## Практикум по добыче данных

*Дмитрий Игнатов*

Обучение современным методам анализа данных невозможно без применения специализированных программных средств и выполнения практических заданий, но возможно ли создание такого лабораторного практикума без использования коммерческого ПО? Образовательный проект кафедры анализа данных и искусственного интеллекта ГУ-ВШЭ призван ответить на этот вопрос.

Специалист, не только владеющий теоретическими знаниями, но и умеющий решать различные задачи анализа с помощью специализированного программного обеспечения, более востребован, поэтому курс, разработанный на кафедре анализа данных и искусственного интеллекта, прежде всего ориентирован на практическое применение современных методов разработки (добычи) данных на реальных массивах, причем с помощью свободно распространяемых программных инструментов. Во внимание были приняты многочисленные просьбы со стороны коллег и студентов факультета бизнес-информатики и отделения прикладной математики, и в курс были включены лекции и практические занятия по современным аналитическим пакетам.

Практикум можно пополнять; например, в будущем в него планируется включить лабораторные работы, связанные со специальными видами данных: категоризация текстов, анализ графовых структур и т.п. Часть лабораторных работ практикума подготовлена научно-учебной группой «Рекомендательные интернет-сервисы и интеллектуальный анализ данных» факультета бизнес-информатики ГУ-ВШЭ.

## Методика обучения

Квалифицированный аналитик должен уметь самостоятельно провести необходимую работу с данными, определить тип задачи (классификация, кластеризация, прогнозирование, поиск зависимостей и т.п.), решить ее адекватно выбранным методом с оптимально определенными параметрами, оценить результаты, сделать содержательные выводы и интерпретировать. Кроме обучения таких специалистов практикум должен способствовать формированию культуры оформления аналитических отчетов и освоению поискового и проблемно-ориентированного подхода к решению задач анализа данных.

Студент изучает необходимый теоретический минимум, изложенный в описании работы, отвечает на вопросы для проверки готовности к выполнению лабораторной работы, получает данные, использует программное обеспечение, выбирает подходящую модель и метод, пытается решить задачу. Результаты работы метода могут быть как удовлетворительными, например метод успешно решает задачу прогнозирования для 92% тестовой выборки, или нет, например когда количество правильных предсказаний низко – 28%. Возникает вопрос, почему задача не решена. Причиной низкого качества предсказаний могут быть: неправильная спецификация модели, шумы и ошибки в данных, неадекватный выбор метода анализа данных и/или его параметров, некорректный способ оценки качества предсказаний и т.п. Как и в случае с научными гипотезами, необходимо подвергать сомнению правильность действий аналитика на каждом из этапов работы и предлагать шаги по улучшению схемы анализа данных. Принципы, лежащие в основе научных гипотез, как нельзя лучше согласуются с понятием схемы анализа данных: проверяемость, максимальная общность, предсказательная сила и простота.

Немаловажным аспектом обучения анализу данных является формирование умения интерпретировать полученные результаты, например объяснять причинно-следственные

связи на основе найденных закономерностей (поиск ассоциативных правил). Следует также отметить дифференцированный характер такого подхода к обучению, так как студент в рамках лабораторной работы решает задачу индивидуально, отвечая на вопросы преподавателя по конкретной теме работы.

При таком построении курса устраняется разрыв между знанием теории метода и его использованием на реальных данных. От преподавателя требуется контролировать выполнение студентами лабораторных работ практикума, проверять знания студентов после изучения материала теоретического минимума, проверять итоговые отчеты, консультировать студента. Для полноценного проведения практикума преподаватель должен быть знаком с применяемыми программными системами и владеть математическими моделями и алгоритмами, лежащими в основе методов анализа данных этого курса.

Предполагаемое количество часов курса рассчитывается исходя из выбранного для проведения числа лабораторных работ. Примерно 2-4 академических часа отводится на выполнение одной лабораторной работы и столько же на защиту всех работ. Оптимальное количество студентов в компьютерном классе – 15-20 человек на одного преподавателя.

В учебном плане бакалавриата четвертого курса на 2010/11 учебный год отделения прикладной математики и информатики курс называется «Системы разработки данных и машинного обучения», на него отводится 22 лекционных часа и 24 часа практических занятий, а в качестве форм контроля указана одна контрольная работа и зачет по итогам практикума.

Перед выполнением лабораторной работы студент отвечает на вопросы и выполняет задания для допуска к практикуму (простые модельные расчеты, производимые вручную). Здесь оценивается уровень понимания студентом выбранной модели или метода, правильность сделанных вручную расчетов для учебного примера. После выполнения работы оценивается соблюдение формальных требований к отчету, правильность выполнения работы (обработка данных, спецификация модели, оценка качества результатов и т.п.), верность и значимость выводов, приемлемость предлагаемой интерпретации результатов. Далее преподаватель проверяет знания студентов по материалам предоставленных ими отчетов с учетом замечаний и ошибок, выявленных ранее.

Студенты получают в качестве задания одну из списка лабораторных работ, текст этой работы в электронном виде или на бумажном носителе. Далее, следуя инструкции по выполнению лабораторной работы, студент отвечает на вопросы теоретического минимума и для предложенного набора данных проводит исследование по шагам, фиксируя результаты в электронной форме отчета.

## Содержание курса и программные системы анализа данных

Лабораторные работы проводятся по следующим темам:

1. Исследование объектно-признаковых данных с помощью программных средств анализа формальных понятий.
2. Поиск ассоциативных правил и частых (замкнутых) множеств признаков.
3. Деревья решений.
4. Задачи кластеризации:
   4.1. иерархическая кластеризация;
   4.2. метод k-средних;
   4.3. спектральная кластеризация.
5. Неточные множества (Rough Sets).
6. ДСМ-метод в системе QuDA.
7. Наивная байесовская классификация (Naï□ve Bayes Classifier).
8. Методы OLAP.

С одной стороны, все это наиболее востребованные на практике методы, а с другой – среди

них есть алгебраические методы, которые успели завоевать популярность в научных кругах для решения задач разработки данных и машинного обучения, но еще не так хорошо известны рядовым аналитикам.

Вопросы для допуска к лабораторной работе могут включать дополнительные задания в виде модельных учебных расчетов, выполняемых вручную для наборов данных размерами семь-десять объектов на пять-шесть признаков для различных предметных областей (выдача кредита, предсказание угона автомобиля, определение съедобности грибов, выбор партнера для знакомства и т.п.). Такой подход позволяет привлечь и сконцентрировать внимание учащегося на сути метода и разобрать его работу в подробностях.

В качестве инструментов исследования предполагается использовать свободное ПО для добычи данных (data mining) и машинного обучения. Действительно, использование только промышленного программного обеспечения не позволяет сделать курс доступным для изучения в течение одного-двух учебных модулей – этому препятствует сложность установки и настройки программного обеспечения (Microsoft SQL Server, Oracle Data Miner и т.д.). Кроме того, сложность промышленных технологий для обработки больших объемов данных может скрыть суть изучаемых методов анализа данных. Коммерческие аналитические пакеты часто содержат излишнюю функциональность, так как ориентированы на использование статистических методов (Statistica, Stata, SPSS и т.п.), а данный курс сосредоточен на методах data mining и машинного обучения. Бесплатно распространяемые программные системы для анализа данных позволяют избежать указанных сложностей – обычно они создаются учеными-практиками в ведущих лабораториях и потому часто обладают наиболее актуальной на сегодняшний день функциональностью.

В лабораторных работах курса используются следующие открытые программные системы: Weka 3 – Data Mining Software in Java (разработана командой специалистов Университета Вайкато, Новая Зеландия); Orange – Data Mining Fruitful & Fun (пакет создан лабораторией искусственного интеллекта Университета Любляни, Словения); QuDA – Data Miner Discovery Environment (разработана в техническом Университете города Дармштадта, Германия); Coron System – платформа добычи данных (разработана коллегами из группы Orpailleur в лаборатории LORIA Университета Нанси, Франция); Concept Explorer – один из основных инструментов анализа формальных понятий (разработан в Техническом университете Дармштадта, Германия); RSES2 – Rough Set Exploration System (разработана в Институте математики Университета Варшавы, Польша). Каждая программная система используется как минимум в одной лабораторной работе, а все перечисленные средства могут работать под управлением большинства современных ОС.

Другая проблема для такого курса – нехватка реальных данных, поэтому предлагается использовать репозитории, сформированные научным сообществом, в частности UCI Machine Learning Repository, созданный для нужд исследователей в области машинного обучения в Калифорнийском университете Ирвина и содержащий 190 наборов данных по разным областям физики, техники, биологии, медицины, социологии, бизнеса и др. Другой тип репозиториев характерен для соревнований в рамках конференций по анализу данных, например, Frequent Itemset Mining Implementations Repository, в котором помимо данных содержатся исходные коды алгоритмов. Хранимые в них наборы данных получены при решении реальных задач, многие из которых представляют собой актуальную научно-практическую проблему – ученые применяют эти наборы данных для доказательства качества и пригодности предложенных ими новых методов анализа данных.

Все программы, а также наборы данных и тексты лабораторных работ доступны в электронном виде, в том числе на сайте факультета. Возможность выполнять практикум вне аудиторных условий делает его пригодным для самостоятельной работы в рамках тех курсов, где аудиторное число часов ограничено или нет возможности использовать компьютерное оборудование.

## Предварительные требования к знаниям, умениям и навыкам студентов

Студенты должны владеть основными понятиями из курса дискретной математики: множество, отображение, бинарное отношение, свойства бинарных отношений, частичный порядок, диаграмма частичного порядка, функция, исчисление высказываний и предикатов первого порядка, граф и алгоритм. Знания из курса линейной алгебры включают вычисления с матрицами, нахождение собственных чисел и собственных векторов, решение матричных уравнений. Знания из курса теории вероятностей предполагают предварительное знакомство студентов с понятием вероятности, алгеброй событий, независимости событий и теоремой Байеса. Дополнительным требованием является знакомство с понятием информационной энтропии Шеннона.

***

Несмотря на появление учебной литературы по методам машинного обучения и добычи данных, предлагаемый лабораторный практикум уникален на российском образовательном рынке в силу открытости используемого ПО, предоставляемого ведущими международными научно-исследовательскими коллективами, и ориентацией именно на выработку умений по его применению в учебных и реальных задачах.

*Дмитрий Игнатов (dignatov@hse.ru) – преподаватель кафедры анализа данных и искусственного интеллекта, ГУ-ВШЭ (Москва).*

**Знания, умения и навыки**

Основные знания, необходимые для свободного выполнения практикума, получены студентом в рамках лекций соответствующих курсов, тем не менее для каждой лабораторной работы приводится необходимый теоретический минимум. В перечень основных знаний, активно использующихся в курсе, входят:

- добыча данных (data mining) и машинное обучение (machine learning) как области современного анализа данных;
- задачи предварительной обработки данных – очистка, шкалирование, дискретизация и другие методы классификации, кластеризации и прогнозирования;
- методы поиска ассоциаций и частых множеств признаков, модели и методы прикладной теории решеток для анализа данных (анализ формальных понятий);
- упорядоченные множества для анализа данных;
- способы оценки качества результатов анализа данных (скользящий контроль, точность и полнота и т.п.).

Студенты должны обладать навыками установки и настройки свободного ПО для анализа данных (Concept Explorer, Coron, Orange, Weka, QuDA, RSES2 и т.д.); загрузки учебных и исследовательских наборов данных из открытых репозиториев, например UCI и FIMI и т.п.; работы с наборами данных и программным обеспечением. Особое внимание уделяется таким аспектам, как: умение выбрать метод анализа данных в соответствии с поставленной целью, характером задачи и данных; понимание математических моделей, лежащих в основе методов, описанных в базовых терминах теории множеств, упорядоченных структур, прикладной алгебры и т.п.; способность студента сформулировать и выполнить простые модельные расчеты, поясняющие суть конкретного метода; написание учебных (аналитических) отчетов, представляющих собой мини-исследование по применению конкретной модели, метода и данных, с результатами экспериментов, промежуточными отчетами и выводами (фактически протокол выполнения лабораторной работы); поисковые умения, направленные на исследование актуальной проблемы или задачи, которые активно обсуждаются научным сообществом; чтение дополнительной научной и учебной литературы, в том числе на английском языке, изучение нового ПО (не описанного в текстах практикума); умение правильно интерпретировать полученные результаты.

**Пример расчетного задания для задачи классификации: угоняемость автомобилей**

| Цвет | Тип | Производство | Повреждения | Угоняют? |
|---|---|---|---|---|
| Красный | Спортивный | США | нет | Да |
| Желтый | Спортивный | Япония | нет | Да |
| Желтый | Джип | Япония | нет | Да |
| Красный | Спортивный | Япония | есть | Да |
| Желтый | Спортивный | США | есть | Нет |
| Желтый | Джип | США | нет | Нет |
| Красный | Джип | Япония | есть | Нет |
| Желтый | Спортивный | США | есть | Нет |
| Красный | Спортивный | Германия | нет | Да |
| Черный | Джип | Япония | нет | Неопределенно |

Требуется предсказать факт угона, и, как видно в этом случае, без предварительного шкалирования справиться с задачей сложно. Если решать эту задачу с помощью ДСМ-метода (метод назван в честь английского философа Джона Стюарта Милля и основан на обучении гипотезам по положительным и отрицательным примерам явления с помощью операции сходства), то можно получить несколько гипотез в пользу положительной (угоняют) и отрицательной (не угоняют) классификации объектов. Положительные: {красный, спортивный}, {желтый, Япония, нет повреждений} и {спортивный, Япония}. Отрицательные: {желтый, США} и {красный, джип, Япония, есть повреждения}. Согласно найденным гипотезам примеры 8, 9 и 10 классифицируются соответственно отрицательно, положительно и неопределенно. Подобные задачи студент решает, выполняя вычисления вручную во время сдачи допуска к лабораторной работе.

Различные методы обладают своими особенностями, например, ДСМ-метод строит прогнозы очень осторожно, что делает его полезным, например, в задачах прогнозирования токсичности веществ – меньше ошибка отнесения ядовитых веществ к нетоксичным. Задачи для вычислений с помощью программных систем проводятся на более крупных наборах данных: когда ясна суть метода, очень важно научить аналитиков умению интерпретировать результаты, среди которых может оказаться не так много новых нетривиальных знаний.

Предлагать наборы данных более крупных размеров, содержащие несколько миллионов объектов или признаков, не входит в задачи курса, так как для успешного овладения методами снижения размерности и отбора релевантных («интересных») объектов или признаков достаточно исследования массивов размерами порядка 1 тыс. объектов на 100 признаков.

02.09.2010г.

---

**Постоянный URL статьи:** http://www.osp.ru/os/2010/Управление облаками/13003739/