АЛГОРИТМЫ DATA MINING В ЗАДАЧАХ УПРАВЛЕНИЯ ДИНАМИЧЕСКИМИ ПРОЦЕССАМИ

А. А. Мусаев[1], И. А. Барласов[2]

[1]Санкт-Петербургский институт информатики и автоматизации РАН, [2]НОУ «Международный Банковский Институт»

[1]СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178; [2]НОУ «МБИ», Невский пр., д., 58, Санкт-Петербург, 191023

[1]<amusaev@szma.com>, [2]<barlasov@yandex.ru>

УДК 681.3.01

Мусаев А. А., Барласов И. А. **Алгоритмы Data Mining в задачах управления динамическими процессами** // Труды СПИИРАН. Вып. 5. — СПб.: Наука, 2007.

Аннотация. *В статье рассматриваются основные алгоритмы интеллектуального анализа данных (Data Mining, DM), лежащие в основе нового типа автоматизированного управления многомерными динамическими процессами — аналитического. Существенным отличием аналитического управления является сочетание оперативных управленческих решений, формируемых должностными лицами на основе данных мониторинга текущей ситуации, с результатами глубокого количественного анализа ретроспективных данных (накопленного опыта), реализуемого средствами DM. Сформулированы концептуальные основы аналитического управления, позволяющие выделить DM в качестве самостоятельного подкласса информационных технологий.* — Библ. 18 назв.

UDC 681.3.01

Musayev A. A., Barlasov I. A. **Data Mining algorithms in the dynamic processes control tasks** // SPIIRAS Proceedings. Issue 5. — SPb.: Nauka, 2007.

Abstract. *The article is devoted to DM (Data Mining) algorithms which are the basis of new type of automatic control of multivariate dynamic processes. DM methods combine immediate control decisions with deep numerical analysis of retrospective data. New conceptual principles of analytic control allow DM to be separate part of information technology.* — Bibl. 18 items.

1. Введение

Проблема алгоритмизации управления многомерными динамическими процессами имеет достаточно давнюю историю и, по-видимому, ее можно было бы начать со времен английской промышленной революции. Тем не менее, ограничимся очень кратким обзором более поздних достижений, а точнее — разработок второй половины 20 века. Именно в это время обильный поток исследований в области теории управления, в целом, и процессов оптимизации управления многомерными динамическими процессами, в частности, оказался подкрепленным мощным развитием средств электронной и вычислительной техники.

Разработка семейства ЭВМ IBM 360 и 370 серий и отвечающий им ряд машин серии ЕС сформировали техническую платформу для создания АСУ. Достаточно очевидно, что столь громоздкие и сложные в обслуживании ЭВМ, как мейнфреймы (mainframe) 70–80-х годов, можно было использовать только в системе централизованного управления предприятием. В то же время, высокий уровень централизации требовал применения крайне сложных многопараметрических и многосвязных алгоритмов управления, для реализации которых вычислительные мощности имеющихся ЭВМ были недостаточными.

Возникшее противоречие при создании эффективной системы алгоритмического обеспечения привело к необходимости перехода к децентрализованным системам управления, в которой различные частные задачи можно было

бы решать с определенной степенью автономности локальными средствами автоматизированного или автоматического управления. В 90-е годы данную идею удалось реализовать благодаря появлению нового поколения ЭВМ — персональных компьютеров, и широкому использованию промышленных контроллеров. Однако любая децентрализация управления может сохранять эффективность лишь в определенных пределах. Основополагающие принципы системного подхода требовали нового витка на обратном пути к централизации управления. В результате этого во второй половине 90-х годов возникли новые стандарты автоматизированного управления предприятиями (IRP ERP, ERP-II, MRP-II, EAM и др.) [3].

Разработка соответствующих алгоритмов управления привела к созданию концепции MMI (man-machine interface), с некоторой долей условности разделяющуюся на технологии SCADA (Supervisory Control And Data Acquisition) и DCS (Distributed Control System). Одновременно формировались разнообразные решения по построению системы оперативного управления для стандарта MES (*Manufacturing Execution Systems*). Представленные стратегические изменения в принципах автоматизированного управления предприятиями вели к соответствующим изменениям в структуре алгоритмов управления. При этом, если содержательная часть алгоритмов управления оставалось, в достаточной степени, неизменной (последовательная коррекция режимов на основе результатов оперативного мониторинга), то их оптимизационная компонента претерпела существенные изменения и вылилась в целый ряд новых разработок, объединенных общей методологией APC (*advance process control*).

Примерно в это же время в сфере автоматизации финансовых и бизнес-приложений возник целый ряд новых решений, объединивший перспективные идеи в области компьютерной математики, систем хранения, представления и визуализации данных в рамках общего направления — *аналитических информационных технологий* (АИТ).

В настоящей статье приведено краткое описание основных идей аналитического управления многомерными динамическими процессами, базирующегося на DM, и представлены некоторые типы алгоритмов управления, используемые при решении задач автоматизированной поддержки принятия решений.

Идеи аналитического управления постепенно охватывали все новые предметные области и в сфере автоматизации частично пересеклись с математическими средствами APC технологий. В частности, алгоритмы нейросетевого прогнозирования и многомерного статистического анализа, применяемые в аналитических технологиях управления, активно используются при создании виртуальных анализаторов, автоматизированных *систем поддержки принятия решений* (СППР или Decision Support Systems, DSS) и других APC средствах [12].

Характер дальнейшего применения алгоритмов АИТ в сфере управления многомерными динамическими процессами покажет время. Весьма вероятно их использование на оперативном (MES) уровне управления, в частности, в системе автоматизированной диспетчеризации многомерными динамическими процессами. Однако наибольший эффект от внедрения АИТ следует ожидать при использовании в системе стратегического управления предприятием, например, в задачах экономического и маркетингового анализа.

2. Аналитические информационные технологии: новые альтернативы

Современный подход к оптимизации управления многомерными динамическими процессами предполагает создание алгоритмического обеспечения во взаимосвязи с вопросами развития информационных технологий. Это означает переход к качественному объединению разнородных технологий, позволяющему осуществлять разработку математических алгоритмов с учетом возможности модификации и развития соответствующих систем хранения и переработки информации.

Примером реализации подобного подхода является DM, представляющий собой подкласс информационных технологий, ориентированных на задачи автоматизированной поддержки принятия решений и прогнозирования состояния сложных динамических систем в нестационарных и неоднородных средах. Разумеется, вопросы построения СППР и связанные с ними задачи ситуационного анализа и прогностики рассматривались и ранее. Однако эффективность их решения оставалась невысокой. Классические математические технологии анализа и прогнозирования развития ситуаций использовалось в управлении многомерными динамическими процессами крайне незначительно в виду низкой достоверности получаемых результатов. Особенно остро это проявлялось в неспособности формальных алгоритмов отследить качественные, скачкообразные изменения контролируемых процессов. В свою очередь, низкая эффективность алгоритмов анализа, прогнозирования и оптимизации обуславливалась, как правило, недостаточной полнотой и оперативностью мониторинга состояния объекта управления, связанная с ограниченными возможностями средств цифровой технике по быстродействию и объему необходимой памяти.

В настоящее время основным и наиболее распространенным инструментом прогнозирования развития сложных ситуаций и выработки управляющих решений является эмпирический анализ. Большинство автоматизированных систем управления работали и работают по единой методологии: осуществляется сбор информации, ее визуализация и представление специалистам по оперативному управлению в заданной предметной области. По существу, информационная система работает в этом случае в режиме оперативной визуализации и передачи команд к исполнительным подсистемам управления.

Однако в сложных, нестационарных ситуациях, обусловленных большим числом разнообразных гетерогенных факторов влияния, специалисты далеко не всегда находят рациональные решения, их мнения оказываются субъективными и противоречивыми. Человеческий мозг, как правило, не способен прогнозировать развитие ситуаций, находящихся под воздействием более 3–5 независимых факторов. Для взаимосвязанных воздействий даже опытный специалист способен корректно учесть не более трех факторов. В тоже время большинство реальных управленческих ситуаций требуют учета, как минимум, от 6 до 50 (и более) значимых факторов влияния.

Таким образом, возникает актуальная проблема создания качественно новой, *аналитической системы управления*, ориентированной на решение задач выработки управляющих решений на основе комплексного анализа оперативных ситуаций и прогнозирования их развития в интересах формирования и реализации оптимальных режимов управления.

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Используемые для решения указанной проблемы АИТ включают в себя ряд новых информационных и математических технологий, представленных на рис. 1.

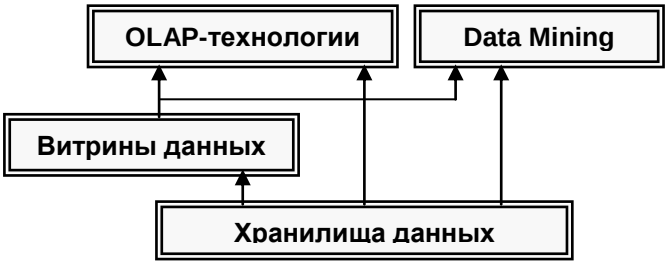

Рис. 1. Основные компоненты АИТ.

Возникновение новых технологий хранения и обработки данных (хранилища или склады данных, витрины данных) связано с необходимостью накопления и оперативной обработки сверхбольших объемов ретроспективной информации (единицы Гбайт) [2, 10, 11]. В частности, необходимость ускорения обработки аналитических запросов привела к разработке многомерного формата представления данных и созданию новых концепций хранения и обработки данных, представленных в виде постулатов Б. Инмона [17] и Е. Кодда [16].

Оперативный анализ текущей ситуации, ее сопоставление с данными ретроспективного анализа, в рамках АИТ, реализуются средствами *on-line analytical processing* (OLAP) [16]. Предполагается, что OLAP технологии могут стать мощным инструментом для визуализации оперативной ситуации в интересах совершенствования текущего управления.

Однако для более глубокого понимания протекающих многомерных процессов необходимо проведение соответствующих аналитических исследований, основанных на применении сложных математических средств анализа ситуаций, прогнозирования их развития и выработки оптимизирующих рекомендаций. Для решения этих задач используются, как уже указывалось, программно-алгоритмические средства «раскопок знаний в базах данных» или Data Mining. В некоторых отечественных публикациях данные технологии получили наименование *интеллектуального анализа данных.* Рассмотрим подробнее основные алгоритмы DM и его отличительные черты, как самостоятельного направления в области информационных технологий.

3. Data Mining: Инструмент формализованного анализа и прогнозирования динамических процессов

Современная прикладная математика предоставляет разработчикам алгоритмического обеспечения систем управления многомерными динамическими процессами обширный арсенал эффективных математических средств, используемых для решения задач оценивания, идентификации, распознавания, прогнозирования и т.п. Данные средства могут быть классифицированы на основе различных иерархических или фасетных представлений. В качестве одной из простейших схем классификации может быть предложена схема разделения, основанная на различных подходах к обучению математических моделей. В частности, различают:

статистические методы, основанные на использовании усредненного накопленного опыта, отраженного в массивах ретроспективных данных,

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

методы компьютерной математики, охватывающие множество разнородных вычислительных подходов (технологий искусственных нейронных сетей, ассоциативной памяти, нечеткой логики и т.п.).

Недостатки такой классификации достаточно очевидны. Перечисленные подходы, так или иначе, опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации. В то же время она достаточно удобна для интерпретации и, de facto, используется при описании математических средств современного подхода к извлечению знаний из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах Data Mining.

Важное отличие DM от известных методов оперативного анализа (OLTP — *one-line transaction processing*), используемых в существующих транзакционных *системах обработки данных* (СОД), состоит в переходе от технологии визуализации текущих ситуаций к фундаментальным методам исследований, опирающимся на мощный аппарат современной прикладной математики.

Основными задачами DM в управлении многомерными динамическими процессами являются комплексный системный анализ оперативных ситуаций, краткосрочный и долгосрочный прогноз их развития и выработка вариантов оптимизационных решений. Анализ оперативных (или текущих) ситуаций включает в себя:

- обнаружение и прогнозирование скрытых тенденций и закономерностей развития наблюдаемых процессов;
- обнаружение и распознавание скрытых факторов влияния (в том числе, факторов угрозы);
- обнаружение и идентификацию ранее неизвестных взаимосвязей между многомерными динамическими параметрами и факторами влияния;
- анализ среды взаимодействия динамических процессов и прогнозирование изменения ее характеристик;
- выработку оптимизационных рекомендаций по управлению многомерными динамическими процессами;
- визуализацию результатов анализа, подготовку предварительных отчетов и проектов допустимых решений с оценками достоверности и эффективности возможных реализаций.

На рис. 2 представлен математический арсенал DM. Рассмотрим подробнее представленные математические средства.

Рис. 2. Математический арсенал Data Mining.

4. Статистические методы DM

В качестве важнейшего направления развития средств DM следует выделить мощный арсенал статистических методов обработки данных. В соответствии с классификационной традицией, их удобно разделить на четыре взаимосвязанных раздела [13]:

- Предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- Выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- Многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластер-анализ, компонентный анализ, факторный анализ и др.);
- Динамические модели и прогноз на основе временных рядов.

Значимость статистических методов DM крайне велика — ведь именно в них наиболее последовательно отрабатывается мысль о принципиальной важности использования больших массивов ретроспективных данных для формирования эффективных управленческих решений.

Среди наиболее известных и популярных пакетов статистического анализа следует отметить Statistica, SPSS, Systat, Statgraphics, SAS, BMDP, TimeLab, DataDesk, S-Plus, Scenario (BI).

5. Кибернетические методы DM

Второе крупное направление развития связано с кибернетическими методами, основанными на идеях компьютерной математики и методах теории искусственного интеллекта. К этому направлению следует отнести методы нейронных сетей, эволюционного моделирования, генетические алгоритмы и методы нечеткой логики и другие.

Нейросетевые технологии. Формирование прогнозов и решений на основе нейросетевых технологий осуществляется путем применения математических моделей нейронных сетей, узлами которых являются модели нервных клеток (нейронов). Выходной сигнал нейрона определяется нелинейной функцией взвешенной суммы входных сигналов. В свою очередь входные сигналы представляют собой выходные сигналы нейронов предыдущего уровня. Входными сигналами всей сети являются параметры текущей ситуации. Ретроспективные данные используются в качестве обучающих выборок, формирующих значения весовых коэффициентов входов (синапсы) нейронов.

Эволюционное моделирование. Получение оптимальных решений осуществляется путем имитации процесса размножения и эволюции биологической популяции [14]. В исходные варианты решения вносятся различные, случайные изменения, имитирующие изменения наследственных свойств предыдущего поколения. Совокупность модифицированных данных образует новое поколение возможных решений, которое подвергается «естественному отбору» (или селекции), основанному на экзогенном «критерии выживания» (критерий допустимости решения). Сохранившиеся после селекции решения вновь модифицируются («размножаются»), образуя третье поколение, и процесс итерационно повторяется. При этом образуется неконтролируемая алгоритмистом, самоорганизующаяся последовательность, приводящая к наилучшему решению (прогно-

зу). При этом оптимум может быть найден самым неожиданным образом: наиболее эффективное решение может оказаться результатом последовательной эволюции, на промежуточных этапах которой ему предшествовали далеко не наилучшие (хотя и допустимые) решения.

По существу, ЭВМ в системе эволюционного моделирования, как и при использовании нейронных сетей, перестает быть программируемым калькулятором, а становится полноправным участником решения задачи. Это позволяет находить новые, нетривиальные и эффективные решения в самых сложных ситуациях: в задачах прогнозирования качества выходных продуктов ТП, выборе оптимальных технологических параметров установок, распознавании изменений в качестве исходного сырья и т.п.

Разновидностью эволюционного программирования являются метод эвристической самоорганизации и метод группового учета аргументов (МГУА), позволяющие выбирать наиболее эффективную структуру полиномиальной модели, описывающей эволюцию состояния изучаемого объекта [6].

Нечеткая логика. Особое направление в спектре DM составляют методы, основанные на нечетких множествах (*fuzzy sets*).

Традиционная вероятностно-статистическая методология базируется на классической колмогоровской аксиоматике. В ее основе лежит понятие меры, определенной на множестве σ-алгебр F в пространстве элементарных событий W. Однако в ряде практических задач, связанных, например, с лингвистическими переменными, подобную аксиоматику построить не удается. В связи с этим в 1961г. Л. Заде [5] была предложена концепция нечетких множеств, позволяющая оперировать с понятием неопределенности в неметрических системах. Применение теории нечетких множеств в системе DM позволяет ранжировать данные по степени близости к желаемому результату, осуществить, так называемый, нечеткий поиск в базах данных. Однако плата за повышенную универсальность всегда была достаточно велика и проявлялась в снижении уровня достоверности и точности получаемых результатов. Поэтому число специализированных приложений данной методологии, несмотря на повышенный интерес к ней со стороны математиков-прикладников в течении последних 35лет, весьма ограничено.

Перечисленные алгоритмы DM в настоящее время реализованы в качестве специализированных пакетов анализа данных. Среди основных программных продуктов, содержащих в себе кибернетические методы DM, следует назвать системы NeuroShell, GeneHunter, BrainMaker, OWL, PolyAnalyst, 4Thought (BI).

Ассоциативная память. Особое место среди алгоритмических средств DM занимают математические технологии, основанные на методах ассоциативной памяти. Некоторые авторы считают, что ассоциативные методы, основанные на поиске решений-аналогов в массивах ретроспективных данных, являются характеристическими атрибутами DM [4].

Работа с ассоциативной памятью обычно связана с формированием шаблонов изучаемых ситуаций. При этом предполагается, что имеется массив ретроспективных данных, позволяющий построить дискретный набор возможных ситуаций. Указанные ситуации можно формировать последовательно, например, используя «скользящее окно» над темпоральным массивом ретроспективы. Если каждому i-ому решению d_j (или ситуации из скользящего окна) из ретроспективы сопоставить точку в n-мерном фазовом пространстве решений

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

D_n, а новую ситуацию описать n-мерным вектором состояния x_n (с тем же набором параметров), то в качестве наиболее приемлемого решения d^* предлагается выбирать то, которое наиболее близко к сложившейся ситуации x. При этом мера близости оценивается по величине a'priori выбранной метрики μ, т.е. $d^* = \arg\min \mu(x_n d_j)$ для $\forall d_j$ [18].

Подобный подход, получивший наименовании метода «ближайшего соседа» («nearest neighbor», CBR — case based reasoning), и лег в основу таких программных продуктов, как Pattern Recognition Workbench или KATE tools.

Очевидно, что любое решение-прецендент обладает определенной информационной значимостью. Однако, чтобы оценить его достоверность и корректность интерпретации необходимы повторные опыты, позволяющие провести дискриминантные гиперповерхности раздела между «областями притяжения» опорных решений, что и предопределяет необходимость применения все тех же статистических методов в процессе обучения.

Деревья решений. Другой подход к выбору ситуационного решения связан с построением последовательного логического вывода на основе деревьев решений. В каждом узле дерева эксперт осуществляет простейший логический выбор («да»–«нет», «true»–«false»). В зависимости от принятого выбора поиск решения продвигается по правой или левой ветви дерева и, в конце концов, приходит к терминальной ветви, отвечающей конкретному окончательному решению. Очевидно, что и здесь процесс статистического обучения выведен за пределы программы и сконцентрирован в виде некоторого априорного опыта, заключенного в наборе ветвей-решений.

Одной из разновидностей метода деревьев решений, является алгоритм деревьев классификации и регрессии (CART, classification and regression trees), предлагающий набор правил для решения задачи дихотомической классификации совокупности исходных данных. Данный метод обычно применяется для определения возможных последовательности событий с заданным исходом.

Более тщательной предварительной подготовки данных требует метод CHAID (chi square automatic interaction detection), предназначенный для выявление зависимости между переменными по критерию хи-квадрат на основе традиционной схемы проверки гипотез.

Отсутствие обучения в отношении новых ситуаций, к сожалению, не исчерпывает недостаток такого подхода. Ведь для ее реализации, как минимум, нужен высококвалифицированный эксперт, способный в каждом узле принять содержательно верное решение. В тоже время реальная СППР, как правило, функционирует в условиях неопределенности, нестационарности и нерегулярности наблюдаемых процессов, в которых даже опытный эксперт не может чувствовать себя достаточно уверенным при формировании решений. Тем не менее, данный подход обеспечивает определенное удобства для экспертов, что и породило разработку на основе технологии деревьев решений (decision trees) таких программных продуктов, как IDIS, C5.0, SIPINA и т.п.

Системы обработки экспертных знаний. В завершении обзора алгоритмических средств DM, отметим предметно-ориентированные системы анализа ситуаций и прогноза, основанные на фиксированных математических моделях, отвечающих той или иной теоретической концепции. Роль эксперта состоит выборе наиболее адекватной системы и интерпретации полученного алгоритма. Достоинства и недостатки таких систем очевидны — предельная простота и доступность применения и расплата достоверностью и точностью за эту про-

стоту. Примерами программных продуктов, отвечающих предметно-ориентированным системам в области финансов являются Wall Street Money, MetaStock, SuperCharts, Candlestick Forecaster и другие [7].

К числу специализированных программных продуктов, позволяющих формировать предварительные отчеты и визуализировать результаты следует отнести системы Mineset, Impromptu (BI) и др. В частности, системы Mineset содержит в себе такие инструменты, как ландшафтный визуализатор, визуализаторы дисперсии, деревьев, правил, свидетельств и другие средства [1, 8].

6. Особенности системы аналитического управления

Для формирования общего представления о структуре аналитического управления с использованием средств DM, рассмотрим его отличительные особенности, роль и место в общей системе управления. С этой целью рассмотрим структурную схему, приведенную на рис. 3 и отражающие взаимодействие всех основных составляющих информационной системы формирования управляющих решений.

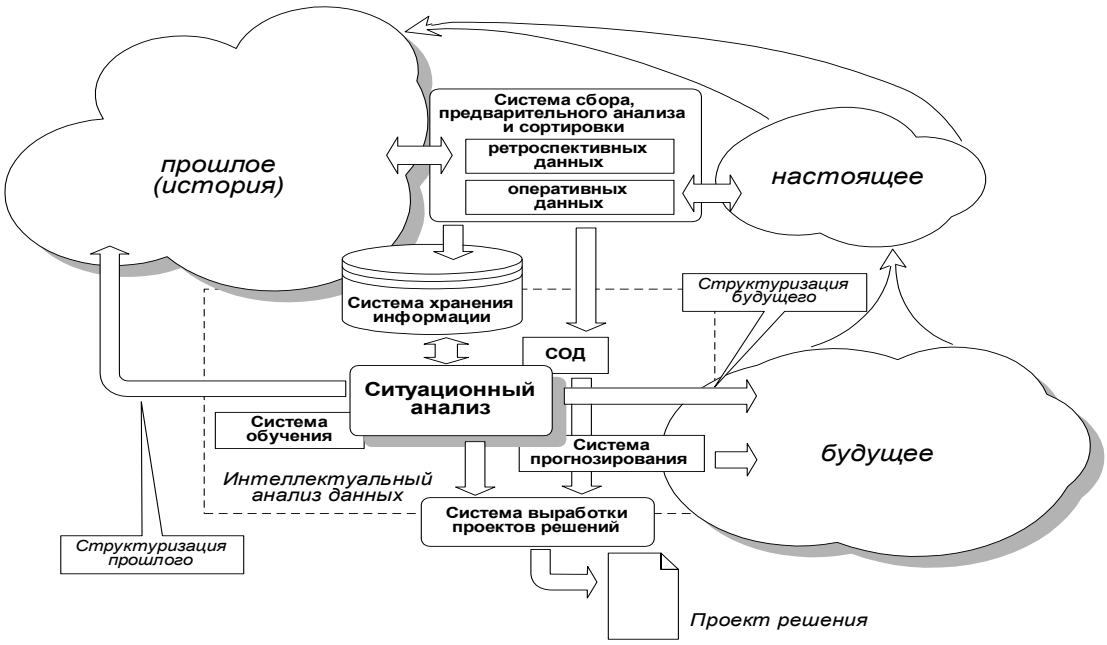

Рис. 3. Структура автоматизированного формирования управляющих решений.

Из рисунка видно, что основная идея DM состоит в сопоставлении результатов мониторинга текущей ситуации с предшествующим опытом управления (с «прошлым»), накопленным в форме массивов ретроспективных данных в информационном хранилище (хранилище данных, Data Warehouse, DW) предприятия. При этом DW выполняется в виде темпоральной БД, создавая основу для структуризации «прошлого» в форме, наиболее пригодной для поиска ситуаций-аналогов и ситуаций-прототипов. В частности, в накопленном ретроспективном опыте, методами ассоциативной памяти, формируются некоторые шаблоны («patterns») аналогов, позволяющие сопоставлять прошлое с текущей ситуацией и, тем самым, осуществлять прогноз развития нестационарных (в том числе скачкообразно изменяющихся) динамических процессов.

Разумеется, ассоциативный поиск не отвергает возможность одновременного классического регрессионного прогноза, отражающего динамику средних и несущего важную информацию об общих тенденциях эволюции оперативной ситуации. По всей видимости, неаддитивная смесь результатов статистической экстраполяции с выводами ассоциативного поиска является наиболее достоверным средством «структуризации будущее», создающего основу для построения прогностических сценариев и, тем самым, для выбора наилучших управленческих решений.

Следует заметить, что формирование (обоснование) накопленного решения является центральным элементом системы аналитического управления. Указанная функциональность помимо аналитической обработки (выявление скрытых закономерностей и взаимосвязей, оценка их влияния на основные показатели исследуемой ситуации, прогнозирование развития ситуаций и т.п.) включает в себя функции сбора и предварительной обработки информации, ее хранения, самообучения, подготовки наглядных отчетов и другие.

Таким образом, главным отличием DM от известных СОД, широко применяемых в современных информационных системах, по-видимому, следует считать попытку подойти к задаче формирования решения с позиции историзма, т.е. на основе полномасштабного количественного анализа всего ретроспективного опыта, предшествовавшего текущей ситуации, и позволяющего перенести результаты прецедентов на прогнозируемый сценарий.

Заметим, что DM не является альтернативой к традиционным СОД; транзакционная система обработки сохраняет за собой функции поддержки оперативных управленческих решения и подготовки экспресс-отчетов.

Акцент DM на количественной методологии позволяет перенести центр тяжести процедуры выработки проекта решения с эвристических логико-интуитивных методов, характерных для эмпирической технологии, на мощную, глубоко формализованную платформу прикладной математики. При этом качественный, эмпирический анализ также сохраняется, и его основным приложением остается вполне обозримый объем конечного перебора уже сформированных вариантов решений на фоне подготовленных прогностических сценариев. Таким образом, речь, по существу, идет о новой форме *гибридного интеллекта*, в которой машине отводится роль сверхмощного количественного анализатора, оставляя за человеком вопросы окончательных решений.

7. Концептуальные основы DATA MINING

Важной характеристической чертой DM является их технологическая основа — симбиоз разнородных средств прикладной математики с последними достижениями в области информационных систем. Роль алгоритмической базы в аналитическом управлении, как уже отмечалось, выполняют уже известные математические средства, и, прежде всего, методы прикладной статистики и перечисленные во втором разделе кибернетические алгоритмы (нейронные сети, эволюционное моделирование, генетические алгоритмы и т.п.). Объединяющим началом для возникновения аналитического управления послужила конкретная, крайне важная для практики цель — создание высокоэффективной автоматизированной СППР, способной учитывать в процессе выработки формализованного решения многолетний ретроспективный опыт, отраженный в сверхбольших объемах накопленных данных.

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Значимость решаемой задачи ускорило разработку и внедрение методов и средств DM. При этом практика настолько опередила теорию, что DM до последнего времени не имела собственной концептуальной платформы, определяющей ее как самостоятельную отрасль прикладных знаний. В связи с этим возникла необходимость в восполнении данного пробела, то есть в решении задачи формировании общих концептуальных принципов аналитического управления. Рассмотрим вариант формирования такой теоретической платформы в виде совокупности базовых принципов построения аналитических информационных систем.

Принцип историзма. Основным «сырьем» для аналитической обработки информации являются большие и сверхбольшие массивы ретроспективных данных, охватывающие поведение как самого *объекта управления* (ОУ), так и всей инфраструктуры, в которую он был погружен, внутри которой он развивался и с которой он активно взаимодействовал. При этом глубина ретроспективного анализа может быть весьма большой — от нескольких месяцев до нескольких лет и даже десятилетий.

Функционирование DM на множестве ретроспективных данных можно разбить на два этапа: поиск прецедентов и анализ их структуры. Результаты структурного анализа прецедентов трансформируются в формализованные выводы, используемые для корректировки результатов оперативной обработки текущих данных. Полученный скорректированный материал, в свою очередь, представляет собой основу для формирования проекта решения (или нескольких проектов решений) по рассматриваемому вопросу.

Очевидно, что механистический перенос исторического опыта на текущую ситуацию может привести к сугубо негативным результатам. Отсюда возникает необходимость в применении человеко-машинной технологии реализации DM.

Принцип системности. В задачах автоматизации управления (как и для большинства других предметных областей) в качестве основного ОУ выступает открытая динамическая система, погруженная в неоднородную и нестационарную эволюционирующую среду и активно с ней взаимодействующую. При этом предполагается, что ОУ отвечает всем основным признакам понятия «система» (а именно, целостности, структурированности и целенаправленности).

В сочетании с принципом историзма, идея системности предполагает формирование и хранение массивов ретроспективных данных, отражающих (количественно и качественно) процессы изменения состояний системы (ОУ) и среды в их историческом и текущем взаимодействии. Указанные массивы представляют собой «сырье», исходные данные, на основании которых средствами DM выявляются скрытые системные связи, неявные закономерности, совокупность значимых для развития системы факторов, условия их реализации и т.п.

Принцип гибридного человеко-машинного интеллекта. По своей природе методология DM опирается на сочетание автоматического компьютерного анализа сверхбольших объемов данных с экспертными заключениями, ориентированными на семантические аспекты решаемой задачи. Как правило, естественный интеллект подключается в наиболее критические моменты процедур анализа и выработки решений. Обычно это происходит, когда количественный подход не позволяет сформировать метрическую систему предпочтений, либо при отсутствии достаточного объема исходных данных для построения формализованного вывода.

Примером такой ситуации может служить задача предварительного выбора глубины ретроспективного поиска. Еще более явным примером может слу-

жить задача качественной отбраковки прецедентов, выявленных компьютерной программой на основе предварительного ассоциативного поиска.

Реализация данного принципа в системах аналитического управления требует решения проблемы рационального распределения функций в человеко-машинных системах и формирования интеллектуально-эргономических интерфейсов, наиболее согласованных с профессиональными представлениями предметных экспертов и *лиц, принимающих решение* (ЛПР).

Принцип симбиоза математических и информационных технологий. DM представляет собой область знаний, в которой в полной мере гармонично соединились методы прикладной математики, кибернетики и новейшие информационные технологии, позволяющие хранить и в разумные сроки обрабатывать сверхбольшие объемы информации.

Следует заметить, что указанная гибридизация происходит не только между математическими и информационными технологиями, но и между различными математическими методами анализа данных. В частности, анализ результатов применения нейронных сетей в задачах прогнозирования состояния изучаемого объекта, как правило, осуществляется статистическими методами. И, наоборот, для решения традиционной статистической задачи регрессионного анализа может использоваться методология, основанная на эволюционном моделировании.

Принцип использования шаблонов. Наличие большого объема упорядоченных опытных данных позволяет существенно снизить влияние «проклятия среднего», характерного для большинства статистических методов обработки данных. При этом в качестве важнейшего инструментального средства DM используется описанная выше технология ассоциативной памяти. В частности, включение шаблонов в структуру аналитического запроса позволяет осуществлять ассоциативный поиск ситуаций аналогов и, на его основе, осуществлять прогноз нестационарного развития изучаемых процессов.

Перечисленные системные принципы образуют общую методологическую платформу, позволяющую выделить DM в качестве самостоятельного подкласса информационных технологий. Для построения более строгой, формализованной структуризации аналитических информационных технологий необходимо рассмотреть на содержательном уровне класс основных математических задач, решение которых составляет основу количественного DM анализа.

8. Заключение

В настоящее время аналитические информационные технологии нашли множество самых разнообразных практических приложений: в экономике, промышленности, торговле, системах здравоохранения, страхования, в различных областях, связанных с контролем и прогнозированием состояния сложных динамических систем. Однако наиболее интенсивное развитие данная методология нашла в сфере экономики, финансов и бизнеса [2, 4, 7, 8, 9]. Внедрение DM в область промышленных приложений пока что находится на ранней стадии и проявляется в создании виртуальных анализаторов, СППР должностных лиц и, разумеется, в использовании в системах стратегического звена управления (маркетинговые исследования, анализ экономических и финансовых ситуаций и т.п.).

В частности, системы DM используется при решении таких задач, как выявление скрытых закономерностей в финансовых данных, при разработке про-

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

гностических моделей (Lockheed), при верификация данных по курсам валют (Reuters), выявление новых потенциальных клиентов (Dickinson Direct) или определение зависимостей между основными показателями и характеристиками сегментов рынка при проведении маркетинговых исследований (Reader's Digest Canada), выявление счетов потенциально платежеспособных дебиторов (Internal Revenue Service), в различных задачах прогнозирования, например, при определении возможных невыплат в сделках с недвижимостью (Leeds) и многих других [15].

Инвестиции на внедрение комплексов DM достаточно велики, однако ожидаемый от их реализации выигрыш по данным ряда компаний может достигать 1000%. При этом инвестиции на внедрение данных технологий (при их правильном использовании) могут окупиться за несколько месяцев. В [8] приводятся данные ряда американских компаний, получивших экономический эффект от внедрения DM, который в 10–70 раз превысил первоначальные затраты, составлявшие от 350 до 750 тыс. US$. А сравнительно небольшой проект стоимостью в 20 млн. $ окупился всего за четыре месяца. По оценкам экспертов, более 95% компаний из списка Fortune 1000 уже внедрили системы хранилищ данных [8].

Как указывается в [9], интерес к DM не меньше, чем в свое время к вопросам искусственного интеллекта и систем автоматизированного проектирования. Однако, если первые два направления развивались преимущественно небольшими поставщиками, то в рассматриваемом нами случае разработку осуществляют такие гиганты, как IBM, AT&T и Microsoft. В частности, Microsoft уже выпустила DM-server, на котором реализована математика кластерного анализа и деревьев решений. Идеи OLAP реализованы не только в виде самостоятельного MS OLAP Server, но уже интегрированы в офисные электронные таблицы MS Excel.

Однако наряду с очевидным прогрессом в области DM, практическая реализация данной технологии выявила и ряд новых проблем, связанных с особенностями реализации сложных алгоритмических комплексов. Настоящая статья, по существу, является лишь введением в DM технологии, предложенные в качестве платформы для создания системы аналитического управления многомерными динамическими процессами.

Литература

1. *Аджиев В*. Mineset — визуальный инструмент аналитика. Открытые системы, 1997. № 3, С. 72–77.
2. *Бирюков А.* Системы принятия решений и хранилища данных // СУБД, 1997. № 4. С. 37–41.
3. *Гершберг А. Ф., Мусаев А. А., Нозик А. А., Шерстюк Ю. М.* Концептуальные основы информационной интеграции АСУ ТП нефтеперерабатывающего предприятия. СПб: Альянс-строй, 2003. 128 с.
4. *Дюк В., Самойленко А.* Data Mining: Учебный курс. СПб.: Питер, 2001. 366с.
5. *Заде Л.* Основы нового подхода к анализу сложных систем и процессов принятия решений // В кн.: Математика сегодня. М.: Знание, 1974. С. 5–48.
6. *Ивахненко А. Г., Зайченко Ю. Г., Димитров В. Д.* Принятие решений на основе самоорганизации. М.: Сов. радио, 1976. 280 с.
7. *Киселев М., Соломатин Е.* Средства добычи знаний в бизнесе и финансах // Открытые системы, 1997. № 4. С. 41–44.
8. *Кречетов Н., Иванов П*. Продукты для интеллектуального анализа данных // Computer Week, 1997. № 14–15. С. 32–39.
9. *Кривда Ш.* Раскопки сокрытых знаний. ЛАН, 1996. № 4. С 17–23.

10. *Львов В.* Создание систем поддержки принятия решений на основе хранилищ данных // СУБД, 1997. № 3. С. 30–40.
11. *Львович О.* Data Warehousing – выход из кризиса оперативного анализа // Read Me, 1998. № 6. С. 44–45, 66.
12. *Мусаев А. А*. Виртуальные анализаторы: концепция построения и применения в задачах управления непрерывными технологическими процессами. Автоматизация в промышленности, 2003. № 8. С. 28–33.
13. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / *С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин.* М.: Финансы и статистика, 1983. 471 с.
14. *Фогель Дж., Оуэнс Дж., Уолш Л.* Эволюционное моделирование и искусственный интеллект. М.: Мир, 1969. 219 с.
15. *Шапот М. И*нтеллектуальный анализ данных в системах поддержи принятия решений. Открытые системы, 1998. № 1. С. 30–35.
16. *Codd E. F., Codd S. B., Salley C. T.* Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd Associates, 1993. 18 p.
17. *W. H. Inmon.* Building the Data Warehouse. Wellesley, MA: QED Publishing Group, 1992.
18. *Musaev A.* Intelligent Control Systems for Refinery Technological Processes // Proceedings of conf. ICPI'02 (Intelligent computing for the petroleum industry, vol. 2. Mexico: 2002. P. 6–17.