

## ПОИСК, ВИЗУАЛИЗАЦИЯ СКРЫТЫХ ЗАВИСИМОСТЕЙ И ПРОГНОЗИРОВАНИЯ РАЗВИТИЯ СИТУАЦИЙ НА БАЗЕ ТЕХНОЛОГИЙ DATA MINING & KNOWLEDGE DISCOVERY



**М.М. Татур<sup>1</sup>**

Профессор кафедры электронных вычислительных машин БГУИР, доктор технических наук, профессор



**В.А. Пархименко<sup>1</sup>**

Заведующий кафедры экономики БГУИР, кандидат экономических наук, доцент,



**Д.И. Самаль<sup>1</sup>**

Заведующий кафедры электронных вычислительных машин БГУИР, кандидат технических наук, доцент,



**А.Т. Кусаикова<sup>2</sup>**

Докторант Евразийского национального университета имени Л.Н. Гумилева, магистр технических наук



**Л.П. Князева<sup>1</sup>**

Декан инженерно-экономического факультета, кандидат физико-математических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан

E-mail: [uladzimir.parkhimenka@gmail.com](mailto:uladzimir.parkhimenka@gmail.com)

**Проблема скрытых зависимостей.** Управление любым объектом или процессом (регионом, политической кампанией, военной операцией, финансовыми активами и т.п.) предполагает постоянное принятие решений, касающихся выбора последующих действий управляющего субъекта относительно этого объекта или процесса.

Анализ сложившейся практики управления показывает, что в подавляющем большинстве случаев управленческие решения высокого уровня принимаются исключительно исходя из опыта и интуиции ответственных работников, специалистов, экспертов с использованием доступной оперативной информации. В процессе принятия решений крайне редко применяются специальные системы поддержки, моделирования и прогнозирования, хотя на текущий момент разработан достаточный перечень подходов и инструментов, позволяющих если

не полностью автоматизировать процедуру принятия решения, то, по крайней мере, существенно повысить степень ее обоснованности за счет визуализации выявленных зависимостями. При этом большинством специалистов констатируется «разрыв» между огромным потоком данных, поступающих в аналитические центры от различных источников, с одной стороны, и применяемыми традиционными методами их обработки для принятия управленческих решений конечными потребителями (органами государственного управления, силовыми структурами, хозяйствующими субъектами), с другой.

В таких условиях высока вероятность того, что будут приниматься не самые лучшие, оптимальные решения, вследствие чего могут быть упущены потенциальные выгоды. Более того, своевременно не обнаруженные негативные тенденции в развитии ситуации, могут привести к катастрофическим для объекта управления или процесса результатам.

Общеизвестно, что на «развитие ситуаций» влияет большое число факторов, и наряду с явными и очевидными причинно-следственными связями, присутствуют скрытые, которые проявляются в виде «случайных, слабо-предсказуемых» событий. Порой человеческое подсознание опытных специалистов-практиков фиксирует и анализирует подобные закономерности, выдавая их за интуицию или «инсайты». Описать такие закономерности в аналитическом виде и гарантировать воспроизводимость не представляется возможным, но именно такие события представляют особую опасность, обычно их называют «стечением обстоятельств».

*Методология интеллектуального анализа данных.* Одними из первых разработки в области моделирования и прогнозирования были применены в анализе финансовых рынков, с целью предсказания котировок акций, валют и др. На начальной стадии использовался математический аппарат статистики. Со временем финансовые системы прогнозирования и поддержки решений все больше расширяли применяемый аппарат в сторону интеллектуализации анализа данных.

Data Mining & Knowledge Discovery (DM&KD) – это методология интеллектуального анализа данных, которая позволяет обнаружить скрытую и нетривиальную информацию, полезную для принятия управленческих решений[1]. DM&KD, как самостоятельное направление, сложилось в 90-х годах прошлого века и базируется на методах распознавания образов, математической статистики, искусственного интеллекта и других смежных областей информатики и информационных технологий.

Общая процедура применения DM&KD включает в себя четыре этапа:

- постановку задачи, т.е. описание прикладной (управленческой) проблемы из предметной области в терминах типовых задач;
- сбор и подготовку данных, т.е. формирование наборов данных (datasets) в том виде, который требуется для корректной работы методов DM&KD;
- процессинг, использование одного из существующих или

модифицированных алгоритмов для непосредственного анализа данных;

– оценку и интерпретацию результатов, т.е. проверку валидности результатов и их «перевод» на управленческий язык для принятия решения в рамках той прикладной проблемы, которая послужила исходной причиной для проведения анализа.

Здесь уместно упомянуть подход Cross Industry Standard Process for Data Mining (CRISP-DM), в рамках которого выделяются 6 этапов, однако, по мнению авторов, это не оказывает существенного влияния на понимание методологии DM&KD в целом. Важно другое, что множество прикладных задач может быть сведено к решению ограниченного числа типовых, для которых известны формальные алгоритмы реализации. В научной литературе обычно выделяют следующие типовые задачи DM&KD:

- кластеризацию (clustering), т.е. группировку объектов по схожести;
- ранжирование (ranking), т.е. упорядочение, выстраивание объектов по определенным (заданным) критериям;
- регрессию (regression), т.е. количественную оценку статистической связи между зависимыми и независимыми признаками объектов;
- классификацию (classification), т.е. построение алгоритма отнесения объекта к тому или иному классу;
- поиск ассоциативных правил (association rules learning), т.е. поиск всех значимых зависимостей между признаками объектов;
- прогнозирование.

*Проблема взаимодействия data scientists с управленцами-практиками.* Методы DM&KD потенциально позволяют выявить в большом массиве данных неявные, неочевидные, скрытые, но существенные зависимости, тем самым представить лицу, принимающему решение только необходимую информацию в объеме и форме, адекватных для принятия управленческих решений. Однако, в реальности присутствует субъективный фактор взаимодействия специалистов DM&KD и управленцев-практиков: первые, как правило, имеют техническое образование и ориентированы на этап алгоритмического процессинга данных, вторые, как правило, мыслят только в терминах предметной области, не понимая и не имея желания вникать в детали «математики и информатики». Зачастую проблема состоит в том, что ученые не могут объяснить доступным языком управленцам-практикам реальные выгоды от применения интеллектуального анализа данных, а те в свою очередь испытывают недоверие к современным средствам моделирования.

*Текущие исследования в сфере Data Mining & Knowledge.* Работы связанные с интеллектуальным анализом данных с использованием параллельных вычислений ведутся в БГУИР с 2000 г. За это время выполнен ряд проектов по созданию математических моделей и спецпроцессоров идентификации образов, нейроподобных параллельных вычислителей [2-6]. По данной тематике ведется активное сотрудничество с Евразийским национальным университетом им.

Л.Н.Гумилева (г.Астана). В 2011г. выполнена совместная НИР «Разработка научной концепции построения вычислительных систем с параллельной архитектурой и возможностью программирования прикладных задач для применения в различных областях», в 2015 г. – совместная НИР «Отображение алгоритмов семантической обработки информации на архитектурах компьютеров с массовым параллелизмом»

В рамках ГПНИ «Информатика» выполняется НИР «Разработка интеллектуальной вычислительной системы обработки больших объемов данных», которая закладывает теоретические основы будущих прикладных систем. На этом пути важным этапом является формирование междисциплинарных команд, включающих информатиков, экономистов, управленцев различных уровней.

В качестве первостепенных стоят следующие научно-практические задачи:

1 Формализовать типовые задачи принятия управленческих решений в конкретных прикладных областях (в экономике, при ликвидации чрезвычайных ситуаций, прогнозирование развития социальных ситуаций и т.п.) и представить в виде композиции формальных задач Data Mining & Knowledge Discovery.

2 Разработать программное обеспечение (прототип системы) с функциями выявления и визуализации скрытых зависимостей и прогнозирования развития ситуаций.

3 Оценить уровень вычислительной сложности реализуемых алгоритмов и реальные ограничения. Реализовать экспериментальный программно-аппаратный комплекс с использованием суперЭВМ и/или вычислительного кластера.

4 Провести апробацию экспериментальной системы поиска, визуализации скрытых зависимостей и прогнозирования развития ситуаций в реальных условиях.

#### Литература

[1]. J. Cios, W. Pedrycz, Roman W. Swiniarski, L. A. Kurgan. Data Mining: A Knowledge Discovery Approach. Springer, November 2007.

[2]. D. Adzinets, M. Lukashevich, S. Bairak, M. Tatur. Synthesis and Analysis of Classifiers Based on Generalized Model of Identification//Advances in intelligent and soft computing. Springer.- 2010, Vol. 71.-P.529-536.

[3]. M. Tatur Problem-Oriented Processors for the Solving of Classification Tasks Journal of Information, Control and Management Systems (Slovakia). 2013. vol.11 No.2 P.155-164.

[4]. В.П. Ивашенко, Н.Л.Вереник, А.И.Гирель, Е.Н.Сейткулов, М.М.Татур Представление семантических сетей, алгоритмы их организации и семантической обработки на вычислительных системах с массовым параллелизмом // 5-я Междунар. конф. «Открытые семантические технологии проектирования интеллектуальных систем». Минск, 2015. – С. 133-140.

[5]. М.М. Татур Особенности построения вычислителей интеллектуальной обработки данных // Минск. Информатика. 2015. - №1(45).-С.39-44.

А.И. Демидчук, Д.Ю. Перцев, Д.В. Кришталь, Д.И. Самаль, М.М. Татур Проявление закона Амдаля-Густавсона на примере реализации алгоритма k-средних Материалы НПК /Междунар. НПК «Big Data and Predictive Analytics. Использование Big Data для оптимизации бизнеса и информационных технологий», 16-19 июня 2015, Минск, Беларусь, С.151-154.