Владимирский государственный университет

Методические указания к практическим занятиям по дисциплине «Современные проблемы прикладной математики и информатики»

Владимир 2015

Министерство образования и науки РФ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»
Кафедра физики и прикладной математики

Методические указания к практическим занятиям по дисциплине «Современные проблемы прикладной математики и информатики»

Составители:
А.Ю. Лексин
С.В. Рощин
С.И. Абрахин
А.В.Духанов

Владимир 2015

Практическое занятие №1

АЛГОРИТМЫ КЛАССИФИКАЦИИ И РЕГРЕССИИ

I. Цель занятия: закрепление знаний об алгоритмах классификации и регрессии, рассмотренных в рамках лекционного курса, а также освоение ряда новых алгоритмов.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с группой методов и алгоритмов глубинного анализа данных (Data Mining), предназначенных для решения задач классификации и регрессии и не рассмотренных в рамках лекционных занятий. К таковым относятся:

- алгоритм C4.5;
- алгоритм покрытия;
- линейные методы построения математических функций классификации и регрессии, в частности, метод наименьших квадратов;
- нелинейные методы построения математических функций;
- метод SVM (Support Vector Machines);
- метод карт Кохонена.

Алгоритм C4.5 является развитием алгоритма классификации ID3, решая присущую последнему проблему сверхчувствительности (overfitting) за счёт введения некоторой нормализации характеристик информации.

Алгоритм покрытия является альтернативным подходу «разделяй и властвуй», используемому в алгоритмах ID3 и C4.5 и заключается в построении деревьев решений для каждого класса по отдельности.

Линейные и нелинейные методы построения математических функций используются тогда, когда классифицируемые данные являются не категориальными, а числовыми. Такая «специализация» этих методов позволяет получить при наличии подобных данных более качественные результаты по сравнению с алгоритмами ID3, C4.5, алгоритмом покрытия.

Метод SVM основан на теории распознавания образов и предположении о том, что наилучшим способом разделения точек в m-мерном про-

странстве является m-1 плоскость (заданная некоторой функцией), равноудалённая от точек, принадлежащих разным классам.

Метод карт Кохонена предназначен для решения задач, в которых данные трудно представимы в математической числовой форме.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты с описанием принципов и примеров применения одного из перечисленных выше алгоритмов (или группы алгоритмов), остальные студенты должны ознакомиться с этими алгоритмами, используя литературные источники.

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, достоинств и недостатков рассмотренных методов, сравнительный анализ методов.

Практическое занятие №2

АЛГОРИТМЫ ПОИСКА АССОЦИАТИВНЫХ ПРАВИЛ

I. Цель занятия: закрепление знаний об алгоритмах поиска ассоциативных правил, рассмотренных в рамках лекционного курса, а также освоение ряда новых алгоритмов.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с группой методов и алгоритмов глубинного анализа данных (Data Mining), предназначенных для решения задач поиска ассоциативных правил и не рассмотренных в рамках лекционных занятий. К таковым относятся:

- алгоритм AprioriTid;
- алгоритм MSAP.

Алгоритм AprioriTid является разновидностью алгоритма Apriori. Отличительной чертой данного алгоритма является подсчет значения поддержки кандидатов не при сканировании множества транзакций, а с по-

мощью множества, являющегося множеством кандидатов (*k*-элементных наборов) потенциально частых, в соответствие которым ставится идентификатор TID транзакций, в которых они содержатся.

Другой разновидностью алгоритма Apriori является алгоритм *MSAP* (Mining Sequential Alarm Patterns), специально разработанный для выполнения сиквенциального анализа сбоев телекоммуникационной сети.

Он использует следующее свойство поддержки последовательностей: для любой последовательности L_k ее поддержка будет меньше, чем поддержка последовательностей из множества L_{k-1}.

Алгоритм MSAP для поиска событий, следующих друг за другом, использует понятие «срочного окна» (Urgent Window). Это позволяет выявлять не просто одинаковые последовательности событий, а следующие друг за другом. В остальном данный алгоритм работает по тому же принципу, что и Apriori.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты с описанием принципов и примеров применения одного из перечисленных выше алгоритмов, остальные студенты должны ознакомиться с этими алгоритмами, используя литературные источники.

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, достоинств и недостатков рассмотренных методов, сравнительный анализ методов.

Практическое занятие №3

КЛАСТЕРИЗАЦИЯ ПО ГЮСТАФСОНУ-КЕССЕЛЮ

I. Цель занятия: закрепление знаний об алгоритмах кластеризации, рассмотренных в рамках лекционного курса, а также освоение алгоритма нечёткой кластеризации по Гюстафсону-Кесселю.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с алгоритмом нечеткой кластеризации по Гюстафсону-Кесселю.

В ходе подготовки к занятию студенты должны программно реализовать данный алгоритм, а также алгоритм Fuzzy C-Means, после чего выполнить их сравнительный анализ.

Один из студентов группы (по желанию или по указанию преподавателя) готовит реферат и устное сообщение с описанием алгоритма кластеризации по Гюстафсону-Кесселю.

Данный алгоритм нечеткой кластеризации ищет кластеры в форме эллипсоидов (рис. 1), что делает его более гибким при решении различных задач по сравнению с алгоритмами *k*-means и Fuzzy C-Means.

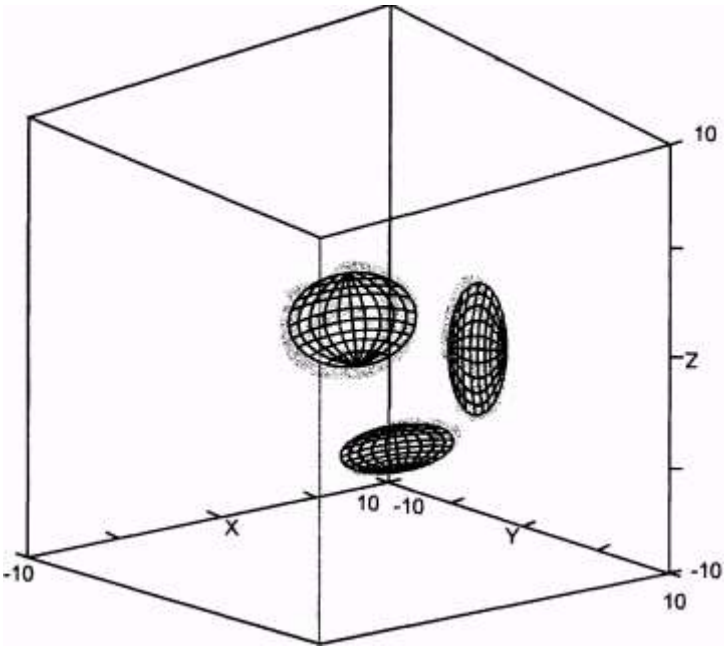

Рис. 1. Форма кластера в алгоритме кластеризации
по Гюстафсону-Кесселю

Необходимо, однако, понимать, что алгоритму Гюстафсона-Кесселя присущи те же недостатки, что и для других неиерархических алгоритмов:

- допущение о том, что все кластеры всегда имеют некоторую, определяемую алгоритмом форму, а это, очевидно, далеко не всегда выполняется. Аппроксимация пространства входных данных некоторыми задан-

ными фигурами на данных, имеющих сложное взаимное расположение, может привести к неинтерпретируемым результатам;

- допущение о том, что в кластере всегда есть некоторая узловая точка (центр кластера), степень принадлежности которой кластеру равна единице, в то время как остальные точки (не равные центру кластера) не могут принадлежать кластеру с такой же высокой степенью принадлежности, что, опять же, при сложном взаимном расположении точек данных является неприемлемым;
- данные алгоритмы строятся не на основе взаимного расположения точек, а лишь на отношении точек к центрам кластеров.

III. Ход занятия.

В ходе занятия студент, подготовивший реферат, делает устное сообщение. По результатам сообщения проходит дискуссия с обсуждением возникших неясных моментов, достоинств и недостатков рассмотренного алгоритма, его сравнительный анализ с другими алгоритмами, иллюстрируемый результатами, полученными студентами в ходе самостоятельной работы при подготовке к занятию.

Практическое занятие №4

КЛАСТЕРИЗАЦИЯ ПРИ ПОМОЩИ НЕЧЁТКИХ ОТНОШЕНИЙ

I. Цель занятия: закрепление знаний об алгоритмах кластеризации, рассмотренных в рамках лекционного курса, а также освоение ряда новых алгоритмов.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с группой алгоритмов Data Mining, предназначенных для решения задач кластеризации данных и основанных на использовании математического аппарата нечётких отношений. При этом отдельно следует рассмотреть ряд перечисленных ниже вопросов.

Свойства нечетких бинарных отношений применительно к анализу данных. Важным отличием нечетких отношений от классических тео-

ретико-множественных отношений является определение характеристической функции $\mu_X(x)$ множеств. В случае классической теории множеств характеристическая функция принимает одно из двух допустимых дискретных значений, смысл которых в идентификации принадлежности/непринадлеж-ности данного элемента универсального множества множеству *X*. Теория нечетких множеств дает обобщение характеристической функции множества, которая носит название функции принадлежности. Для понимания принципов кластеризации с использованием этого понятия следует определить смысл ряда классических бинарных отношений, а также ввести некоторые новые.

Принципы сравнения данных. Следует ввести ряд базовых понятий и мер, необходимых для выполнения процедуры сравнения данных, подвергаемых кластеризации. Такими понятиями являются: образец данных, мера сходства по расстоянию, нормальная мера сходства, отношение α-толерантности.

Отношение α-квазиэквивалентности. В классической теории множеств при помощи отношения толерантности можно построить классы эквивалентности на множестве образцов данных *X* и построить покрытие этого множества. Для этого необходимо найти совокупности транзитивно зависимых образцов данных на отношении толерантности. Для случая нечетких отношений определить свойство транзитивности, в силу недискретной характеристической функции отношения, можно различными способами, которые бы учитывали желаемую степень нечеткости проявления свойства транзитивности. Одним из ключевых понятий при этом становится понятие α-квазиэквивалентности.

Построение и использование шкалы α-квазиэквивалентности для анализа данных. Данный вопрос, собственно, и касается непосредственно алгоритма кластеризации на основе нечетких отношений. Его рассмотрение подразумевает также анализ конкретных примеров использования данного метода.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит устные сообщения по одному из перечисленных выше вопросов (или группы вопросов), остальные студенты должны ознакомиться с этими вопросами, используя литературные источники.

III. Ход занятия.

В ходе занятия студенты делают устные сообщения. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, сравнительный анализ с другими известными им алгоритмами кластеризации. В конце занятия подводится итог изучения всего раздела курса, посвященного глубинному анализу данных.

Практическое занятие №5

ИНТЕГРАЦИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ. ТЕХНОЛОГИЯ SEMANTIC WEB

I. Цель занятия: систематизация и закрепление знаний о группе технологий Semantic Web, предназначенных для глобальной интеграции информационных ресурсов.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство со стеком технологий Semantic Web (SW), разработанных в рамках инициативы консорциума World Wide Web по созданию «интеллектуального Интернета». Подлежат рассмотрению следующие технологии:

- URI (Universal Resource Identifier) – универсальные идентификаторы, определяющие способ записи адреса произвольного ресурса. В Semantic Web каждое понятие определяется через URI-идентификатор, что позволяет любому определить новое понятие, просто разместив в нужном месте его URI;
- язык разметки XML (eXtensible Markup Language) – синтаксическая основа Semantic Web. Этот язык позволяет создавать собственные произвольные структуры данных;
- язык описания ресурсов RDF (Resource Description Framework) и его расширение RDF Schema. Позволяют на основе XML описать различные ресурсы, а также создавать каталоги и словари понятий;

- языки описания онтологий DAML (DARPA Agent Markup Language) и OWL (Ontology Web Language). Предоставляют средства для создания онтологий. Обеспечивают более полную автоматическую обработку содержимого Сети, чем та, которую поддерживают XML и RDF;
- агенты – множество программ, предназначенных для работы в Semantic Web. Агенты, знакомясь с содержимым Сети из различных источников, обрабатывают полученную информацию и обмениваются ею с другими программами;
- отдельным вопросом является математическая основа функционирования Semantic Web – алгоритмы логического вывода и доказательства, построенные на принципах описательной логики.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты с описанием принципов и примеров применения указанных выше элементов SW, остальные студенты должны ознакомиться с технологиями SW, используя литературные источники и источники на сайте World Wide Web Consortium (www.w3c.org).

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, проблем и перспектив внедрения технологий «интеллектуального Интернета».

Практическое занятие №6

ПРОБЛЕМЫ ИНФОРМАЦИОННОГО ПОИСКА

I. Цель занятия: систематизация и закрепление знаний о математических и алгоритмических основах построения поисковых систем, о существующих в этой области задачах и перспективах развития.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с основами работы поисковых систем, а также обзор акту-

альных на текущий момент задач в данной области. К таковым можно отнести:

- разработку алгоритмов поиска изображений и музыки по содержанию;
- создание алгоритмов кластеризации и классификации изображений;
- развитие математико-алгоритмических методов борьбы с незапрашиваемой корреспонденцией (спамом);
- создание алгоритмов автоматической классификации веб-ресурсов;
- совершенствование алгоритмов и методов автоматического аннотирования и реферирования;
- разработка алгоритмов анализа предпочтений пользователей и рекомендательных систем;
- создание лгоритмов выявления дубликатов документов и «зеркал» сайтов;
- разработка алгоритмов и методов жанровой классификации веб-текстов;
- Развитие методов, подходов и алгоритмов автоматического построения тезаурусов и онтологий;
- задачи и алгоритмы корпусной лингвистики.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты, в которых раскрывается состояние научных исследований и практических разработок, касающихся одного из перечисленных выше направлений, остальные студенты должны также предварительно кратко ознакомиться с данными вопросами.

Помимо перечисленных возможных тем рефератов, один из студентов должен подготовить реферат, содержащий обзор ресурсов, научных групп, изданий и форумов (конференций, семинаров, научных школ и т.п.), посвящённых технологии Semantic Web и информационному поиску.

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с

обсуждением возникших неясных моментов, проблем и перспектив развития рассмотренных направлений исследований, их сравнительный анализ.

Практическое занятие №7

ЗАДАЧИ БОЛЬШОЙ ВЫЧИСЛИТЕЛЬНОЙ ЕМКОСТИ

I. Цель занятия: получение знаний о важном направлении исследований в области прикладной математики и информатики – алгоритмах и методах решения задач большой вычислительной ёмкости.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с математическими и алгоритмическими основами решения задач, требующих существенных вычислительных затрат, применения аппаратных и программных технологий параллельных вычислений, метакомпьютинга. К таковым относятся:

- мониторинг и предсказание погоды;
- мониторинг и предсказание землетрясений;
- прогнозирование чрезвычайных ситуаций, обусловленных причинами природного и техногенного характера;
- задачи взлома шифров;
- поиск внеземных цивилизаций с помощью распределенной обработки данных, поступающих с радиотелескопа;
- расшифровка генома животных и человека;
- анализ информации для целей астрономии и космологии;
- исследования в области ядерной физики;
- решение сложных технических и экономических задач.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты с описанием постановки и состояния исследований по одной из перечисленных задач с акцентом на вклад прикладной математики и информационных технологий, остальные студенты должны ознакомиться с данными вопросами, используя литературные источники.

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, перспектив решения рассмотренных задач, возможной в связи с ними эволюции математики в целом и прикладной математики в частности, а также информационных технологий.

Практическое занятие №8

ПРОБЛЕМЫ БЕЗОПАСНОСТИ В ИНФОРМАЦИОННОМ ОБЩЕСТВЕ

I. Цель занятия: закрепление знаний о методах и алгоритмах защиты информации в условиях глобального проникновения информационных технологий во все сферы общественных отношений.

II. Подготовка к занятию.

Подготовка к занятию предполагает предварительное самостоятельное знакомство с технологиями и алгоритмами защиты информации, применяемыми в локальных и глобальных компьютерных сетях, при организации электронного документооборота, при размещении электронных публикаций.

Информационная безопасность является одной из проблем, с которой столкнулось современное общество в процессе массового использования автоматизированных средств ее обработки.

Проблема информационной безопасности обусловлена возрастающей ролью информации в общественной жизни. Современное общество все более приобретает черты информационного общества.

С понятием «информационная безопасность» в различных контекстах связаны различные определения. Так, в Законе РФ «Об участии в международном информационном обмене» информационная безопасность определяется как состояние защищенности информационной среды общества, обеспечивающее ее формирование, использование и развитие в интересах

граждан, организаций, государства. Подобное же определение дается и в Доктрине информационной безопасности Российской Федерации, где указывается, что информационная безопасность характеризует состояние защищенности национальных интересов в информационной сфере, определяемых совокупностью сбалансированных интересов личности, общества и государства.

Оба эти определения рассматривают информационная безопасность в национальных масштабах и поэтому имеют очень широкое понятие.

Наряду с этим характерно, что применительно к различным сферам деятельности так или иначе связанным с информацией понятие «информационная безопасность» принимает более конкретные очертания. Так, например, в «Концепции информационной безопасности сетей связи общего пользования Российской Федерации» даны два определения этого понятия.

Информационная безопасность – это свойство сетей связи общего пользования противостоять возможности реализации нарушителем угрозы информационной безопасности.

Информационная безопасность – свойство сетей связи общего пользования сохранять неизменными характеристики информационной безопасности в условиях возможных воздействий нарушителя.

Необходимо иметь в виду, что при рассмотрении проблемы информационной безопасности нарушитель необязательно является злоумышленником. Нарушителем информационной безопасности может быть сотрудник, нарушивший режим информационной безопасности или внешняя среда, например, высокая температура, может привести к сбоям в работе технических средств хранения информации и т. д.

Всевозможные аспекты, связанные с информационной безопасностью, и являются предметом рассмотрения в рамках данного практического занятия.

Ряд студентов группы (по желанию или в соответствии с указанием преподавателя) готовит рефераты с описанием математических и алгоритмических принципов и примеров решения актуальных задач обеспечения информационной безопасности, остальные студенты должны ознакомиться с этими вопросами, используя литературные источники.

III. Ход занятия.

В ходе занятия студенты, подготовившие рефераты, делают устное сообщение. По результатам каждого из сообщений проходит дискуссия с обсуждением возникших неясных моментов, достоинств и недостатков рассмотренных методов, сравнительный анализ методов. Обязательным вопросом, требующим обсуждения, является взаимное влияние информационных технологий и системы общественных отношений в современной цивилизации, перспективы развития в связи с этим человеческого общества.

ОГЛАВЛЕНИЕ