

Государственный комитет по высшему образованию
Московский физико-технический институт

УТВЕРЖДАЮ

Проректор по учебной работе

_____ Т. В. Кондранин
" ___ " _____ 200_ г.

Факультет управления и прикладной математики
Кафедра интеллектуальных систем

ПРОГРАММА

по курсу: Анализ данных в метрических пространствах

по направлению 511656

курс 6

семестр 11

лекции 34 часа

практические (семинарские)

Экзамен 11 семестр

занятия нет

лабораторные занятия нет

Программу составил: к.ф.-м.н. А. И. Майсурадзе

Программа обсуждена на заседании кафедры 16 мая 2007 г.

Программа обсуждена и одобрена на методической комиссии факультета
" ___ " _____ 200_ г.

Председатель методической комиссии ФУПМ

чл.-корр.РАН

Ю. А. Флеров

Аннотация. В курсе дисциплины представлена теоретическая основа для построения, реализации и анализа широкого спектра методов, применяющихся в интеллектуальном анализе данных (ИАД, data mining) и базирующихся на понятиях сходства, близости, аналогии. Идея сходства свойственна человеческому мышлению, это породило целый комплекс подходов для всех фундаментальных задач ИАД, среди которых внимание в курсе уделено классификации, восстановлению регрессии, кластеризации, восстановлению пропущенных данных. Рассмотрены методы построения и вычисления функций сходства, согласование сходства на различных множествах объектов, синтез новых способов сравнения объектов на базе уже имеющихся. Рассмотрен комплекс технологий, предназначенный для эффективного представления и обработки метрической информации вычислительными системами. Изложение материала дополняется примерами из прикладных областей, демонстрирующими возникновение рассматриваемых идей или их использование.

Основные подходы к заданию сходства. Функциональный подход: двуместные функции, удовлетворяющие аксиомам. Геометрический подход: определение в пространстве множеств точек. Табличный подход: матрицы попарного сходства над конечными множествами.

Классическое определение метрики и метрического пространства. Аксиоматическое задание метрики. Построение топологии по метрике. Пространства сходящихся последовательностей. Фундаментальные последовательности и полные пространства. Роль аксиомы треугольника и непрерывность метрики. Роль аксиомы сепарабельности и единственность предела сходящейся последовательности. Сопоставление метрик и отношений эквивалентности, $0,1$ -метрики. Различные модификации системы аксиом метрики и их интерпретация: расстояние, полуметрика, ультра-метрика, квази-метрика, неравенство Птолемея.

Локальные метрики и их продолжение на всё пространство. Формализация понятия «между» в метрическом пространстве. Выпуклость метрического пространства по Менгеру. Аксиомы существования и единственности точек между заданными точками. Аксиомы существования и единственности продолжения луча. Теорема о единственности продолжения локально совпадающих метрик. Практический пример проверки аксиом и использования локального продолжения метрики.

Геометрические подмножества общих метрических пространств. Понятия открытого и замкнутого шара, их согласованность с топологией метрического пространства. Понятия открытого и замкнутого обобщенного эллипсоида. Клетки Дирихле («сферы влияния»), автоматическое исправление ошибок. Геометрическое место точек, равноудаленных от заданных точек, проблема меры указанного подмножества. Понятие кривой в метрическом пространстве, длина кривой. Геодезическая линия, кривая наименьшей длины, сегмент. Свойство совпадения геодезических с множествами равноудаленных точек в обобщенных евклидовых пространствах.

Примеры метрических пространств. Пространство изолированных точек, дискретная топология. Метрики l_1 (городских кварталов), l_2 (евклидова), l_∞ (Чебышёва). Их

физический смысл. Метрика l_p (Минковского). Форма шаров, вложенность единичных шаров. Зависимость объема шара от размерности пространства. Проблема сопоставления объема шаров в разных метриках с ростом размерности. Проблема единственности кратчайшего пути. Хаусдорфова метрика и другие метрики между подмножествами метрического пространства, индуцированные исходной метрикой между точками. Расстояния между функциями (графиками). Метрики на декартовом произведении метрических пространств. Случай конечного и бесконечного числа сомножителей, метрики на последовательностях.

Классификация функций сходимости. Сопоставление значений: номинальные, порядковые, арифметические (интервальные, относительные, разностные, абсолютные) шкалы. Понятие о граничных объектах. Аксиомы сходимости, главный и вспомогательный аргументы. Классификация мер сходимости по одному свойству (признаку). Функции сходимости на декартовом произведении пространств со значениями в различных шкалах.

Характеристики метрик. Инвариантность расстояния относительно сдвига. Инвариантность расстояния относительно поворота. Инвариантность формы шаров относительно положения центра и направления на центр. Инвариантность объема шаров относительно положения центра и направления на центр. Ограниченность метрики. Ограниченность шаров. Понятие полностью абсолютных и полностью относительных метрик, промежуточные метрики. Выпуклость шаров. Односвязность шаров. Существование и единственность сегментов, непрерывность сегментов.

Преобразования метрик. Изометрические преобразования пространств. Преобразования функций, сохраняющие метрические свойства. Некоторые достаточные условия преобразований, сохраняющих метрические свойства. Ограничение значений метрики (range componders). Примеры универсальных компандеров. Возможность монотонного преобразования произвольной функции в метрику. Возможность линейного преобразования произвольной ограниченной функции в метрику. Нормализация (релятивизация) метрик, зависимость от точки отсчета. Примеры для пространств множеств и евклидовых пространств. Переход от булеанов конечных множеств к пространствам бинарных векторов, соответствие мощности множества и длины вектора.

Реализация метрик. Реализация конечных метрик точками ЛВП, точечные конфигурации. Алгоритмическая сложность решения задачи точного вложения в линейные пространства с метриками l_1 , l_2 , l_{∞} . Примеры МК, имеющих или не имеющих точную реализацию. Задача поиска оптимальной точечной конфигурации в пространстве малой размерности, методы метрического и неметрического многомерного шкалирования. Реализация многомерных данных элементами функциональных пространств. Методы визуализации многомерных данных: параллельные координатные оси, графики Эндрюса, шкалирование и иерархии, таблицы проекций, параметризованные глифы (звезды, лица Чернова).

Принцип самоорганизации при построении эвристических информационных моделей. Понятие представителей, мера сходимости между объектами и представителями. Функции представительства и назначений, структура метода. Самоорганизация в задаче кластеризации. Самоорганизация и задача факторного анализа, самоорганизация и задача дискриминантного анализа. Модификация прецедентной информации, понятие типологического дискриминантного анализа. Самоорганизация и задача восстановления пропусков.

Метрики на конечных множествах. Традиционное представление метрик таблицами попарных расстояний. Понятие метрической конфигурации (МК). Специальное линейное

пространство метрических конфигураций. Система неравенств треугольника как определение полиэдрального конуса полуметрик. Грани и экстремальные лучи полуметрического конуса, проблема их определения. Векторное представление метрических конфигураций. Достаточные условия сохранения метрических свойств покомпонентными корректорами метрических конфигураций. Примеры использования достаточных условий. Несовместимость метрических свойств и ортогональности метрических конфигураций.

Разложение МК по конечным системам МК. Полные системы, базисы МК. Проблема использования переполненных систем МК. Гомогенные базисы, интерпретация коэффициентов разложения. Ранг МК. Ранговые и полуметрические ранговые базисы. Неполные системы, оптимальная аппроксимация МК. Разложение по системе «отдельных объектов», метрика попарных сумм, эффективное вычисление признака «общая удаленность» для индивидуальных объектов.

Основная литература

1. Воронин Ю.А. Начала теории сходства. Новосибирск: Наука, СО, 1991.
2. Воронин Ю.А. Теория классификации и её приложения. Новосибирск: Наука, СО, 1985.
3. Деза М., Лоран М. Геометрия разрезов и метрик. М.: МЦНМО, 2001.
4. Скворцов В.А. Примеры метрических пространств. М.: МЦНМО, 2002.
5. Майсурадзе А.И. Гомогенные и ранговые базисы в пространствах метрических конфигураций // Ж. вычисл. матем. и матем. физ. (ЖВМиМФ). 2006. Т.46. №2. С.344-361.
6. Майсурадзе А.И. О поиске оптимального коллективного слагаемого для набора метрических конфигураций // Искусственный интеллект (ИИ). 2006. №2. С.183-187.
7. Майсурадзе А.И. О свойствах оптимальных точечных конфигураций для одного семейства функционалов сравнения метрических конфигураций // Ж. вычисл. матем. и матем. физ. (ЖВМиМФ). 2005. Т.45. №9. С.1741-1748.
8. Майсурадзе А.И. Об оптимальных разложениях конечных метрических конфигураций в задачах распознавания образов // Ж. вычисл. матем. и матем. физ. (ЖВМиМФ). 2004. Т.44. №9. С.1697-1707.

Дополнительная литература

9. Буземан Г. Геометрия геодезических. М.: Физматгиз. 1962.
10. Дидэ Э. Методы анализа данных. М.: Финансы и статистика, 1985.
11. Дэйвисон М. Многомерное шкалирование. М.: Финансы и статистика, 1988.
12. Кочетков Д.В. О функциях близости. Сообщения по прикл. матем. ВЦ АН СССР. 1978.
13. Кочетков Д.В. Построение алгоритма вычисления расстояний для одного класса метрических пространств. Сообщения по прикл. матем. ВЦ АН СССР. 1978.
14. Тылкин М.Е. О геометрии Хэмминга единичных кубов. // Доклады АН СССР, 1960, Т.134, С.1037-1040.
15. Тылкин М.Е. О реализуемости матриц расстояний в единичных кубах. // Проблемы кибернетики, 1962, Т.7, С.31-42.
16. Шрейдер Ю.А. Что такое расстояние? М.: Физматгиз. 1963.
17. Basalaj W. Proximity Visualization of Abstract Data. Dissertation work. 2001.
18. Yianilos P.N. Normalized Forms for Two Common Metrics. Princeton: NEC Research Institute. 2002.