На правах рукописи

Киреев Василий Сергеевич

МЕТОДЫ ДВУХЭТАПНОЙ И МНОГОКРИТЕРИАЛЬНОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ ВЫБОРОК БОЛЬШИХ ОБЪЕМОВ

05 13 01 - системный анализ, управление и обработка информации
(научное обслуживание)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Автор 

Москва 2008

Работа выполнена в Московском инженерно-физическом институте (государственном университете)

Научный руководитель	кандидат технических наук, доцент Синицын Сергей Владимирович
Официальные оппоненты	доктор физико-математических наук, профессор Крянев Александр Витальевич Московский инженерно-физический институт (государственный университет), кафедра «Прикладная математика»
	доктор технических наук, Емец Евгений Павлович Госкорпорация «РОСАТОМ»
Ведущая организация	Московский государственный университет экономики, статистики и информатики

Защита диссертации состоится 4 июня 2008 года в 15 часов 00 минут на заседании диссертационного совета Д 212 130 03 в Московском инженерно-физическом институте (государственном университете) по адресу 115409, г Москва, Каширское ш , 31, тел (495) 323-95-26, 324-84-98, ауд 408, главный корпус

С диссертацией можно ознакомиться в библиотеке института

Автореферат разослан 30 апреля 2008 года

Ученый секретарь
диссертационного совета — Шумилов Ю Ю

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Решение задачи кластеризации, то есть разбиения исходной совокупности объектов на группы со схожими в смысле какого-либо критерия свойствами, является актуальным для многих приложений, где возникает проблема анализа большого объема информации – в экологических, медицинских, социологических, экономических, в т ч маркетинговых исследованиях и т д Кластеризация позволяет среди всей совокупности объектов и их свойств уловить определенные закономерности и тенденции Разработка простых и быстрых методов кластеризации, не зависящих от параметров, значения которых редко можно знать априорно, имеет особую актуальность при решении практических задач в области социальных и экономических приложений, когда точность полученных кластерных решений имеет решающее значение

Задача кластеризации или таксономии впервые была рассмотрена в 1930-х годах Эту проблему в ее различных аспектах изучали как зарубежные так и отечественные исследователи, в том числе МакКуин Д , Ланс У , Уильямс Д , Хартиган Д , Вонг М , Кохонен Т , Фрицке Б , и Колмогоров А Н , Загоруйко Н Г , Елкина В Н , Айвазян С А , Мхитарян В С , Шумский С А , и другие Проблема кластеризации данных обычно рассматривается в двух различных вариантах постановки – нахождения естественного расслоения кластеров и нахождения кластеров в виде групп близких объектов Первый вариант может и не иметь решения, например, в случае, если исходные данные представляют собой один большой кластер, однако второй вариант имеет решение всегда, и представляет наибольший интерес для исследователей

Кластерные процедуры для нахождения схожих объектов разделяют на два основных типа – агломеративные (дивизимные) и итеративные Агломеративные процедуры основаны на пошаговом объединении пары ближайших кластеров (наблюдений) и пересчете матрицы расстояний В итеративных процедурах на каждом шаге работы рассматривается только один объект, и производится его отнесение к одному из кластеров, пока не будет получено устойчивое разбиение Пересчет матрицы расстояний приводит к тому, что вычислительная сложность иерархических процедур резко возрастает при увеличении объема выборки Итеративные методы сильно зависят от выбора начального разбиения, что приводит к необходимости повторного решения задачи с новыми условиями Недостатки рассмотренных подходов не позволяют применять их как универсальные, круг их применения ограничивается данными сравнительно небольших объемов ($\sim 10 - 10^2$ объектов) при априори известной информации о кластерной структуре

Вышесказанное определяет актуальность настоящей работы, связанной с разработкой математического, алгоритмического и программного обеспечения методов обработки больших объемов информации(~ 10^3 объектов), определением критериев и моделей описания кластерных разбиений

Целью диссертационной работы является решение важной научной задачи, заключающейся в систематизации известных методов кластерного анализа и разработке новых, предназначенных для достижения точности решения на больших объемах информации, их теоретическое исследование, экспериментальное обоснование и анализ эффективности на основе статистических исследований

Методы исследования

При разработке математического аппарата в диссертационной работе используются методы теории математической статистики, методы дискретной оптимизации и методы многокритериальной оптимизации При разработке программного обеспечения используются методы объектно-ориентированного программирования

Научная новизна работы заключается в следующем

1 Создана модель унифицированного формального описания наиболее известных методов кластеризации для проведения их сравнительного анализа С целью определения областей эффективного использования различных методов определена их вычислительная сложность Было показано, что эффективность традиционных иерархических методов решения задачи таксономии резко снижается при увеличении объема исходных данных, а итеративные методы не могут обеспечить качество решения для выборок больших объемов по нескольким критериям

2 Впервые предложен метод «карманной» кластеризации для решения задачи таксономии на выборках большого объема и построения разбиения с возможностью выбора оптимального числа кластеров Оценка предложенного метода на предмет вычислительной сложности показала, что метод «карманной» кластеризации решает задачу таксономии на выборках большого объема за субквадратичное время, в отличие от иерархических методов, характеризующихся кубической сложностью Использование в методе двухэтапной процедуры позволяет получать стабильные кластерные решения независимо от условий проведения первого этапа

3 Впервые предложен метод Q-кластеризации, основанный на решении задачи дискретной многокритериальной оптимизации, позволяющий построить оптимальное разбиение с учетом нескольких критериев качества Исследование свойств кластерных решений, получаемых с помощью метода Q-кластеризации,

проведенное на сгенерированных тестовых примерах показало, что примерно в 70% случаев найденные решения соответствуют истинной структуре кластеров объектов в пространстве признаков В частности, метод адекватно определяет истинное число кластеров

4 Разработано математическое и алгоритмическое обеспечение для реализации предложенных методов, что позволило исследовать их свойства и границы применимости, а также провести анализ точности на искусственных тестовых и практических примерах Результаты проведенного исследования показали, что предложенные методы могут успешно применяться для решения практических задач, связанных с исследованием особенностей пространственного распределения объектов, заданного массивами большой размерности, в условиях отсутствия априорной информации

Практическая значимость состоит в следующем:

1 Выполнена программная реализации метода «карманной» кластеризации, которая использовалась для решения задачи сегментации потребителей банковских услуг на данных исследования, проведенного в 2003 году Международным Агентством Социальных и Маркетинговых Исследований (МАСМИ) В дальнейшем это программное обеспечение было внедрено в маркетинговый процесс торговой организации ООО «Мегагранд XXI Век», что подтверждается соответствующим актом о внедрении

2 Выполнена программная реализации метода Q-кластеризации, которая использовалась для решения задачи многофакторного анализа и выделения сегмента учителей-новаторов в рамках выполнения Национального проекта «Образование» в 2006г по программе «Совершенствование системы повышения квалификации и профессиональной переподготовки педагогических, инженерно-технических кадров общеобразовательных школ в области информационных и коммуникационных технологий (ИКТ) и смежных областей» В этом случае была собрана и обработана информация по использованию информационных и коммуникационных технологий (ИКТ) в профессиональной деятельности среди 5500 участников программы из семи федеральных округов, прошедших в 2006г повышение квалификации В результате полученных кластерных решений был сформирован пул учителей-инноваторов для дальнейшего использования их методического опыта в области ИКТ, что подтверждено соответствующими актами о внедрении.

Эксплуатация указанных программных реализаций показала, что все они обладают большой практической значимостью и могут быть рекомендованы к дальнейшему использованию Предложенные автором методы кластерного анализа несколько лет используются в лабораторных практикумах по курсам «Теория вероятностей и математическая статистика» кафедры «Кибернетика» и «Маркетинг и маркетинговые исследования» кафедры «Экономика и

управление» МИФИ, что подтверждается соответствующим актом о внедрении Результаты данной работы вошли в проект разработки информационно-образовательного портала МИФИ для самостоятельной работы студентов, выполняемый по Инновационной программе инженерно-физического образования для нового этапа развития ядерной науки и промышленности в рамках реализации Приоритетного национального проекта «Образование» С помощью предложенных методов осуществляется статистическая обработка результатов обучения и определяется рейтинг студентов

На защиту выносятся:

1 Новый метод «карманной» кластеризации для решения задачи таксономии на выборках большого объема, в котором успешно комбинируются характеристики точности иерархических схем и простота реализации итерационных процедур кластеризации

2 Новый метод Q-кластеризации для решения задачи построения оптимального разбиения с учетом нескольких критериев качества, обеспечивающий соответствие кластерных решений истинной структуре распределения объектов в пространстве признаков

3 Математическое, алгоритмическое и программное обеспечение нового метода «карманной» кластеризации, примененное для аналитического исследования социально-демографического и психологического статуса потребителей банковских услуг с целью более эффективной работы банков на финансовом рынке

4 Математическое, алгоритмическое и программное обеспечение нового метода Q-кластеризации, примененное в ряде образовательных проектов в целях совершенствования образовательного процесса

5 Содержательные результаты кластеризации данных, полученные в ходе использования программного обеспечения метода Q-кластеризации при реализации программы «Совершенствование системы повышения квалификации и профессиональной переподготовки педагогических, инженерно-технических кадров общеобразовательных школ в области информационных и коммуникационных технологий (ИКТ) и смежных областей» в рамках Национального проекта «Образование» в 2006г

6 Содержательные результаты, полученные в ходе использования программного обеспечения метода «карманной» кластеризации для решения задачи сегментации потребителей банковских услуг на данных исследования, проведенного Международным Агентством Социальных и Маркетинговых Исследований в 2003г

Достоверность разработанного математического, алгоритмического и программного обеспечения методов «карманной» и Q-кластеризации

подтверждается проведенными в работе экспериментальными и теоретическими исследованиями, сравнительным анализом результатов кластеризации разными методами, соответствующими актами о внедрении, представлением основных результатов диссертации на международных конференциях и выставках

Реализация и внедрение результатов работы

Научные результаты, полученные в диссертационной работе в виде методов и алгоритмов кластеризации данных выборок больших объемов и их программных реализаций для конечных пользователей, были использованы и внедрены·

- в ООО «Мегагранд XXI век» в маркетинговый процесс для обработки данных исследований потребительского рынка,
- в ЗАО «Академия АйТи» в рамках работы по Федеральной целевой программе развития образования на 2006-2010 годы при проведении многофакторного анализа компетенций в области информационно-коммуникативных технологий (ИКТ),
- в проекте разработки информационно-образовательного портала МИФИ для самостоятельной работы студентов, выполняемый по Инновационной программе инженерно-физического образования для нового этапа развития ядерной науки и промышленности в рамках реализации Приоритетного национального проекта «Образование»

Апробация работы

Основные результаты диссертационного исследования докладывались и обсуждались на международных и всероссийских конференциях и семинарах, в том числе

- Научные сессии МИФИ 2005-2008,
- XIV-XVI Международные научно-технические семинары «Современные технологии в задачах управления, автоматизации и обработки информации» (г. Алушта, 2005-2007 гг)

Структура работы

Диссертационная работа состоит из введения, четырех глав, заключения, трех приложений, списка использованной литературы и содержит 66 рисунков, 16 таблиц Общий объем без приложений 128 с (вместе с приложениями – 148 с)

Публикации

По материалам диссертации опубликованы 12 печатных работ общим объемом 1,5 печатных листа, в том числе статья в журнале из перечня изданий,

рекомендованных ВАК для опубликования основных результатов диссертационных работ

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность диссертационной работы и приводится ее краткая характеристика Формулируются цели работы и задачи исследования, представляются основные положения, выносимые на защиту

В первой главе дан обзор имеющейся литературы по теме исследования и рассмотрены научно-методологические и технические аспекты исследуемой проблемы Рассмотрены различные постановки задачи кластеризации, освещены проблемы определения близости наблюдений и приведены основные способы измерения расстояния между кластерами

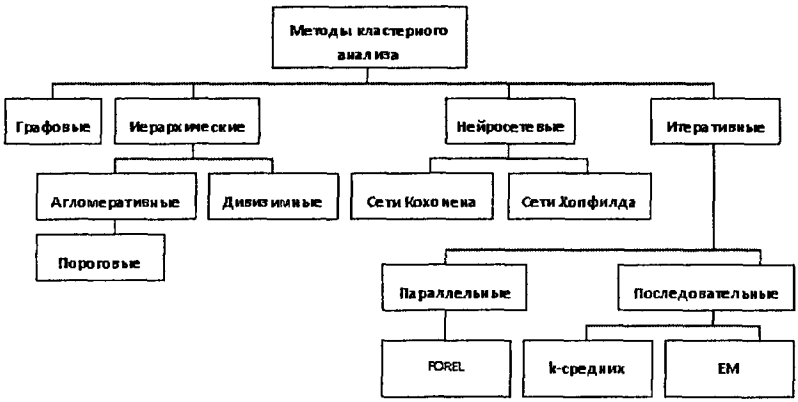

Рис 1 Классификация методов кластерного анализа

Для проведения сравнительного анализа наиболее известных методов кластеризации была создана модель их унифицированного формального описания С целью определения областей эффективного использования различных методов определена их вычислительная сложность В ходе исследования предметной области были выделены основные подходы к решению задач кластерного анализа иерархический, итеративный, нейросетевой и основанный на представлении выборки ориентированным графом (см рис. 1)

Проведенный на основе модели унифицированного описания сравнительный анализ позволил определить условия и границы применимости указанных подходов В частности, были выявлены виды задачи кластеризации, практически не решаемые в рамках традиционных подходов, а именно 1) построение дерева таксономии для выборок больших объемов, 2) построение

кластерного решения, оптимального по нескольким критериям качества, при отсутствии априорной информации об истинном числе кластеров

Полученные в данной работе результаты сравнительного анализа позволили сформулировать формальную постановку указанных выше задач, с учетом следующих определений

1 имеется выборка объектов объема N, характеризуемых M признаками ($M << N$) $o_i = (o_i^1, o_i^2, \ldots, o_i^M)^T$, где $i = \overline{1,N}$, признаки измеряются в числовой шкале $o_i^j \in R$, где $j = \overline{1,M}$ (либо в порядковой шкале),

2 под разбиением C подразумевается множество из k кластеров c_p, где $p = \overline{1,k}$, т е $C = \{c_1, c_2, \ldots, c_k\}$, S – множество возможных разбиений,

3 под кластером c_p подразумевается множество из N_p объектов $\sum_{p=1}^{k} N_p = N$, $\rho(o_x, o_y) < \rho(o, o_z)$, где $o = o_x, o_y$, а $o_x, o_y \in c_p$, при $o_z \notin c_p$,

4 под $\rho(o_x, o_y)$ подразумевается расстояние (мера близости) между объектами o_x и o_y.

Задача **построения таксономического дерева** формулируется так в указанных обозначениях по данным исходной выборки ($N \sim 10^3$) построить систему вложенных разбиений $C_1, C_2, \ldots, C_{N-1}$, где $C_j = [C_{j-1} \cup \{c_p^* \cup c_q^*\}] / \{c_p^*, c_q^*\}$, при $\rho(c_p^*, c_q^*) = \min \rho(c_p, c_q), \forall c_p, c_q \in C_{j-1}$, $C_1 = \{o_1, \ldots, o_N\}$

Задача нахождения **кластерного решения, (локально) оптимального по нескольким критериям** представляется в виде в указанных обозначениях по данным исходной выборки ($N \sim 10^3$) найти разбиение C^*, оптимизирующее значения заданных критериев $F_j(C^*) = \mathrm{opt} F_j(C), \forall C \in S$, где $j = \overline{1,l}$

Во второй главе для решения задачи построения таксономического дерева по выборке данных большого объема был предложен метод «карманной» кластеризации Этот метод основан на двухэтапной схеме, включающей этап *снижения размерности исходной задачи* и этап *собственно кластеризации* с использованием агломеративной иерархической процедуры с использованием метода Варда в качестве метода измерения расстояния между кластерами

Этап снижения размерности исходной задачи опирается на разбиение исходной выборки данных на l выборок одинакового меньшего размера (выборки извлекаются без возвращения) Каждая из этих выборок кластеризуется с помощью метода k-средних при заданном одинаковом значении k На выходе этапа имеются центры тяжести кластеров и $N - [N/l]$

наблюдений, не вошедших в рассматриваемые l выборок, и возникающих, если объем выборки N не делится на l нацело.

На втором этапе полученные центры тяжести кластеров подаются на вход иерархической агломеративной процедуры для построения таксономического дерева Дальнейшие действия зависят от требуемого результата так для получения конкретного решения обычно используется эмпирическое правило оптимального числа кластеров – по номеру шага r скачка расстояния агломерации – расстояния, на котором объединяются два ближайших кластера

Таким образом, алгоритм предложенного метода можно представить в виде *следующей схемы*

1 Первый этап снижение размерности исходной задачи

1 1 Разбиение исходной выборки на l выборок объема $[N/l]$,

1 2 Цикл по всем l выборкам,

1 2 1 Кластеризация выборки $\iota = \overline{1,l}$ методом k-средних;

2 Второй этап построение таксономического дерева по полученным кластерам и наблюдениям

2 1 Иерархическое агломеративное объединение $k \times l$ кластеров и $N - [N/l]$ наблюдений (расстояние d между кластерами c_p и c_q определяется по методу Варда)

$$d(c_p, c_q) = \frac{N_p N_q}{N_p + N_q} \rho^2(\mu_p, \mu_q) \qquad (1),$$

где N_p – объем кластера c_p, μ_p – центр тяжести кластера c_p;

2 2 Определение оптимального числа кластеров как $k^* = n - r$, где число r соответствует номеру шага иерархической процедуры, предшествующего скачку расстояния агломерации

Для изучения свойств метода «карманной» кластеризации было проведено специальное исследование, с применением как специально сгенерированных модельных примеров, так и классического набора данных – ирисов Фишера Так, с применением статистического метода *однофакторного* анализа была определена вычислительная сложность «карманной кластеризации» как субквадратичная

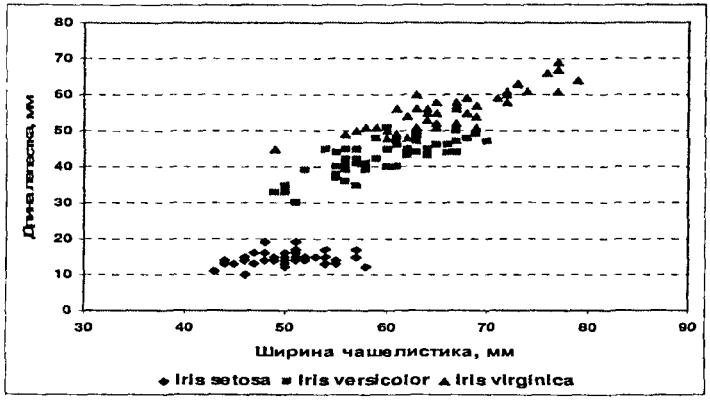

Рис 2 Диаграмма рассеяния для ирисов Фишера в выбранных координатах

На примере ирисов Фишера (см рис 2) было проведено сравнение «карманной» кластеризации с традиционным итеративным методом k-средних МакКуина Полученные результаты свидетельствуют о большем числе успешно классифицированных объектов (при известной структуре кластеров) в случае применения метода, предложенного автором

Для решения задачи многокритериальной оптимизации был предложен метод Q-кластеризации, основанный на итеративной оптимизации вектора значений критериев качества Постановка задачи рассматривалась с учетом двух критериев - среднего межкластерного расстояния F_1 (найти локальный максимум) и среднего внутрикластерного расстояния F_2 (найти локальный минимум)

$$F_1(C)=\frac{2}{K(K-1)}\sum_{p=1}^{k-1}\sum_{q=p+1}^{k}\left[\frac{1}{N_p N_q}\sum_{i=1}^{N_p}\sum_{j=1}^{N_q}\rho(o_i,o_j)\right] \quad (2),$$

$$F_2(C)=\frac{1}{k}\sum_{l=1}^{k}\left[\frac{2}{N_l(N_l-1)}\sum_{i=1}^{N_l-1}\sum_{j=i+1}^{N_l}\rho(o_i,o_j)\right] \quad (3),$$

где K – число кластеров, N_p – объем p-ого кластера, $\rho(o_i,o_j)$ – расстояние между наблюдениями i и j

На каждой итерации осуществляется последовательное разбиение исходной выборки данных, так чтобы значения обоих критериев не ухудшались

к каждым новым шагом Цель каждой новой итерации состоит в улучшении значений критериев, полученных на предыдущей итерации, до тех пор, пока не будут получены стабильные значения (в смысле заданной точности) Результирующее число и состав кластеров считается окончательным В качестве меры, определяющей точность решения, предлагается использовать функционал следующего вида

$$Q_j = \left(F^{ideal}(C_j) - F_2(C_j)\right)\left(F_1(C_j) - F^{ideal}(C_j)\right) \qquad (4)$$

С учетом сказанного выше, алгоритм предложенного метода можно представить в виде следующей общей схемы

1 Инициализация параметров метода
   1 1 Разбить исходную выборку на два кластера – c_1 объема $N-1$, и c_2 объема 1
   1 2 Рассчитать начальные значения критериев $(F_1(0),\ F_2(0))$, $t=0$
2 Основной цикл по j, пока разница в решениях не станет меньше заданного порога, т е $\left|Q_j - Q_{j-1}\right| < \delta$
   2 1 Цикл по всем $o_t \in c_1, t = \overline{1, N-1}$ наблюдениям,
      2 1 1 Рассчитать значения $\left(F_1^1(t),\ F_2^1(t)\right)$, если наблюдение o_t присоединяется к ближайшему кластеру c_j, где $j = \overline{1,k}$,
      2 1 2 Рассчитать значения $\left(F_1^2(t),\ F_2^2(t)\right)$, если наблюдение o_t выделяется в новый кластер c_{k+1},
      2 1 3 С помощью метода смещенного идеала выбрать наилучший вариант разбиения по значениям $\left(F_1^1(t),\ F_2^1(t)\right)$, $\left(F_1^2(t),\ F_2^2(t)\right)$ и $\left(F_1(t-1),\ F_2(t-1)\right)$ При выборе текущего «идеала» учитывать рассчитанный на предыдущем шаге $\left(F_1^{ideal}(t-1),\ F_2^{ideal}(t-1)\right)$
      2 1 4 Осуществить перенос наблюдения o_t в соответствии с выбранным вариантом разбиения

На основе сгенерированных данных была исследована зависимость метода Q-кластеризации от выбора начального разбиения. Результирующие разбиения сравнивались с решениями метода k-средних по значениям рассмотренных критериев качества Метод Q-кластеризации оказался более эффективным, т к значения критериев были ближе к оптимальным, чем в случае метода k-средних

Третья глава посвящена решению практической задачи – сегментированию рынка банковских услуг с помощью метода «карманной» кластеризации. Данные о потребителях банковских услуг были получены в ходе соответствующего маркетингового исследования, проведённого компанией МАСМИ (Международного Агентства Маркетинговых и Социальных исследований) в 2003 году.

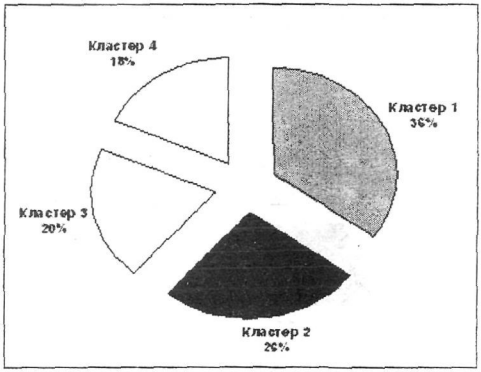

Рис. 3 Распределение респондентов по кластерам

В рамках исследования было опрошено 2000 респондентов из семи федеральных округов. Для сегментации респондентов были выбраны психографические переменные (степень согласия респондентов с высказываниями о поведенческой реакции). Ввиду большого числа переменных были применен метод главных факторов для сокращения их числа.

Преобразованные таким образом данные были успешно кластеризованы с помощью метода «карманной» кластеризации. В результате было получено четыре кластера респондентов, для которых были построены профили по социально-демографическим переменным и которые были проинтерпретированы с точки зрения средних значений главных факторов.

На основании полученных результатов были сформулированы управленческие рекомендации к дальнейшей стратегии действий банков-участников рассмотренного рынка банковских услуг.

Четвёртая глава посвящена решению практической задачи – исследованию профессиональных компетенций в области ИКТ в рамках Национального проекта «Образование» в 2006 г. по программе «Совершенствование системы повышения квалификации и профессиональной переподготовки педагогических, инженерно-технических кадров общеобразовательных школ в области информационных и коммуникационных технологий (ИКТ) и смежных областей».

В ходе исследования была собрана и обработана информация по применению ИКТ в профессиональной деятельности среди 5500 участников программы из семи федеральных округов, прошедших в 2006г. повышение квалификации. На основании полученных данных был проведён многофакторный анализ с использованием метода Q-кластеризации. Предварительно, с целью сокращения размерности исходной задачи методом главных факторов было сокращено число переменных кластеризации.

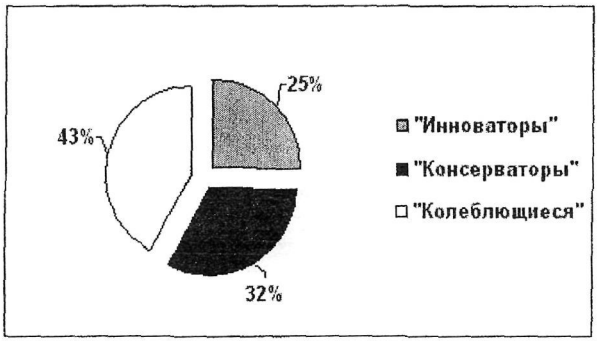

Рис. 4 Распределение респондентов по кластерам

На основании полученных результатов среди респондентов был выделен сегмент «инноваторов» (см. рис. 4), способных к применению ИКТ в профессиональной деятельности. Для контроля качества полученных решений была произведена кластеризация методом k-средних, для трёх кластеров, а также разбиение методом Q-кластеризации для половин исходной выборки с последующим сравнением с исходным результатом (см. таб. 1).

Таблица 1 Сравнение кластерных решений по значениям критериев

	Критерий F_1	**Критерий F_2**
Метод Q-кластеризации		
Исходное решение	5,036	0,187
Объединённое решение	5,031	0,192
Метод k-средних		
Исходное решение	4,089	0,203

По результатам исследований, связанных с накопленным опытом учителей-инноваторов, использования ИКТ в учебном процессе были разработаны: рекомендации по использования электронных образовательных ресурсов в типовом учебном заведении; рекомендации по использованию интернет-порталов (как вертикальных, так и горизонтальных); рекомендации по

использованию хранилищ учебных и методических материалов, рекомендации по использованию программно-технических средств поддержки учебного процесса

По результатам данной программы также были сформированы рекомендации по методическому, программному и технологическому оснащению типовых образовательных учреждений системы общего образования, а также их включении в единую образовательно-информационную систему, основанных на полученном позитивном опыте профессионального и педагогически осознанного применения ИКТ в системе российского образования

В заключении отражены основные результаты, полученные в диссертационной работе

В приложениях содержатся элементы анкет, использованных при проведении практических исследований, выходные формы программных реализаций предложенных методов, акты о внедрении

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Среди основных результатов работы можно выделить следующие

1 Проведено детальное исследование проблемной области кластеризации данных Для проведения сравнительного анализа традиционных методов решения задач кластеризации созданы формальные алгоритмические описания наиболее используемых методов кластеризации по унифицированному образцу, с целью выявления областей эффективного использования определена вычислительная сложность этих методов, что позволило определить границы применимости традиционных методов

2 Предложен двухэтапный метод «карманной» кластеризации для решения задачи таксономии на выборках большого объема и построения разбиения с возможностью выбора оптимального числа кластеров Оценка вычислительной сложности показала, что метод «карманной» кластеризации решает задачу таксономии на выборках большого объема за субквадратичное время Использование в методе двухэтапной схемы позволяет получать стабильные кластерные решения независимо от условий проведения первого этапа

3 Реализованный с помощью средств VBA 6 0 для Microsoft Excel метод «карманной» кластеризации использовался для решения задачи сегментации потребителей банковских услуг на данных исследования, проведенного в 2003 году Международным Агентством Социальных и Маркетинговых Исследований (МАСМИ) Разработанное программное обеспечение было внедрено в маркетинговый процесс торговой организации ООО «Мегагранд XXI Век», что подтверждается соответствующим актом о внедрении

4 Предложен метод Q-кластеризации для решения задачи построения оптимального разбиения с учетом нескольких критериев качества разбиения Исследование свойств кластерных решений, получаемых с помощью метода Q-кластеризации, проведенное на сгенерированных тестовых примерах показало, что, в среднем в 70% случаев найденные решения соответствуют истинной структуре данных В частности, адекватно определяется истинное число кластеров

5 Реализованный с помощью средств VBA 6 0 для Microsoft Excel метод Q - кластеризации использовался для решения задачи многофакторного анализа и выделения сегмента учителей-новаторов в рамках выполнения Национального проекта «Образование» в 2006 г по программе «Совершенствование системы повышения квалификации и профессиональной переподготовки педагогических, инженерно-технических кадров общеобразовательных школ в области информационных и коммуникационных технологий (ИКТ) и смежных областей» В результате полученных кластерных решений был сформирован пул учителей-инноваторов для дальнейшего использования их методического опыта в области ИКТ, что подтверждено соответствующими актами о внедрении

6 Предложенный метод Q - кластеризации использован в проекте разработки информационно-образовательного портала МИФИ для самостоятельной работы студентов, выполняемый по Инновационной программе инженерно-физического образования для нового этапа развития ядерной науки и промышленности в рамках реализации Приоритетного национального проекта «Образование» (2007)

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1 Киреев В С Алгоритм кластеризации данных с минимаксной оптимизацией критериев качества разбиения /Киреев В С //Информационные технологии – 2007, № 7 – С 47-49

2 Киреев В С Сегментация потребителей банковских услуг с помощью метода Q-кластеризации /Киреев В С , Синицын С В // Информационная математика – 2007, № 1(6) – С 81-90

3 Киреев В С Оптимальность кластерных решений, получаемых методом «карманной кластеризации» /Киреев В С , Синицын С В // XVI Международный научно-технический семинар «Современные технологии в задачах управления, автоматики и обработки информации»· сб научных трудов – Алушта, 2007 – С 58

4 Киреев В С Метод усечения матрицы расстояний в решении задачи кластерного анализа /Киреев В С// XV Международный научно-технический семинар «Современные технологии в задачах управления, автоматики и обработки информации» сб научных трудов – Алушта, 2006 –С 73

5 Киреев В С Объединенные итеративный и агломеративный подходы в процедуре кластеризации с априорно неизвестным числом кластеров /Киреев В С, Синицын С В // XIV Международный научно-технический семинар «Современные технологии в задачах управления, автоматики и обработки информации» сб научных трудов – Алушта, 2005 – С 50

6 Киреев В С Информационная поддержка самостоятельной работы студентов система МИФИСТ /Гусева А И, Киреев В С, Тихомирова А Н, Филиппов С А, Цыплаков А С // Научная сессия МИФИ-2008 сб научных трудов – М МИФИ, 2008 –Том 6 – С 21-22

7 Киреев В С Разработка информационно-образовательного портала для поддержки самостоятельной работы студентов /Гусева А И, Киреев В С, Тихомирова А Н, Филиппов С А, Цыплаков А С // Научная сессия МИФИ-2008 XII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании» сб научных трудов – М МИФИ, 2008 – С 13-14

8 Киреев В С «МИФИСТ» информационно-образовательный портал для поддержки самостоятельной работы студентов /Гусева А И, Киреев В С, Маслий, Н П, Тихомирова А Н, Филиппов С А, Цыплаков А С // Научная сессия МИФИ-2008 XII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании» Каталог – М МИФИ, 2008 – С 14

9 Киреев В С Q-алгоритм сегментирования данных при неизвестном исходном числе сегментов /Киреев В С // Научная сессия МИФИ-2007 сб научных трудов – М МИФИ, 2007 –Том 2 –С 93-95

10 Киреев В С Двухэтапный алгоритм кластеризации данных /Киреев В С, Синицын С В // Научная сессия МИФИ-2006 сб научных трудов – М МИФИ, 2006 –Том 2 –С.14-15

11 Киреев В С Нейросетевая модель потребителя в маркетинговом ценовом исследовании ВРТО/Киреев В С, Синицын С В // Научная сессия МИФИ-2005 сб научных трудов – М. МИФИ, 2005 –Том 2 –С 142

12 Киреев В С Маркетинг и маркетинговые исследования/Киреев В С, Маковеев Н П //Электронные учебные материалы –М, МИФИ, 2007