СОВРЕМЕННЫЕ ТЕХНОЛОГИИ

В. А. Дюк, А. В. Флегонтов, И. К. Фомина

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЕСТЕСТВЕННОНАУЧНЫХ, ТЕХНИЧЕСКИХ И ГУМАНИТАРНЫХ ОБЛАСТЯХ

В статье рассматриваются различные области применения информационных технологий интеллектуального анализа данных. На практических примерах демонстрируется эффективность и общность методологических подходов и применения инструментария извлечения знаний на основе компьютерных технологий. В излагаемом материале характеризуются ключевые моменты обширного направления в анализе данных — data mining и особенности его применения в естественнонаучных, технических и гуманитарных областях.

Ключевые слова: интеллектуальный анализ данных, компьютерные технологии, инструментарий извлечения знаний.

V. Dyuk, A. Flegontov, I. Fomina

APPLICATION OF DATA MINING TECHNOLOGIES IN THE SCIENTIFIC, TECHNICAL AND HUMANITARIAN AREAS

The article discusses the various fields of application of information technology data mining. Practical examples demonstrate the effectiveness and generality of methodological approaches and tools of knowledge extraction based on computer technology. The key points of the vast areas in the analysis of data — data mining, and particularly its applications in scientific, technical and humanitarian areas are described.

Key words: data mining, computer technology, knowledge extraction tools.

В современном обществе центр экономического развития переносится с материальных сфер производства (энергетическо-сырьевой базис) на наукоемкие и высокотехнологичные сферы. Поступательное движение, в том числе в области экономики, определяется сегодня и будет определяться в ближайшее десятилетие совершенствованием информационных технологий. *Информационное общество* — это нынешний этап социальной эволюции человечества.

Движущей силой информационного общества являются *знания — интеллектуально-информационный ресурс (ИИР)*. Это новая и непривычная категория, активно включаемая сегодня в сферу деятельности человека. Относительно ИИР человеку неизвестны законы сохранения или ограничения, так характерные для *вещественно-энергетической (материальной) субстанции*. По многим параметрам (динамика развития, эффективность внедре-

ния и др.) ИИР имеет неоспоримые преимущества по сравнению с материальными ресурсами.

Общество, базирующееся на информационной экономике, уже по своей структуре избегает большинства социально-экономических и экологических проблем, ситуационно тяготеющих над нами сегодня, и в потенциале предполагается его экспоненциальное развитие по всем основным параметрам («*знания — порождают знания*»).

Важнейшим проявлением качественного технологического рывка, приведшего к возникновению информационного общества, и одновременно одной из его существенных черт является возникновение и стремительное распространение так называемых «метатехнологий» или «гипертехнологий». Это кардинально снижает значение финансовых ресурсов с точки зрения конкурентоспособности обществ и корпораций: если раньше они были главным источником могущества, то теперь превращаются в его следствие. Главным источником рыночной силы становится интеллект, воплощенный в организационных структурах исследовательских и рыночных корпораций, создающих метатехнологии и удерживающих контроль за ними. В информационном обществе все больший вес приобретают высококвалифицированные специалисты — «золотые воротнички», владеющие метатехнологиями.

В настоящей статье мы сконцентрируем внимание на Data Mining — одной из современных аналитических метатехнологий, предназначенной для переработки сырой информации с целью получения продуктивных знаний.

По теме Data Mining написаны десятки, если не сотни, книг. Количество статей тоже весьма велико — в поисковике Google на это словосочетание на момент подготовки статьи выдавалось 2 540 000 ссылок.

Попытаемся кратко охарактеризовать ключевые моменты этого обширного направления в анализе данных.

В связи с совершенствованием технических средств для получения, записи и хранения информации на специалистов обрушились колоссальные потоки разнородных данных. Вместе с тем традиционная математическая статистика оказалась неспособной обеспечить продуктивное решение ряда актуальных задач из различных предметных областей (поиск закономерностей в многомерных данных, построение диагностических и прогностических моделей, выявление сложных непериодических паттернов в динамических рядах и др.). Одна из причин — концепция усреднения по выборке, приводящая к операциям над фиктивными величинами. Кроме того, практически отсутствуют аналитические критерии для оценки достоверности взаимосвязей и регулярностей в многомерных данных и др.

Направление Data Mining родилось как ответ на сложившуюся проблемную ситуацию. В настоящее время термин «Data Mining» (раскопка данных) является синонимом появившегося позже (1989) термина «обнаружение знаний в базах данных» (Knowledge Discovery in Databases — KDD). В русском языке область, очерченная вышеупомянутыми терминами, нередко обозначается словосочетанием «интеллектуальный анализ данных» (ИАД).

Исходное определение дал наш бывший соотечественник Григорий Пятецкий-Шапиро:

«Data mining — это процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» (G. Piatetsky-Shapiro).

В настоящее время ИАД существует в двух ипостасях. Ряд специалистов делает акцент на обработке сверхбольших объемов данных. Здесь предъявляются повышенные требования к быстродействию алгоритмов, естественно, в ущерб оптимальности результатов.

Другая группа специалистов концентрирует внимание на глубине раскопки данных. В понимании этой группы основные отличия технологии ИАД следующие:

– ИАД — это всегда сугубо многомерные задачи — поиск связи между значением целевого показателя и набором значений группы других показателей БД.

– Технологии ИАД предназначены для обработки разнородной информации, то есть поля могут быть представлены количественными, качественными и текстовыми переменными.

– Технология ИАД, в отличие от традиционных статистических методов, не претендует на поиск взаимосвязей, характерных для полного объема данных (всей выборки). Ищутся правила, связывающие значения показателей, для подвыборок данных. При этом эти правила всегда высокоточные, а не «размытые» по всей выборке, общие и неточные статистические тенденции.

– Алгоритмы ИАД производят поиск указанных выше подвыборок данных и точных взаимосвязей для этих подвыборок в автоматическом режиме.

Таким образом, ключевые слова ИАД: точность, многомерность, разнотипность данных, автоматический поиск. Здесь, конечно, еще нужно добавить важное требование интерпретируемости получаемого результата.

Дополнительную информацию можно найти на многочисленных сайтах Интернета, из которых один из самых информативных — сайт упомянутого Г. Пятецкого-Шапиро — www.kdnuggets.com. Также весьма полезным является популярный ресурс — репозиторий данных UCI университета г. Ирвин (Калифорния, США), история которого началась в 1987 г. Адрес — http://archive.ics.uci.edu/ml/. Здесь можно найти массу данных и ссылок на примеры решения задач из самых разных областей, в том числе относящихся к теме исследования живых систем.

Методы ИАД имеют много общего с методами решения задач классификации, диагностики и распознавания образов. Но одной из главных их отличительных черт, как отмечалось выше, является функция интерпретации закономерностей, кладущихся в основу правил вхождения объектов в классы эквивалентности. Поэтому сегодня все большее распространение получают логические методы. Есть еще одна важная причина, обусловившая приоритет логических методов. Она заключается в сложной системной организации областей, составляющих предмет приложения современных информационных технологий. Эти области относятся, как правило, к надкибернетическому уровню организации систем [9], закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей. Гибкость и многообразие логических конструкций индуктивного вывода позволяют нередко добиваться успешных результатов при описании таких сложных систем.

Другие методы ИАД для построения диагностических и прогностических моделей имеют менее прозрачную интерпретацию. Сюда относятся байесовские классификаторы, дискриминантный анализ, нейросетевой подход, метод ближайших соседей, метод опорных векторов, генетические алгоритмы и др. Как показала практика последнего десятилетия, в ряде задач (особенно в бизнес-приложениях, где требуется анализировать огромные базы данных) требование интерпретируемости результатов стало отступать на задний план. Акцент здесь стал делаться на стабильности получаемых решений. Более того, на передний план начали выходить методы работы с комитетами, содержащими сотни и тысячи методов и алгоритмов. Как выяснилось, подобные комитеты, состоящие даже из «слабых» алгоритмов, способны превосходить по точности изолированные «сильные» алгоритмы, нацеленные на поиск глубоких закономерностей в массивах данных. Эта тенденция современного

ИАД нуждается в самостоятельном рассмотрении. Здесь наблюдается явное отступление от изначальных идеалов ИАД, связанных с попытками извлечения знаний из данных, а не с построением моделей в виде «черных ящиков».

При работе с комитетами алгоритмов сегодня широко используются 2 общих технологических приема или метода, имеющих чрезвычайную важность для ИАД. Это «бустинг» (boosting) и «бэггинг» (bagging — сокращение от bootstrap aggregation). Эти приемы предназначены для повышения «обобщающей способности» получаемых моделей — способности выдавать правильные результаты не только для примеров, участвовавших в процессе обучения, но и для любых новых, не участвовавших в процессе обучения данных. Кратко охарактеризуем эти два приема.

Идея бустинга предложена в конце 1980-х гг. [12] в контексте фундаментального вопроса об эквивалентности слабого и сильного обучения. Бустинг реализует процедуру последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. В течение последних 10 лет бустинг остается одним из наиболее популярных методов ИАД. Основные причины: простота, универсальность, гибкость (возможность построения различных модификаций) и, главное, высокая обобщающая способность.

Бустинг деревьев решений считается одним из наиболее эффективных методов решения задач классификации. В ряде экспериментов наблюдалось практически неограниченное уменьшение частоты ошибок на независимой тестовой выборке по мере наращивания композиции. Более того, качество на тестовой выборке часто продолжало улучшаться даже после достижения безошибочного распознавания всей обучающей выборки. Это изменило существовавшие долгое время представления о том, что для повышения обобщающей способности необходимо ограничивать сложность алгоритмов. На примере бустинга стало понятно, что хорошим качеством могут обладать сколь угодно сложные композиции, если их правильно настраивать.

Теоретическое обоснование эффективности бустинга связано с тем, что взвешенное голосование сглаживает ответы алгоритмов, входящих в комитет. Эффективность бустинга объясняется тем, что по мере добавления базовых алгоритмов увеличиваются отступы обучающих объектов. Причем бустинг продолжает раздвигать классы даже после достижения безошибочной классификации обучающей выборки.

Бэггинг — это метод формирования ансамблей классификаторов с использованием случайной выборки с возвратом, или бутстрепа. Он был предложен в 1994 г. [11].

При формировании бутстреп-выборок из множества данных случайным образом отбирается несколько подмножеств. Так как отбор производится случайно, набор примеров в этих подмножествах будет различным: некоторые из них могут быть отобраны по несколько раз, а другие — ни разу. Затем на основе каждого подмножества (выборки) строится классификатор. Выходы полученных классификаторов комбинируются (агрегируются) путем голосования или простого усреднения. Считается, что результат будет намного точнее любой одиночной модели, построенной на исходном наборе данных.

Известно много работ по сравнительному анализу обобщающей способности бустинга и бэггинга. Бэггинг направлен исключительно на уменьшение вариации модели, в то время как бустинг способствует уменьшению и вариации и смещения [10]. Эмпирические исследования этих методов на реальных задачах показали, что бустинг работает лучше на больших обучающих выборках, бэггинг — на малых. Бустинг лучше воспроизводит границы классов сложной формы. При увеличении длины выборки бустинг повышает разнооб-

разие классификаторов активнее, чем бэггинг, хотя этот недостаток бэггинга может быть восполнен методом генерации случайных подпространств.

В целом, как было отмечено выше, в области ИАД за последнее десятилетие произошли существенные изменения. Слово «интеллектуальный» теперь нужно воспринимать скорее в контексте автоматического построения классифицирующих и прогнозирующих моделей. Поиск сильных индивидуальных методов и алгоритмов для основной массы специалистов ИАД стал не столь актуальным — их интересы сместились в сторону умений работать с большими коллективами «слабых» методов и алгоритмов.

Вместе с тем сфера применения ИАД, как и прежде, ничем не ограничена — она везде, где имеются какие-либо данные. В первую очередь сегодня ИАД представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов ИАД они могут получить ощутимые преимущества в конкурентной борьбе. Выборочно опишем некоторые возможные бизнес-приложения ИАД [3; 6; 7].

Розничная торговля. Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля. Вот типичные задачи, которые можно решать с помощью ИАД в сфере розничной торговли:

– *анализ покупательской корзины* (анализ сходства) предназначен для выявления товаров, которые покупатели стремятся приобретать вместе. Знание покупательской корзины необходимо для улучшения рекламы, выработки стратегии создания запасов товаров и способов их раскладки в торговых залах;

– *исследование временных шаблонов* помогает торговым предприятиям принимать решения о создании товарных запасов. Оно дает ответы на вопросы типа: «Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?»;

– *создание прогнозирующих моделей* дает возможность торговым предприятиям узнавать характер потребностей различных категорий клиентов с определенным поведением, например, покупающих товары известных дизайнеров или посещающих распродажи. Эти знания нужны для разработки точно направленных, экономичных мероприятий по продвижению товаров.

Банковское дело. Достижения технологии ИАД используются в банковском деле для решения следующих распространенных задач:

– *выявление мошенничества с кредитными карточками*: путем анализа прошлых транзакций, которые впоследствии оказались мошенническими, банк выявляет стереотипы такого мошенничества;

– *сегментация клиентов*: разбивая клиентов на различные категории, банки делают свою маркетинговую политику более целенаправленной и результативной, предлагая различные виды услуг разным группам клиентов;

– *прогнозирование изменений клиентуры*: ИАД помогает банкам строить прогнозные модели ценности своих клиентов и соответствующим образом обслуживать каждую категорию.

Телекоммуникации. В области телекоммуникаций методы ИАД помогают компаниям более энергично продвигать свои программы маркетинга и ценообразования, чтобы удерживать существующих клиентов и привлекать новых. Среди типичных мероприятий отметим следующие:

– *анализ записей о подробных характеристиках вызовов*: назначение такого анализа — выявление категорий клиентов с похожими стереотипами пользования услугами и разработка привлекательных наборов цен и услуг;

– *выявление лояльности клиентов*: ИАД можно использовать для определения характеристик клиентов, которые, один раз воспользовавшись услугами данной компании, с большой долей вероятности останутся ей верными. В итоге средства, выделяемые на маркетинг, можно тратить там, где отдача больше всего.

Страхование. Страховые компании в течение ряда лет накапливают большие объемы данных. Здесь обширное поле деятельности для методов ИАД:

– *выявление мошенничества*: страховые компании могут снизить уровень мошенничества, отыскивая определенные стереотипы в заявлениях о выплате страхового возмещения, характеризующие взаимоотношения между юристами, врачами и заявителями;

– *анализ риска*: путем выявления сочетаний факторов, связанных с оплаченными заявлениями, страховщики могут уменьшить свои потери по обязательствам. Известен случай, когда в США крупная страховая компания обнаружила, что суммы, выплаченные по заявлениям людей, состоящих в браке, вдвое превышает суммы по заявлениям одиноких людей. Компания отреагировала на это новое знание пересмотром своей общей политики предоставления скидок семейным клиентам.

Другие приложения в бизнесе. ИАД может применяться во множестве других областей:

– *автомобильная промышленность*: при сборке автомобилей производители должны учитывать требования каждого отдельного клиента, поэтому им нужны возможности прогнозирования популярности определенных характеристик и знание того, какие характеристики обычно заказываются вместе;

– *политика гарантий*: производителям нужно предсказывать число клиентов, которые подадут гарантийные заявки, и среднюю стоимость заявок;

– *поощрение часто летающих клиентов*: авиакомпании могут обнаружить группу клиентов, которых данными поощрительными мерами можно побудить летать больше. Например, одна авиакомпания обнаружила категорию клиентов, которые совершали много полетов на короткие расстояния, не накапливая достаточно миль для вступления в их клубы, поэтому она таким образом изменила правила приема в клуб, чтобы поощрять число полетов так же, как и мили.

Специальные приложения:

Медицина. Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания симптомов различных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как нужно его лечить. Правила помогают выбирать средства медикаментозного воздействия, определять показания/противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения и т. п. Технологии ИАД позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил [5].

Молекулярная генетика и генная инженерия. Пожалуй, наиболее остро и вместе с тем четко задача обнаружения закономерностей в экспериментальных данных стоит в молекулярной генетике и генной инженерии. Здесь она формулируется как определение так называемых маркеров, под которыми понимают генетические коды, контролирующие те или иные фенотипические признаки живого организма. Такие коды могут содержать сотни, тысячи и более связанных элементов [1; 4; 5; 8]. Кроме того, на фронтире современных ис-

следований находятся эксперименты с различными молекулярными (в том числе ДНК) биочипами, содержащими десятки, сотни, тысячи и даже десятки тысяч реагентов с биопробами. Растущий интерес здесь в значительной степени мотивирован многочисленными практическими приложениями знаний, полученных из таких данных, в медицинской диагностике, разработке лекарств и др. При анализе данных биочипов исследователи сталкиваются с ситуациями, когда число изучаемых генов на два порядка превышает количество имеющихся образцов. Большинство стандартных алгоритмов классификации плохо справляются с решением задач большой размерности и при малом числе примеров почти гарантированно переобучаются. Кроме того, как правило только малая часть из огромного числа проверяемых генов актуальна в контексте решаемых задач. Из этого следует актуальность разработанных в ИАД и охарактеризованных выше методов бустинга и бэггинга.

Прикладная химия. Методы ИАД находят широкое применение в прикладной химии (органической и неорганической). Здесь нередко возникает вопрос о выяснении особенностей химического строения тех или иных соединений, определяющих их свойства. Особенно актуальна такая задача при анализе сложных химических соединений, описание которых включает сотни и тысячи структурных элементов и их связей.

Можно привести еще много примеров различных областей знания, где методы ИАД играют ведущую роль. Особенность этих областей заключается в их сложной системной организации. Они относятся главным образом к надкибернетическому уровню организации систем, закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей [2; 9]. Данные в указанных областях неоднородны, гетерогенны, нестационарны и часто отличаются высокой размерностью.

Применение ИАД в гуманитарной сфере приводит к ряду специфических проблем и задач. Многообразие разработанных к настоящему времени моделей, методик и методов анализа данных делает непростой задачу отбора методов для гуманитарного образования и требует изучения новых подходов, на которых они базируются, их особенностей и возможностей.

Потребность практического использования в гуманитарной сфере методов ИАД приводит к необходимости построения и соответствующей адаптации их обобщенных математических и алгоритмических моделей и создания оригинальных методик их применения и обучения. Создание механизма повышения эффективности отдельных методов ИАД выдвигает задачу разработки процедуры, обеспечивающей их адаптацию к различным отраслям гуманитарного применения.

Многообразие программного инструментария требует для своего успешного применения умения проводить сравнительный анализ эффективности его использования для решения той или иной задачи, вопроса о сроках и качестве освоения специалистами-гуманитариями.

СПИСОК ЛИТЕРАТУРЫ

1. *Вейр Б.* Анализ генетических данных: Пер. с англ. М.: Мир, 1995. 400 с.
2. *Гик Дж. ван*. Прикладная общая теория систем. М.: Мир, 1981.
3. *Дюк В. А.* Data Mining — интеллектуальный анализ данных // BYTE (Россия). 1999. № 9. С. 18–24.
4. *Дюк В. А*. Data Mining: Учебный курс. СПб.: Питер, 2001. 368 с.
5. *Дюк В. А., Эмануэль В. Л.* Информационные технологии в медико-биологических исследованиях. СПб.: Питер, 2003. 525 с.

6. *Киселев М., Соломатин Е.* Средства добычи знаний в бизнесе и финансах // Открытые системы. 1997. № 4. С. 41–44.

7. *Кречетов Н.* Продукты для интеллектуального анализа данных // Рынок программных средств. 1997. № 14–15. С. 32–39.

8. Математические методы для анализа последовательностей ДНК: Пер. с англ. / Под ред. М. С. Уотермена. М.: Мир, 1999. 349 с.

9. *Boulding K. E.* General Systems Theory — The Skeleton of Science // Management Science. 2. 1956.

10. *Freund Y., Schapire R. E.* Discussion of the paper «Arcing classifiers» by Leo Breiman // The Annals of Statistics. 1998. Vol. 26. No. 3. P. 824–832.

11. *Breiman L.* Bagging predictors // Machine Learning. 24 (1996). S. 123–140.

12. *Schapire Robert E.* The strength of weak learnability // Machine Learning. 1990.5(2). S. 197–227.

RFERENCES

1. *Vejr B.* Analiz geneticheskih dannyh: Per. s angl. M.: Mir, 1995. 400 s.

2. *Gik Dzh., van.* Prikladnaja obwaja teorija sistem. M.: Mir, 1981.

3. *Djuk V. A.* Data Mining — intellektual'nyj analiz dannyh // BYTE (Rossija). 1999. № 9. S. 18–24.

4. *Djuk V. A.* Data Mining : uchebnyj kurs. SPb.: Piter, 2001. 368 s.

5. *Djuk V. A., JEmanujel' V. L.* Informacionnye tehnologii v mediko-biologicheskih issledovanijah. SPb.: Piter, 2003. 525 s.

6. *Kiselev M., Solomatin E.* Sredstva dobychi znanij v biznese i finansah // Otkrytye sistemy. 1997. № 4. S. 41–44.

7. *Krechetov N.* Produkty dlja intellektual'nogo analiza dannyh // Rynok programmnyh sredstv. 1997. № 14–15. S. 32–39.

8. Matematicheskie metody dlja analiza posledovatel'nostej DNK: Per. s angl. / Pod red. M. S. Uotermena M.: Mir, 1999. 349 s.

9. *Boulding K. E.* General Systems Theory The Skeleton of Science // Management Science. 1956. № 2.

10. *Freund Y, Schapire R. E.* Discussion of the paper «Arcing classifiers» by Leo Breiman // The Annals of Statistics.1998. Vol. 26. No. 3. P. 824–832.

11. *Breiman Leo.* Bagging predictors // Machine Learning. 24 (1996), 123-140.

12. *Schapire Robert E.* The strength of weak learnability // Machine Learning. 1990. № 5(2). S. 197–227.

Ю. А. Жук, В. В. Фомин, Л. В. Уткин

ОПРЕДЕЛЕНИЕ РЕФЛЕКСИВНЫХ ПОКАЗАТЕЛЕЙ ДЛЯ ОЦЕНКИ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ ДИСПЛЕЙНЫХ ФОРМ НАГЛЯДНОСТИ

В статье описывается опыт разработки гипермедийной структуры дисплейной наглядности. Рассматривается ее опытная апробация в процессе преподавания органической химии в вузе. Представлены результаты экспериментов по оценке эффективности использования дисплейных форм наглядности, в которых одним из критериев оценки являлись рефлексивные показатели. Дана методика оценки целесообразности использования дисплейных форм наглядности в учебном процессе вуза на основе многокритериальной задачи принятия решений при смешанной стратегии.

Ключевые слова: дисплейные формы наглядности, когнетика, технические средства обучения.