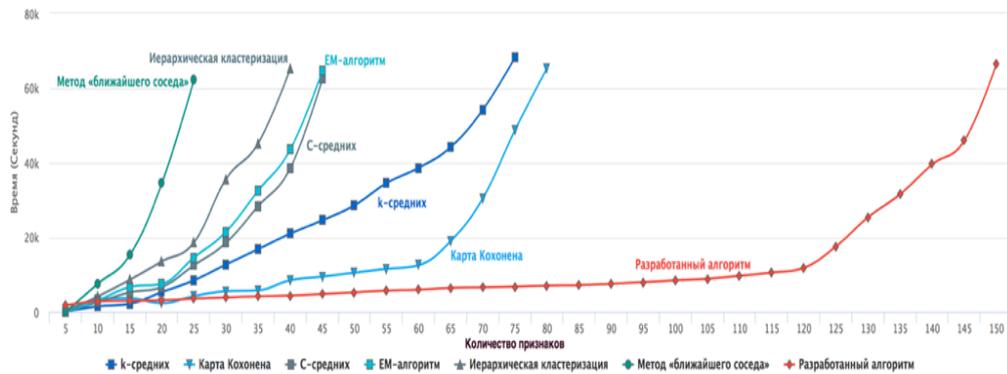


Рисунок 8. Графики зависимостей быстродействия алгоритмов кластеризации в зависимости от числа классифицирующих признаков

Анализ графиков, приведенных на этом рисунке, показывает, что для всех методов существует пороговое значение размерности пространства признаков, начиная с которого наблюдается резкий рост времени, затрачиваемого на обработку данных, что, в конечном счете, и определяет эффективность алгоритмов. Из Рисунка 8 видно, что большинство алгоритмов полностью теряют работоспособность при размерности пространства классифицирующих признаков 40-50. Метод Кохонена полностью эффективен до размерности 60 и сохраняет удовлетворительную работоспособность вплоть до 75-80. Метод же динамической кластеризации безусловно эффективен до размерности 120 и остается приемлемым до 150.

В этом же разделе приведены результаты применения алгоритма динамической кластеризации для анализа предпочтений клиентов телекоммуникационной компании, число которых изменялось от 500000 до 1000000. В качестве классифицирующих признаков (их выявилось 150), здесь выступали различные услуги и опции, предоставляемые компанией пользователям сети. С помощью алгоритма динамической кластеризации в составе клиентской базы удалось выделить девять групп пользователей, устойчивых кластерных образований (см. Рисунок 9).

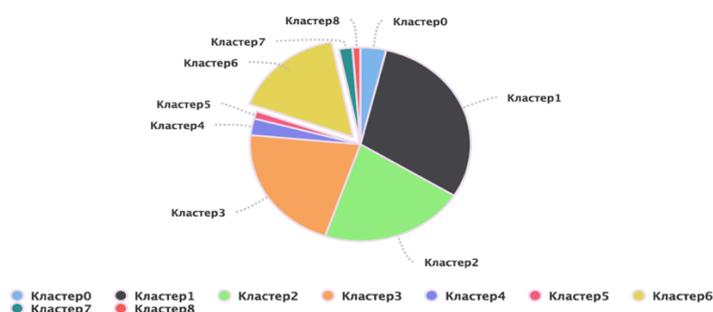


Рисунок 9. Диаграмма распределения элементов обучающей выборки по кластерам

Эти результаты были использованы администрацией компании для решения задач перспективного планирования и выработки ценовой политики.

**В третьей главе** рассматривается функциональная схема специализированного комплекса, предназначенного для решения задач классификации, и две математические модели классификации объектов: математическая модель на основе Байесовского классификатора и математическая модель построения и обучения нейронных сетей; также представлены примеры применения разработанных математических моделей и алгоритмов.

Разработанный комплекс включает в себя три независимых блока классификации: блок динамической кластеризации, Байесовский классификатор и нейронные сети. Каждый из них может работать как самостоятельно, так и в составе комплекса. Поскольку алгоритм динамической кластеризации был подробно рассмотрен в предшествующей главе, здесь будет разобрана и проиллюстрирована только работа Байесовского классификатора и нейронных сетей.

Пусть  $x$  – некоторый объект, обладающий определенным набором признаков  $X$ , а  $\{\omega_i\} i = \overline{1, n}$  множество классов, к которым этот объект может принадлежать. Обозначим  $\{Y_i\}$  – множества признаков, характеризующие соответствующие классы. Если  $X = Y_i$ , то  $x \in \omega_i$  и будем считать этот факт осуществлением события  $A$ . Однако, не всегда сразу и однозначно удастся установить точное совпадение множества  $X$  с каким-нибудь элементом  $\{Y_i\}$ . Это может происходить и по причине нечеткого задания распознаваемого объекта  $x$ , и вследствие того, что  $Y_i \cap Y_j \neq \emptyset$ , и в силу ряда прочих обстоятельств. Обозначим  $H_i$  ситуацию, когда множества  $X$  и  $Y_i$  совпадают частично, и в этом случае объект  $x$  может как принадлежать, так и не принадлежать классу  $\omega_i$ , тогда, по формуле полной вероятности

$$P(A) = \sum_{i=1}^n P(H_i) \cdot P(A|H_i), \quad (10)$$

а вероятность того, что  $x \in \omega_i$  именно потому, что его признаки частично соответствуют признакам этого класса, определится по формуле Байеса

$$P(H_i|A) = \frac{P(A|H_i) \cdot P(H_i)}{\sum_{i=1}^n P(H_i) \cdot P(A|H_i)} \quad . \quad (11)$$

Этот прием позволяет ранжировать классы по вероятности принадлежности им классифицируемого объекта и указать предпочтительную очередность проверки объектов на предмет принадлежности их тому или иному классу. Данный алгоритм особенно эффективен при обработке больших потоков запросов, поскольку, как правило, снимает необходимость полного перебора и сокращает продолжительность процедуры распознавания. На рисунке 10 представлен демонстрационный пример применения описанного алгоритма. Из диаграммы видно, что вероятность принадлежности распознаваемого объекта классу  $\omega_1$  наибольшая, а значит классификацию следует начать с  $\omega_1$  и проводить ее в последовательности  $\omega_1 \rightarrow \omega_2 \rightarrow \omega_4 \rightarrow \omega_3$ .

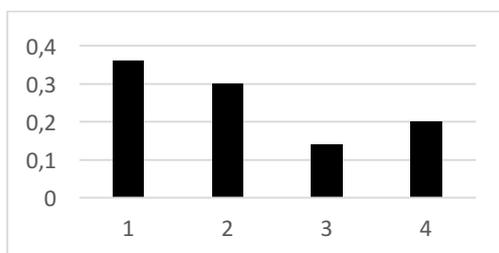


Рисунок 10. Диаграмма вероятностей принадлежности объекта  $i$  – тому классу  $\omega_i, i = \overline{1,4}$ .

В последнем разделе этой главы рассмотрен алгоритм решения задачи распознавания образов с помощью нейронных сетей. Основным принципом обучения сетей базируется на использовании метода обратного распространения (Backpropagation) ошибки, в котором по достижении последнего уровня определяется значение ошибки

$$E = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (12)$$

и, если она оказалась больше заданной степени точности, образ распознаваемого объекта возвращается на вход первого уровня и цикл повторяется заново (см. Рис. 11).

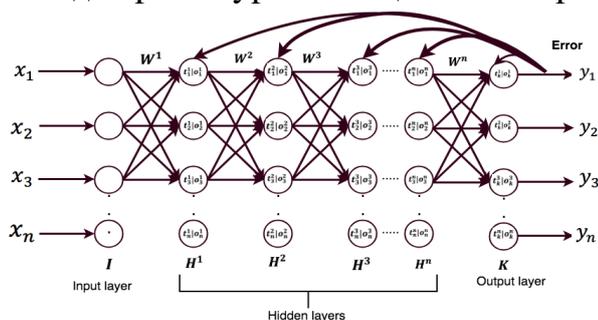


Рисунок 11. Многослойная нейронная сеть с алгоритмом обратного распространения ошибки

Главным и трудноустраняемым недостатком нейронных сетей является длительность процесса их обучения, быстро возрастающая с увеличением размерности пространства классифицирующих признаков. Поэтому, в качестве самостоятельного инструмента, в многомерных задачах распознавания образов они не применяются. Однако, они входят в состав разработанного программного комплекса и используются как вспомогательное средство, например, для дополнительного исследования кластеров малой мощности после обработки обучающей выборки с помощью алгоритма динамической кластеризации.

**Четвертая глава** посвящена описанию программных комплексов, разработанных для реализации моделей и алгоритмов классификации объектов в задачах распознавания образов, и внедрения результатов исследований на предприятиях, занятых в сфере информационных и телекоммуникационных услуг.

Особенностью функционирования всех коммуникационных систем является обработка потоков запросов от пользователей, поступающих через интернет, операторов сотовой связи и т.п. Для обеспечения высокоэффективной работы таких

систем необходимо по некоторым заранее сформулированным критериям своевременно принять решение о наилучшем способе обработки запроса и потребных для этого средствах. На решение именно таких задач ориентированы математические модели и алгоритмы, подробно описанные в предыдущих главах настоящего исследования. Те из них, что основаны на использовании элементов кластерного анализа, обладают свойством самоорганизуемости и малокритичны к размерности пространства классифицирующих признаков.

В данной главе дается подробно описание разработанного автором специализированного комплекса программ, предназначенного для реализации упомянутых выше математических моделей и алгоритмов. Для написания программ использовался язык Java и Javascript с применением объектно-ориентированного подхода в области интернета Web online. Программы классификации объектов выполнены в среде Netbeans. Результаты тестирования программного комплекса на конкретных числовых примерах представлены в предшествующих главах.

Разработанный программный комплекс внедрен и успешно используется в компаниях, активно работающих на рынке телекоммуникационных и информационных услуг. Это система коммуникационных услуг, задействованная в государстве Йемен, и система сбора и анализа данных из сети интернета компании “Дельтаинком” (г. Казань, Россия). В этих системах наглядно проявилась эффективность разработанных математических моделей и алгоритмов, как средства обработки большого объема данных в многомерном пространстве признаков. Размеры и структуру массивов данных, на которых работает программный комплекс, иллюстрирует диаграмма на рисунке 12, где каждый столбец имеет смысл мощности множества задач отдельного кластера, обрабатываемых коммуникационной системы AlsharabyExchange в течении месяца.



Рисунок 12. Диаграмма количества задач в 4-ых коммуникационных кластерах в течении месяца

**В заключении** приведены основные результаты и выводы, имеющие научную и практическую ценность.

1. Проведена экспертиза методов, используемых для решения задач распознавания образов, по результатам которой определены области их эффективного использования.
2. Для решения задачи распознавания образов разработана новая математическая модель самоорганизующейся системы кластеров.

3. Показано, что кластеры являются динамическими образованиями, изменяющимися под воздействием потока распознаваемых образов.
4. Разработан и апробирован алгоритм динамической кластеризации, действующий на заданном множестве характеризующих признаков определенной предметной области.
5. Заложены основы эволюционного формирования кластерной модели, позволяющей использовать ее для решения задач эффективного управления и прогнозирования.
6. Разработаны математическая модель и алгоритм вычислительной процедуры, которые позволяют классифицировать объекты в некоторой предметной области, по их свойствам (признакам), при условии, что часть признаков разных классов совпадают.
7. Для практической реализации алгоритмов разработан специализированный программный комплекс.

### **Рекомендации и перспективы дальнейшей разработки темы**

Перспективы дальнейшей разработки темы автор видит в применении разработанных методов кластеризации для решения задач распознавания образов в режиме реального времени.

### **Основные результаты диссертации изложены в следующих публикациях**

В изданиях из списка ВАК РФ:

- 1 Печеный Е.А. Математическая модель динамической кластеризации в задачах распознавания образов / Е.А. Печеный, А.А. Аль Хашеди, Н.К. Нуриев // Современные наукоемкие технологии. №5, 2018, С. 124-130.
- 2 Аль Хашеди А.А. Экспертиза методов, используемых в различных задачах распознавания образов / А.А. Аль Хашеди, Е.А. Печеный, Н.К. Нуриев // Вестник технологического университета. 2018. №5. С. 130-135.
- 3 Аль Хашеди А.А. Разработка математической модели распознавания запросов-задач коммуникационных услуг / А.А. Аль Хашеди, А.А. Обади, Н.К. Нуриев, Е.А. Печеный // Фундаментальные исследования. – 2017. – № 6. – С. 9-14.
- 4 Нуриев Н.К. Математическое моделирование эволюции кластерных образований / Н.К. Нуриев, А.А. Аль-Хашеди, Е.А. Печеный // Современные наукоемкие технологии. №8, 2018, С. 121-126.
- 5 Обади А.А. Проектирование математической модели и модуля распознавания образов для смарт-обучающей системы / А.А. Обади, А. А. Аль Хашеди, Н.К. Нуриев, Е.А. Печеный // Вестник технологического университета. 2017. №8. С. 95–100.
- 6 Аль Хашеди А.А. Разработка математического и программного обеспечения задач распознавания образов на основе перцептрона / А.А. Аль Хашеди, А.А. Обади, Н.К. Нуриев // Вестник технологического университета. ун-та. 2017. №11. С. 85–89.

- 7 Обади А.А. Моделирование процесса распознавания образов для smart-образовательных систем / А.А. Обади, А.А. Аль Хашеди, Н.К. Нуриев, Е.А. Печеный // *Фундаментальные исследования*. – 2017. – № 11-2. – С. 320-324.
- 8 Обади А.А. Проектирование программного обеспечения smart-образовательной системы на примере решения алгебраических и трансцендентных уравнений / А.А. Обади, А.А. Аль Хашеди, Ф.А. Мурshed // *Вестник технологического университета*. 2016. №10. С.121-124.
- 9 Аль Хашеди А.А. Проектирование программного обеспечения умной образовательной системы на примере обучения решения систем линейных уравнений / А.А. Аль Хашеди, А.А. Обади // *Вестник технологического университета*. 2017. №1. С.102-105.
- 10 Мурshed Ф.А. Использование имитационного моделирования для администрирования систем массового обслуживания / Ф.А. Мурshed, А.А. Обади, А.А. Аль Хашеди // *Вестник технологического университета*. 2017. Т.20, №1. С.125-127.

В других изданиях:

- 11 Обади А.А. Разработка smart-системы для распознавания алгебраических и трансцендентных уравнений и их решений / А.А. Обади, А.А. Аль Хашеди // *«Электронная Казань 2016»*. – Казань: ЮНИВЕРСУМ, 2016. С. 435-441.
- 12 Аль Хашеди А.А. Проектирование программного обеспечения smart-обучающей системы на примере распознавания задач в коммуникационных услугах / А.А. Аль Хашеди, А.А. Обади // *«Электронная Казань 2017»*. – Казань: ЮНИВЕРСУМ, 2017. С. 30-36.
- 13 Аль Хашеди А.А. Проектирование программного обеспечения smart-обучающей системы на примере распознавания правильности решения задач интерполяции и экстраполяции / А.А. Аль Хашеди, А.А. Обади // *«Электронная Казань 2017»*. – Казань: ЮНИВЕРСУМ, 2017. С. 422-428.
- 14 Аль Хашеди А.А. Разработка математической модели и программного обеспечения построения и обучения нейронных сетей для классификации объектов в задачах распознавания образов / А.А. Аль Хашеди, А.А. Обади // *«Электронная Казань 2018»*. – Казань: ЮНИВЕРСУМ, 2018. С. 41-51.
- 15 Обади А.А. Разработка математической модели автоматизированной классификации объектов из определенной предметной области / А.А. Обади, А.А. Аль Хашеди // *«Электронная Казань 2018»*. – Казань: ЮНИВЕРСУМ, 2018. С. 379-385.
- 16 Аль Хашеди А.А. Анализ эффективности алгоритмов классификации на больших данных / А.А. Аль Хашеди, Е.А. Печеный, Н.К. Нуриев // *«Наука и Просвещение»*. – Пенза: МЦНС, 2018. С. 130-134.