Российская Академия Наук
Сибирское Отделение
Институт Систем Информатики

На правах рукописи

ДЕМИН Александр Викторович

ЛОГИКО-ВЕРОЯТНОСТНЫЙ МЕТОД ИЗВЛЕЧЕНИЯ ЗНАНИЙ И ЕГО ПРИМЕНЕНИЕ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ И УПРАВЛЕНИЯ

Специальность 05.13.11 –
Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат диссертации на соискание ученой степени
кандидата физико – математических наук

Новосибирск 2008

Работа выполнена в Институте Систем Информатики
имени А.П. Ершова СО РАН

Научные руководители: Марчук Александр Гурьевич,
доктор физико-математических наук,
профессор

Витяев Евгений Евгеньевич,
доктор физико-математических наук

Официальные оппоненты: Загоруйко Николай Григорьевич,
доктор технических наук, профессор

Васючкова Татьяна Сергеевна,
кандидат физико-математических наук,
доцент

Ведущая организация: Институт автоматики и процессов
управления ДВО РАН

Защита состоится 26 декабря 2008 г. в 15 ч 00 мин на заседании диссертационного совета К003.032.01 в Институте систем информатики имени А. П. Ершова Сибирского отделения РАН по адресу:
630090, г. Новосибирск, пр. Лаврентьева, 6.

С диссертацией можно ознакомиться в читальном зале библиотеки ИСИ СО РАН (пр. Лаврентьева, 6)

Автореферат разослан 24 ноября 2008 г.

Ученый секретарь
диссертационного совета К003.032.01,
к.ф.–м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. В последние годы интенсивно развивается направление исследований Knowledge Discovery in Databases and Data Mining (KDD&DM). К настоящему времени разработано достаточно большое количество KDD&DM методов и реализующих их программных систем. Однако существующие на данный момент KDD&DM методы обладают рядом ограничений, не позволяющих извлечь из данных знания в полном объеме. Впервые эта проблема была сформулирована на международной конференции «Philosophies and Methodologies for Knowledge Discovery» (25 августа 2005, Копенгаген) и продолжает активно обсуждаться.

В результате анализа выясняются следующие основные ограничения существующих KDD&DM методов:

1. каждый метод ориентирован на работу с вполне определенными данными и позволяет использовать для анализа только часть информации, содержащейся в данных;
2. каждый метод явно или неявно обнаруживает на данных только вполне определенные типы закономерностей.

Тем самым становится актуальной проблема разработки такого метода извлечения знаний, который обладал бы достаточно универсальностью, чтобы использовать всю информацию, содержащуюся в данных, и обнаруживать любые виды закономерностей. Разработку такого метода целесообразно осуществлять, основываясь на некотором общем универсальном подходе к проблеме извлечения знаний. На данный момент единственным подходом, в котором в достаточно полном объеме разрешены все теоретические и философские аспекты для разработки такого метода, является реляционный подход к извлечению знаний, предложенный в работах Е.Е. Витяева и Б.Я. Ковалерчука. Данный подход использует язык логики первого порядка с вероятностной мерой для представления данных и формулировки различных видов закономерностей. Идеи реляционного подхода позволяют преодолеть большинство ограничений существующих KDD&DM методов. Однако реляционный подход не был сформулирован в виде конкретного метода извлечения знаний, пригодного для практической реализации в виде программной системы. Таким образом, задача разработки «универсального» метода остается актуальной.

В данной работе предложен логико-вероятностный метод извлечения знаний, основанный на идеях реляционного подхода. Данный метод реализован в виде программной системы и обладает достаточной универсальностью, чтобы использовать всю информацию, содержащуюся в данных, и обнаруживать любые виды закономерностей.

В настоящее время все чаще оказывается целесообразным использовать обнаруженные при помощи KDD&DM методов знания для осуществления прогноза в различных научно-прикладных задачах. Предложенный

метод обладает рядом преимуществ по сравнению с существующими KDD&DM методами, однако пока еще отсутствует необходимый опыт применения данного метода для решения задач прогноза. Поэтому важной задачей является разработка методов предсказания, использующих предложенный метод, и их исследование на примере решения реальных прикладных задач. В данной работе рассматривается применение разработанного метода извлечения знаний для построения прогнозирующих систем, предназначенных для решения ряда актуальных прикладных задач: 1) диагностика фолликулярного рака щитовидной железы (в медицине); 2) прогнозирования курсов ценных бумаг (в финансах); 3) распознавание сайтов связывания транскрипционных факторов (в биоинформатике).

В последнее время KDD&DM методы активно используются в алгоритмах самообучения адаптивных систем управления. Однако анализ показывает, что в сложившихся условиях постоянной тенденции к увеличению сложности и разнообразия задач управления, существующие подходы к построению адаптивных систем управления уже не способны обеспечить необходимый уровень управления и адаптации. Становится актуальной задача разработки универсальной системы управления, основанной на некоторых общих универсальных принципах управления и адаптации.

Разработку универсальной системы управления целесообразно проводить, отталкиваясь от общих концептуальных теорий и схем. Одной из таких общих концепций является теория функциональных систем, разработанная в 1930-70-х годах советским нейрофизиологом П.К. Анохиным. В данной работе предлагается новая модель универсальной адаптивной системы управления, которая включает схему управления на основе теории функциональных систем П.К. Анохина, алгоритм самообучения на основе разработанного логико-вероятностного метода извлечения знаний и возможность автоматического обнаружения новых подцелей.

Цель работы. Целью работы является разработка логико-вероятностного метода извлечения знаний из данных и его применение для создания прогнозирующих систем и разработки модели адаптивной системы управления. Для достижения этой цели необходимо:

1. Разработать метод обнаружения на данных полного множества закономерностей для заданного класса гипотез, выраженных в языке логики первого порядка, пополненного вероятностными оценками. Для решения этой задачи необходимо:

   а) разработать спецификацию фрагмента языка логики первого порядка для задачи обнаружения знаний в таблицах данных;

   б) разработать интерактивный способ задания классов гипотез для рассматриваемых данных;

в) разработать алгоритм обнаружения на данных полного множества закономерностей заданного класса.

2. Разработать метод предсказания, использующий множество вероятностных закономерностей, обнаруженных на данных.

3. Разработать и реализовать программную систему, позволяющую задавать класс обнаруживаемых закономерностей, извлекать из данных множество закономерностей заданного класса, использовать найденные закономерности для прогноза и принятия решений.

4. Используя разработанный метод и программную систему, провести ряд вычислительных экспериментов в исследованиях, связанных с решением следующих задач: 1) диагностика фолликулярного рака щитовидной железы; 2) прогнозирование финансовых временных рядов; 3) распознавание сайтов связывания транскрипционных факторов.

5. Разработать модель адаптивной системы управления, основанной на логико-вероятностном методе извлечения знаний из данных и обладающей возможностями самообучения, формирования иерархии целей и автоматического обнаружения новых подцелей.

6. Провести ряд вычислительных экспериментов по исследованию возможностей разработанной модели адаптивной системы управления.

Методы исследования. В работе использовались аппарат и методы математической логики, теории вероятности и математической статистики. Основным методом исследования являлось представление информации, содержащейся в данных, в виде множества отношений и операций в языке логики первого порядка и разработке специальных методов вычисления и обнаружения вероятностных закономерностей, выраженных в терминах этих отношений и операций. Разработка модели адаптивной системы управления осуществлялась путем формализации основных принципов организации и работы функциональных систем организма, изложенных в теории функциональных систем П.К. Анохина. При проектировании и разработке программных систем использовались методы объектно-ориентированного программирования, проектирования и анализа алгоритмов и программ.

Научная новизна. Следующие результаты, полученные в данной диссертации, раскрывают научную новизну работы:

1. Разработана спецификация фрагмента языка логики первого порядка для задачи обнаружения знаний в таблицах данных.

2. Разработан способ представления классов гипотез для рассматриваемых данных, позволяющий реализовать интерактивную систему задания классов гипотез.

3. Разработан алгоритм обнаружения вероятностных закономерностей, реализующий семантический вероятностный вывод.

4. Разработан метод предсказания и принятия решений, использующий множество вероятностных закономерностей, обнаруженных на данных.

5. Разработана архитектура программной системы, реализующей предложенный метод обнаружения закономерностей.

6. Разработана новая модель адаптивной системы управления, основанной на теории функциональных систем П.К. Анохина.

7. Разработан метод самообучения системы управления, основанный на методе обнаружения по истории деятельности системы множества вероятностных закономерностей, дающих максимально вероятный прогноз достижения цели.

8. Разработан метод автоматического обнаружения новых подцелей и формирования иерархии целей.

Практическая ценность. Разработана программная система «Discovery», предназначенная для обнаружения закономерностей на данных и осуществления прогноза на основе обнаруженных закономерностей. Разработанная система была успешно применена для решения следующих задач:

1. В медицине для диагностики фолликулярного рака щитовидной железы. Автор работы вместе с д.м.н. Т.Л. Полоз и проф. В.А. Шкурупием являются авторами патента на изобретение № 2293524 «Способ дифференциальной диагностики фолликулярной аденомы и фолликулярного рака щитовидной железы».

2. В финансах для прогнозирования курсов ценных бумаг.

3. В биоинформатике для распознавания сайтов связывания транскрипционных факторов. Результаты работы были использованы в работах по гранту РФФИ 05-07-90185в «Разработка научно-исследовательского комплекса программ “Expert Discovery” познания строения всех уровней геномной ДНК» и интеграционному проекту СО РАН № 115 «Разработка интеллектуальных информационных технологий генерации и анализа знаний для поддержки фундаментальных научных исследований в области естественных наук».

Результаты работы по разработке модели адаптивной системы управления использовались в работах по Программе Президента Российской Федерации поддержки научных школ 4413.2006.1.

Апробация работы. Основные результаты диссертационной работы докладывались и обсуждались на следующих научных конференциях: Меж-

дународная конференция «Мальцевские чтения» (Новосибирск, 2006); Конференция-конкурс «Технологии Microsoft в теории и практике программирования» (Новосибирск, 2007); Всероссийская конференция с международным участием «Знания – Онтологии – Теория» (Новосибирск, 2007).

Кроме того, полученные результаты обсуждались на семинарах в Институте систем информатики СО РАН и в Институте математики СО РАН.

Публикации. По материалам диссертации опубликовано 20 печатных работ, среди которых 10 статей в периодических изданиях и журналах, 3 статьи в трудах конференций, 6 тезисов докладов научных конференций, 1 патент на изобретение.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы. Объем работы составляет 171 страницу. Список литературы содержит 85 наименований. Работа включает 25 рисунков и графиков, полученных в результате расчетов на ЭВМ.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность диссертации, формулируются ее цели, характеризуется научная новизна и практическая ценность работы, приводится краткое описание содержания диссертации.

В первой главе представлен обзор существующих KDD&DM методов. Приводится сравнительный анализ числовых и логических методов в свете трех аспектов. Описываются основные идеи реляционного подхода к извлечению знаний. Рассматриваются ограничения существующих KDD&DM методов и указываются идеи реляционного подхода, позволяющие преодолеть эти ограничения. Анализируются нерешенные задачи реляционного подхода и формулируется задача разработки метода извлечения знаний и метода предсказания, основанных на идеях реляционного подхода.

Во второй главе приводится описание разработанного логико-вероятностного метода извлечения знаний и метода предсказания.

Раздел 2.1 посвящен описанию логико-вероятностного метода извлечения знаний.

В разделе 2.1.1 вводится спецификация фрагмента многосортного языка логики первого порядка для задачи извлечения знаний из таблиц данных и вводится определение понятия класса гипотез в терминах введенного языка.

Предполагается, что исходные данные представлены в виде таблицы значений $\boldsymbol{D}$, строки которой соответствуют объектам, а колонки – признакам объектов.

Зафиксируем фрагмент многосортного языка логики первого порядка с сигнатурой $\Sigma^* = \langle \mathbb{S}, \mathbb{P}, \mathbb{F} \rangle$, где

- $\mathbb{S} = \{s_{obj}, s_{\mathbb{R}}\}$ – множество сортов, s_{obj} – сорт объектов таблицы D, $s_{\mathbb{R}}$ – сорт действительных чисел;
- $\mathbb{P}$ – множество предикатных символов, все аргументы которых имеют сорт $s_{\mathbb{R}}$;
- $\mathbb{F} = \mathbb{F}_1 \cup \mathbb{F}_2 \cup \mathbb{F}_3$ – множество функциональных символов, где

  $\mathbb{F}_1$ – множество функциональных символов, аргументы которых имеют сорт s_{obj} или $s_{\mathbb{R}}$, а значения – сорт s_{obj};

  $\mathbb{F}_2$ – множество функциональных символов, таких, что все символы из $\mathbb{F}_2$ одноместные, их аргументы сорта s_{obj}, а значения – сорта $s_{\mathbb{R}}$.

  $\mathbb{F}_3$ – множество функциональных символов, аргументы и значения которых имеют сорт $s_{\mathbb{R}}$;

Введем многосортную алгебраическую систему $\mathcal{A}^* = \langle \{D, \mathbb{R}\}, \pi \rangle$ сигнатуры Σ^*, где D – таблица данных, являющаяся носителем сорта s_{obj}, $\mathbb{R}$ – множество действительных чисел, являющееся носителем сорта $s_{\mathbb{R}}$, π – интерпретация сигнатуры Σ^* на $\{D, \mathbb{R}\}$.

Введем понятия *шаблона термов* и *шаблона предикатов*.

Шаблон термов – это пара $Tf = \langle f, \{\Psi_1, ..., \Psi_n\} \rangle$, где $f \in \mathbb{F}_3$, n – арность символа f, $\Psi_i = \{t_1, ..., t_{m_i}\}$, $i = 1, ..., n$, t_k $(k = 1, ..., m_i)$ – терм сорта $s_{\mathbb{R}}$ языка первого порядка, такой, что любая переменная, входящая в терм t_k имеет сорт s_{obj}. Шаблон терма $Tf = \langle f, \{\Psi_1, ..., \Psi_n\} \rangle$ определяет множество всех термов вида $f(t_1, ..., t_n)$, где $t_i \in \Psi_i$, $i = 1, ..., n$. Будем обозначать через $[Tf]$ – множество термов, определяемое шаблоном Tf.

Шаблон предикатов – это пара $Tp = \langle P, \{\Theta_1, ..., \Theta_n\} \rangle$, где $P \in \mathbb{P}$, n – арность символа P, $\Theta_i = \{Tf_1, ..., Tf_{m_i}\}$, $i = 1, ..., n$, Tf_k $(k = 1, ..., m_i)$ – некоторый шаблон терма. Шаблон предиката $Tp = \langle P, \{\Theta_1, ..., \Theta_n\} \rangle$ определяет множество всех атомарных формул вида $P(t_1, ..., t_n)$, где $t_i \in [Tf_1] \cup ... \cup [Tf_{m_i}]$, $i = 1, ..., n$, $Tf_k \in \Theta_i$, $k = 1, ..., m_i$. Будем обозначать через $[Tp]$ – множество атомарных формул, определяемых шаблоном Tp.

Теперь, используя понятие шаблона предикатов, мы можем определить понятие *класса гипотез*.

Класс гипотез – это пара $Th = \langle \{Tp_1, ..., Tp_m\}, P_0^{\varepsilon} \rangle$, где Tp_i – шаблоны предикатов, P_0^{ε} – целевая литера, P_0 – атомарная формула, $\varepsilon \in \{0, 1\}$ – обозначает наличие отрицания формулы.

Класс гипотез $Th = \langle \{Tp_1, ..., Tp_m\}, P_0^{\varepsilon} \rangle$ определяет множество правил вида

$$\forall i_1,...,i_k (P_1^{\varepsilon} \& ... \& P_n^{\varepsilon} \to P_0^{\varepsilon}),\ k \geq 0 \quad (1)$$

где $i_1,...,i_k$ – переменные сорта s_{obj}, $P_i \in [Tp_1] \cup ... \cup [Tp_m]$, $i = 1,...,n$.

Формулы вида (1) будем записывать в более удобном виде

$$A_1 \& ... \& A_n \to A_0,\ A_i = P_i^{\varepsilon},\ i = 0,...,n, \quad (2)$$

который получается введением вместо кванторов всеобщности и связанных ими переменных индивидных констант $z_1,...,z_k$.

Данное определение понятия класса гипотез позволяет реализовать интерактивную систему задания классов гипотез, проверяемых на данных.

В разделе 2.1.2 вводяться понятия эксперимента, вероятности на множестве экспериментов и вероятностной закономерности.

Подправилом правила $R_1 = A_1 \& ... \& A_m \to A_0$ будем называть любое правило $R_2 = A_{i1} \& ... \& A_{ik} \to A_0$, такое что $\{A_{i1},...,A_{ik}\} \subset \{A_1,...,A_m\}$.

Вероятностной закономерностью называется правило $A_1 \& ... \& A_m \to A_0$, удовлетворяющее следующим условиям:

1. условная вероятность правила определена, т.е. $p(A_1 \& ... \& A_m) > 0$;
2. условная вероятность правила строго больше условных вероятностей каждого из его подправил.

Сильнейшей вероятностной закономерностью называется вероятностная закономерность, которая не является подправилом никакой другой вероятностной закономерности.

В разделе 2.1.3 описывается метод обнаружения вероятностных закономерностей на выборке данных. Приводятся статистические критерии, позволяющие проверить выполнимость вероятностных неравенств, указанных в определении понятия вероятностной закономерности.

В разделе 2.1.4 приводится алгоритм поиска вероятностных закономерностей. Предложенный алгоритм основан на семантическом вероятностном выводе, который был определен в работах Е.Е. Витяева в рамках реляционного подхода к извлечению знаний.

Семантическим вероятностным выводом (СВВ) некоторой сильнейшей вероятностной закономерности R_n называется такая последовательность вероятностных закономерностей $R_1, R_2,..., R_n$, $R_i = A_1 \& ... \& A_k \to A_0$, $i = 1,2,...,n, n \geq 1$, что правило R_i является подправилом правила R_{i+1}, $p(R_{i+1}) > p(R_i), i = 1,2,...,n-1$.

В работе выводится, что для нахождения множества всех закономерностей $Reg(Th)$, которые могут быть получены на основе класса гипотез Th, достаточно обнаружить множество базовых закономерностей $Base(Th) \subseteq Reg(Th)$. Все остальные закономерности могут быть найдены путем последовательного уточнения всех базовых закономерностей с проверкой выполнимости условий для вероятностных закономерностей. Основы-

ваясь на данном свойстве, в работе был предложен алгоритм направленного перебора правил вида (2), позволяющий обнаружить множество $Reg(Th)$.

В разделе 2.2 описывается метод предсказания, использующий множество вероятностных закономерностей, обнаруженных на данных.

В разделе 2.2.1 приводится общая формулировка метода предсказания. Пусть $Reg(Th)$ – множество закономерностей, найденное алгоритмом поиска закономерностей на основе заданного класса гипотез $Th = \langle \{Tp_1, ..., Tp_m\}, P_0^{\varepsilon} \rangle$. Пусть $P(Th) = [Tp_1] \cup ... \cup [Tp_m]$.

Пусть нам дано множество объектов $\mathbb{D}$ и некоторый новый объект b. Предположим, что значения части атомарных формул $P^{И} \subset P(Th)$ на объектах $\mathbb{D}$ и b нам известны. Требуется, используя закономерности из $Reg(Th)$, по известным значениям атомарных формул из $P^{И}$ предсказать неизвестные значения остальных атомарных формул из $P^{П} = P(Th) \setminus P^{И}$. Таким образом, задача предсказания сводится к тому, чтобы по модели $pr_0 = \langle \mathbb{D}, P(Th) \rangle$ и модели $pr^{И} = \langle \mathbb{D} \cup b, P^{И} \rangle$ восстановить модель $pr = \langle \mathbb{D} \cup b, P(Th) \rangle$.

Методом предсказания называется алгоритм $AP : \langle Reg(Th), pr_0, pr^{И} \rangle \to \nu$, преобразующий каждую тройку $\langle Reg(Th), pr_0, pr^{И} \rangle$ в частично определенное отображение $\nu : PS \to [0,1]$, сопоставляющее каждой модели $pr \in PS$ некоторую оценку её вероятности $\nu(pr)$. Где PS – множество всех возможных восстановлений модели $pr = \langle \mathbb{D} \cup b, P(Th) \rangle$.

В разделе 2.2.2 описывается способ определения метода предсказания в общем случае. Все закономерности из $Reg(Th)$ можно разбить на три группы: 1) Reg_1 – множество закономерностей, которые содержат только одну индивидную константу; 2) Reg_2 – множество закономерностей, заключение которых содержит только одну индивидную константу, а посылка содержит по крайней мере две различные индивидные константы; 3) Reg_3 – множество закономерностей, у которых в заключении есть хотя бы две различные индивидные константы. $Reg(Th) = Reg_1 \cup Reg_2 \cup Reg_3$, $Reg_i \cap Reg_j = \varnothing$, $i \neq j$.

Для каждой группы закономерностей Reg_1, Reg_2 и Reg_3 в работе предложены способы определения прогнозов для их закономерностей.

Чтобы оценить вероятность $\nu(pr)$ некоторой модели $pr \in PS$, метод прогноза должен основываться на прогнозах всех закономерностей из $Reg(Th)$. Таким образом, для построения метода предсказания необходимо задать частично определенную функцию $\lambda : \{\langle pr, Reg(pr) \rangle\} \to [0,1]$, которая для каждой модели $pr \in PS$ на основании множества закономерностей $Reg(pr) \subseteq Reg(Th)$, предсказывающих данную модель, вычисляет оценку вероятности $\nu(pr)$. Тогда $\nu(pr) = \lambda(\langle pr, Reg(pr) \rangle)$.

В работе приводится несколько примеров задания функции λ.

В разделе 2.2.3 рассматривается метод предсказания, основанный на оценке максимальной вероятности. Для всех трех видов закономерностей Reg_1, Reg_2 и Reg_3 подсчитываются вероятностные оценки, необходимые для получения предсказаний. Полученные оценки затем используются для определения итоговой оценки вероятности $\nu(pr)$. В работе приводятся доказательства ряда утверждений, позволяющих установить отображение ν и получить оценки вероятности для итогового предсказания $\nu(pr)$.

В разделе 2.2.4 описывается общий механизм принятия решений на основе предсказания. Пусть нам известно множество вариантов решений $S = \{s_1, ..., s_n\}$, и для каждого решения $s \in S$ можно указать множество моделей $PS(s) \subset PS$, для которых должно быть выбрано данное решение. Будем называть множества $PS(s_i)$, $i = 1, ..., n$ *вариантами исхода*.

Для построения метода принятия решений, необходимо задать *функцию оценки вероятностей исходов* $\kappa : \{PS(s_1), ..., PS(s_n)\} \to [0, 1]$, которая для любого исхода $PS(s) \subset PS$ вычисляет результирующую оценку вероятности данного исхода, основываясь на оценках $\nu(pr)$ всех моделей $pr \in PS(s)$, входящих в этот исход $PS(s)$. В работе приводится несколько примеров определения функции κ.

Методом принятия решений называется функция $Dec : \{\langle \kappa(PS(s_1)), ..., \kappa(PS(s_n)) \rangle\} \to \{S \cup \varnothing\}$, которая для каждого набора оценок вероятностей исходов $\langle \kappa(PS(s_1)), ..., \kappa(PS(s_n)) \rangle$ возвращает один выбранный вариант решения $s \in S$ либо $\varnothing$, если решение не может быть принято.

Рассматривается один из возможных вариантов определения функции Dec, основанный на расчете показателей согласованности прогнозов.

В третьей главе описывается программная система «Discovery», реализующая разработанный метод извлечения знаний, и ее применение для решения актуальных задач в медицине, финансах и биоинформатике.

В разделе 3.1 приводится описание программной системы «Discovery». Программа дает возможность пользователю задавать класс гипотез для обнаружения закономерностей, осуществлять поиск закономерностей и использовать найденные закономерности для прогноза и принятия решений.

Раздел 3.2 описывает применение системы «Discovery» для решения задачи диагностики фолликулярной опухоли щитовидной железы. Совершенствование дооперационной диагностики фолликулярной аденомы и рака является актуальной задачей, поскольку, как показывает опыт, в настоящее время совпадение цитологических и окончательных гистологических диагнозов не превышает 56%. Помимо повышения точности цитологической диагностики, особый интерес также вызывает предварительная диагностика заболевания по данным УЗИ обследования. При помощи системы «Discovery» мы провели два исследования возможности диагностики фолликуляр-

ной опухоли, основанные на использовании данных цитологического анализа и данных УЗИ обследования.

Исходными данными для первого исследования послужили цитологические препараты 197 больных (86 раков и 111 аденом), которые были проанализированы по 30 цитологическим признакам. Во втором исследовании были использованы данные об УЗИ обследовании 170 больных (70 раков и 100 аденом), которые были проанализированы по 9 УЗИ признакам.

В обоих исследованиях тестирование системы методом скользящего контроля показало, что обнаруженные системой правила позволяют осуществлять диагностику с точностью до 96 %. Мы провели сравнение точности прогнозов системы «Discovery» с прогнозами, полученными при помощи нейронной сети. На скользящем контроле нейронная сеть показала точность, равную 91% на цитологических данных и 86% – на данных УЗИ. Таким образом, система «Discovery» показала более высокую точность прогнозов, чем нейронные сети.

В разделе 3.3 описывается применение системы «Discovery» для прогнозирования финансовых временных рядов. В данном исследовании рассмотрено применение системы «Discovery» для разработки торговой системы, основанной на предсказании увеличения или уменьшения котировок индекса S&P500 через пять дней по отношению к текущей цене.

Для обнаружения вероятностных закономерностей использовались гипотезы следующего вида:

$$\forall t_1 \exists t_2,...,t_m : (ext(t_1,...,t_m) =< \delta_1,...,\delta_m >) \& (c(t_i) < c(t_j)) \& ... \& (c(t_k) < c(t_l)) \rightarrow T,$$

где $i, j, k, l \in \{1,...,m\}$, $t_1 > ... > t_m$, $c(t_i)$ – значение временного ряда в точке t_i

Предикат $ext(t_1,...,t_m) =< \delta_1,...,\delta_m >$, $\delta_i \in \{-1;1\}$, $i = 1,...,m$ определяет условие, проверяющее, являются ли точки t_i, $i = 1,...,m$ локальными минимумами (при $\delta_i = -1$) или локальным максимумами (при $\delta_i = 1$).

T – целевой предикат, $T = (c(t_1) < c(t_1 + 5))$ или $T = (c(t_1) > c(t_1 + 5))$.

Данные гипотезы проверяют, сформировалась ли в прошлом временного ряда определенная фигура из локальных минимумов и максимумов, и если такая фигура найдена, делают прогноз на пять дней вперед.

Тестирование разработанной нами торговой системы осуществлялось методом скользящего контроля на временном интервале из 2065 торговых дней. Чтобы оценить эффективность работы системы «Discovery», мы также провели сравнение данной системы с методом скользящей линейной регрессии и комитетом нейронных сетей. В результате тестирования система «Discovery» показала лучшие результаты как по проценту правильных прогнозов, так и по показателям финансовой эффективности. Процент прибыльных сделок у системы «Discovery» достигает 69% по сравнению с 57% у нейронных сетей и 41% у линейной регрессии.

В разделе 3.4 описывается применение системы «Discovery» в биоинформатике для распознавания сайтов связывания транскрипционных факто-

ров (ССТФ). Задача обнаружения ССТФ очень важна для понимания механизмов регуляции транскрипции генов. Традиционным методом предсказания ССТФ является позиционная весовая матрица (PWM), основанная на предположении о независимости нуклеотидных позиций. Однако точность предсказания может быть увеличена за счет учета взаимосвязи между нуклеотидами. При помощи системы «Discovery» мы исследовали три семейства транскрипционных факторов: SREBP, EGR1 и HNF4.

Мы использовали систему «Discovery» для поиска вероятностных закономерностей следующего вида:

$$(\boldsymbol{Pos}_1(\boldsymbol{s}) = \boldsymbol{N}_1)^{\varepsilon 1} \,\&\, (\boldsymbol{Pos}_2(\boldsymbol{s}) = \boldsymbol{N}_2)^{\varepsilon 2} \,\&\, \ldots \,\&\, (\boldsymbol{Pos}_k(\boldsymbol{s}) = \boldsymbol{N}_k)^{\varepsilon k} \rightarrow (\boldsymbol{TFBS}(\boldsymbol{s}) = \boldsymbol{1}),$$

где $(\boldsymbol{Pos}_i(\boldsymbol{s}) = \boldsymbol{N})^{\varepsilon}$ – предикат, означающий, что в позиции $\boldsymbol{i}$ последовательности нуклеотидов $\boldsymbol{s}$ находиться (при $\boldsymbol{\varepsilon = 0}$) или не находиться (при $\boldsymbol{\varepsilon = 1}$) символ $\boldsymbol{N} \in \{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{G}, \boldsymbol{T}\}$; $(\boldsymbol{TFBS}(\boldsymbol{s}) = \boldsymbol{1})$ – целевой предикат, означающий, что последовательность нуклеотидов $\boldsymbol{s}$ является сайтом связывания.

Качество распознавания ССТФ оценивалось в сравнении с качеством распознавания оптимальной PWM, путем осуществления специальной процедуры скользящего контроля. Полученные результаты для всех трех семейств транскрипционных факторов показывают, что ошибка перепредсказания, соответствующая системе «Discovery», меньше таковой для PWM при каждом уровне ошибки недопредсказания. Таким образом, учет нуклеотидных позиций не по отдельности, а во взаимосвязи, позволил системе «Discovery» увеличить точность распознавания ССТФ.

В четвертой главе описывается разработанная модель адаптивной системы управления.

В разделе 4.1 представлен обзор существующих подходов и проблем построения адаптивных систем управления. Представлен анализ существующих в рамках направления исследований «Адаптивное поведение» подходов к построению систем управления с точки зрения возможности организации самообучения и адаптации.

В разделе 4.2 приведено описание разработанной модели адаптивной системы управления аниматом (искусственным организмом).

В разделе 4.2.1 описываются основные идеи теории функциональных систем П.К. Анохина, которая является концептуальной основой для разработанной модели адаптивной системы управления.

В разделе 4.2.2 описывается общая архитектура системы управления. Предполагается, что анимат имеет некоторый набор сенсоров $\mathbb{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_n\}$ и набор возможных действий $\mathbb{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_m\}$. Информации об окружающей среде подается в систему управления в виде набора *сенсорных предикатов* $\mathbb{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_k\}$, описывающих текущее состояние сенсоров.

Архитектура системы управления представляет собой иерархию функциональных систем, при которой функциональные системы верхнего

уровня ставят цели системам нижнего уровня. Отдельная функциональная система $\boldsymbol{ФС}$ определяется следующим набором: $\boldsymbol{ФС} =< \boldsymbol{PG}, \mathbb{G}, \boldsymbol{PR} >$.

$\boldsymbol{PG}$ – *предикат-цель*, описывающий цель, достижение которой является задачей данной функциональной системы, $\boldsymbol{PG} = \boldsymbol{P_1} \,\&\, \boldsymbol{P_2} \,\&\, ... \,\&\, \boldsymbol{P_l}$, $\boldsymbol{P_i} \in \mathbb{P}$.

$\mathbb{G}$ – множество предикатов-целей, соответствующих функциональным системам, подчиненным данной системе.

$\boldsymbol{PR}$ – множество закономерностей, принадлежащих данной функциональной системе и имеющих вид:

$$\boldsymbol{P_1} \,\&\, ... \,\&\, \boldsymbol{P_n} \,\&\, \boldsymbol{PG_1} \,\&\, ... \,\&\, \boldsymbol{PG_m} \,\&\, \boldsymbol{A_1} \,\&\, ... \,\&\, \boldsymbol{A_k} \rightarrow \boldsymbol{PG}, \qquad (3)$$

где $\boldsymbol{P_i} \in \mathbb{P}$, $\boldsymbol{PG_i} \in \mathbb{G}$, $\boldsymbol{A_i} \in \mathbb{A}$. Эти закономерности предсказывают, что если из ситуации $\boldsymbol{P_1}, ..., \boldsymbol{P_n}$ достичь цели $\boldsymbol{PG_1}, ..., \boldsymbol{PG_m}$, и затем выполнить действия $\boldsymbol{A_1}, ..., \boldsymbol{A_k}$, то с некоторой вероятностью $\boldsymbol{p}$ будет достигнута цель $\boldsymbol{PG}$.

В разделе 4.2.3 описывается работа отдельной функциональной системы. Задачей каждой функциональной системы $\boldsymbol{ФС} =< \boldsymbol{PG}, \mathbb{G}, \boldsymbol{PR} >$ является нахождение наиболее оптимального способа достижения цели $\boldsymbol{PG}$. Для этого анализируется множество закономерностей $\boldsymbol{PR}$ и для каждой закономерности $\boldsymbol{R} = \boldsymbol{P_1} \,\&\, ... \,\&\, \boldsymbol{P_n} \,\&\, \boldsymbol{PG_1} \,\&\, ... \,\&\, \boldsymbol{PG_m} \,\&\, \boldsymbol{A_1} \,\&\, ... \,\&\, \boldsymbol{A_k} \rightarrow \boldsymbol{PG}$, применимой в текущей ситуации, рассчитывается оценка вероятности достижения цели $\boldsymbol{f(R)}$ по формуле: $\boldsymbol{f(R) = p(R) \cdot f(PG_1) \cdot f(PG_2 \mid PG_1) \cdot ... \cdot f(PG_m \mid PG_{m-1})}$, где $\boldsymbol{f(PG_1)}$ – оценка вероятности достижения подцели $\boldsymbol{PG_1}$ из текущей ситуации; $\boldsymbol{f(PG_i \mid PG_{i-1})}$ – оценка вероятности достижения подцели $\boldsymbol{PG_i}$ после достижения подцели $\boldsymbol{PG_{i-1}}$. Расчет оценок $\boldsymbol{f(PG_1)}, \boldsymbol{f(PG_2 \mid PG_1)}, \ldots, \boldsymbol{f(PG_m \mid PG_{m-1})}$ осуществляется аналогичным образом путем отправки запросов соответствующим функциональным системам. На основании полученных оценок выбирается закономерность $\boldsymbol{R_{best}}$, имеющая максимальную оценку вероятности, которая и будет определять оптимальные действия по достижению цели.

В разделе 4.2.4 описывается работа системы управления в целом. Каждый такт работы системы включает два этапа: 1) формирование плана действий; 2) выполнение действий и контроль качества. Во время первого этапа происходит запрос к функциональной системе, находящейся на вершине иерархии, о достижении доминирующей цели. Ответом будет являться максимально вероятный прогноз достижимости цели и соответствующий ему план действий. Во время второго этапа система в соответствии с выбранным планом передает управление функциональным системам, осуществляющим выполнение действий по достижению цели.

В разделе 4.2.5 описывается метод самообучения системы управления, который позволяет для каждой функциональной системы $\boldsymbol{ФС} =< \boldsymbol{PG}, \mathbb{G}, \boldsymbol{PR} >$ обнаруживать множество закономерностей $\boldsymbol{PR}$ на множестве данных истории деятельности анимата. Метод самообучения основан на адаптированном варианте разработанного логико-вероятностного метода обнаружения закономерностей, учитывающего специфику правил вида (3).

В разделе 4.2.6 представлен метод автоматического формирования новых подцелей. Предложенный метод основан на следующем определении подцели: подцелью является ситуация, достижение которой значительно увеличивает вероятность достижения вышестоящей цели, и последующие действия из этой ситуации не могут быть определены однозначно.

Для выявления подцелей у каждой функциональной системы $\boldsymbol{ФС} =< \boldsymbol{PG}, \mathbb{G}, \boldsymbol{PR} >$ анализируется множество ее правил $\boldsymbol{PR}$. Если ситуация, описываемая предикатом $\boldsymbol{PG}_{New} = \boldsymbol{P}_1 \& ... \& \boldsymbol{P}_k$, $\boldsymbol{P}_i \in \mathbb{P}$ удовлетворяет определенным условиям, то она будет являться подцелью и для нее создается новая функциональная система $\boldsymbol{ФС}_{New}$, находящаяся ниже по иерархии системы $\boldsymbol{ФС}$ и реализующая достижение этой подцели.

В разделе 4.3 описываются эксперименты по оценке эффективности разработанной системы управления и ее сравнение с другими подходами.

Для исследования предложенной системы управления было поставлено два эксперимента по решению классической задачи фуражирования и ее усложненного варианта. Эффективность разработанной системы управления оценивалась в сравнении с системами, основанными на различных реализациях обучения с подкреплением (Reinforcement Learning).

В первом эксперименте агент, двигаясь по плоскости, должен был научиться эффективно собирать пищевые объекты. Результаты первого эксперимента показали, что все системы управления способны научиться решать классическую задачу фуражирования. Однако предложенная система обучается значительно быстрее (примерно в 12 раз).

Второй эксперимент отличался от первого наличием еще одного объекта, условно названного «таблетка». Чтобы съесть еду анимат должен был предварительно найти таблетку. Таким образом, в данной задаче можно выделить очевидную подцель – найти таблетку, и тем самым протестировать способность автоматического обнаружения новых подцелей.

Результаты второго эксперимента показали значительное преимущество разработанной системы в возможностях самообучения и адаптации по сравнению системами Reinforcement Learning. Метод обнаружения подцелей позволил разработанной системе быстро обнаружить новую подцель – найти таблетку, в результате чего система научилась эффективно решать усложненную задачу фуражирования. Системы управления на основе Reinforcement Learning оказались не способны решить эту задачу.

В заключении приведены основные результаты работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

В диссертационной работе были получены следующие основные результаты:

1. Разработан метод обнаружения на данных полного множества закономерностей для заданного класса гипотез.

2. Разработан метод предсказания и принятия решений, использующий множество вероятностных закономерностей, обнаруженных на данных.

3. Разработана и реализована программная система «Discovery», позволяющая задавать вид обнаруживаемых закономерностей, извлекать из данных множество закономерностей заданного вида, использовать найденные закономерности для прогноза и принятия решений.

4. Получены положительные результаты о возможности применения разработанного метода и системы «Discovery» для решения задач: 1) диагностики фолликулярного рака щитовидной железы; 2) прогнозирования финансовых временных рядов; 3) распознавания сайтов связывания транскрипционных факторов.

5. Разработана модель адаптивной системы управления, обладающая возможностями самообучения, формирования иерархии целей и автоматического обнаружения новых подцелей. Проведены вычислительные эксперименты по исследованию возможностей разработанной системы управления и сравнению с другими подходами.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССРЕТАЦИИ

1. Витяев Е.Е., Ковалерчук Б.К., Федотов А.М., Барахнин В.Б., Белов С.Д., Дурдин Д.С., Демин А.В. Обнаружение закономерностей и распознавание аномальных событий в потоке данных сетевого трафика // Вестник НГУ, серия: Информационные технологии. – 2008. – Т. 6. – Вып. 2. – С. 57-68.
2. Khomicheva I.V., Demin A.V., Vityaev E.E. Transcription Factor Binding Site Discovery by the Probabilistic Rules. Proceedings of the 2nd workshop in data mining in functional genomics and proteomics. The 18th European conference on machine learning and the 11th European conference on principles and practice of knowledge discovery in databases. Warsaw, Poland, September 17-21, 2007. – 2007. – p.104-109.
3. Демин А.В., Витяев Е.Е., Полоз Т.Л., Реализация универсальной системы извлечения знаний «Discovery» и ее применение в задачах медицинской диагностики // Труды Всероссийская конференция с международным участием «Знания – Онтологии – Теории». – Новосибирск, 2007. – Т. 1. – С. 63–70.
4. Демин А.В., Реализация универсальной версии системы «DISCOVERY» // Тез. докл. конференции-конкурса «Технологии Microsoft в теории и практике программирования», Новосибирск, 24–26 февраля 2007. – 2007.

– С. 106–108.
5. Demin A.V., Vityaev E.E., Animat control system based on semantic probabilistic inference // Bull. Nov. Comp. Center, Computer Science, 24 (2006). – NCC Publisher, 2006. – p. 57-72.
6. Демин А.В., Витяев Е.Е., Разработка модели адаптивного поведения анимата на основе семантического вероятностного вывода // Молодая информатика. – Новосибирск, 2006. – Выпуск 2. – С. 121–133.
7. Демин А.В., Витяев Е.Е. Реализация модели анимата на основе семантического вероятностного вывода // VIII Всероссийская научно-техническая конференция «Нейроинформатика-2006». – Москва, 2006. – Т. 2. – С. 16-24.
8. Демин А.В., Витяев Е.Е. Система управления аниматом, основанная на теории функциональных систем П.К.Анохина // Анализ структурных закономерностей. – Новосибирск, 2005. – Вып. 174: Вычислительные системы. – с. 93-113.
9. Shapiro N.A., Poloz T.L., Shkurupij V.A., Tarkov M.S., Poloz V.V., Demin A.V. Application of Artificial Neural Network for Classification of Thyroid Follicular Tumors // Anal. Quant. Cytol. Histol. – 2007. – V. 29, – P. 122-119.
10. Полоз Т.Л., Демин А.В., Шкурупий В.А. Патент на изобретение № 2293524 «Способ дифференциальной диагностики фолликулярной аденомы и фолликулярного рака щитовидной железы».
11. Полоз Т.Л., Шкурупий В.А., Полоз В.В., Демин А.В. Результаты количественного цитологического анализа строения фолликулярных опухолей щитовидной железы с помощью компьютерных и нейросетевых технологий // Вестник Российской Академии Медицинских Наук. – Москва, 2006. – № 8. – С.7-10.
12. Демин А.В., Использование нейросетевых технологий для диагностики фолликулярного рака щитовидной железы. // Студент и научно-технический прогресс: тез. докл. XLII международной научной студенческой конференции, Новосибирск, 2004. – 2004. – С. 240-241.
13. Полоз Т.Л., Демин А.В. Опыт применения нейросетевых технологий для цитологической диагностики некоторых заболеваний щитовидной железы // ZEISS СЕГОДНЯ. – 2004. – № 24. – С. 4.
14. Полоз В.В, Полоз Т.Л., Демин А.В. Морфометрия в нейросетевой технологии для цитологической диагностики фолликулярных пролифератов щитовидной железы // Новости клинической цитологии России: тез. докл. VI междунар. семинара патологоанатомов и клинических цитологов, Нижний Новогород, 2004. – 2004. – Т.8 – № 1-2. – С. 56.
15. Пупышева Т.Л., Демин А.В. Применение нейросетевых технологий в цитологической диагностике фолликулярных пролифератов щитовидной железы // Новости клинической цитологии России: тез. докл. V всероссийского съезда Ассоциации клинических цитологов России, Псков, 2003. – 2003. – Т. 7. – № 1-2. – С. 70.

16. Пупышева Т.Л., Демин А.В. Использование ассоциативных правил для решения задач цитологической диагностики фолликулярных новообразований щитовидной железы // Системный анализ и управление в биомедицинских системах. – 2003. – Т.2. – №2. – С.116-119.
17. Пупышева Т.Л., Демин А.В. Возможности нейросетевой технологии при дифференциальной диагностике фолликулярных пролифератов щитовидной железы // Новости клинической цитологии России. – 2003. – Т.7. – № 1-2. – С. 4-7.
18. Пупышева Т.Л., Демин А.В. Цитологическая диагностика фолликулярных неоплазий щитовидной железы с использованием нейросетевых технологий // Актуальные вопросы онкологии: Современные проблемы морфологической диагностики опухолей: тез. докл. междунар. научно-практ. конф., Иркутск, 2003. – 2003. – С.59.
19. Пупышева Т.Л., Демин А.В. Возможности использования искусственных нейронных сетей в цитологической диагностике фолликулярных неоплазий щитовидной железы // Новые направления и разработки в онкоморфологии: тез. докл. IX науч. конф., Москва, 2003. – 2003. – С.50.
20. Пупышева Т.Л., Демин А.В. Применение искусственных нейронных сетей в цитологической диагностике фолликулярных пролифератов щитовидной железы // Системный анализ и управление в биомедицинских системах. – 2003. – Т.2. – №1. – С. 38-40.

Демин А. В.

ЛОГИКО-ВЕРОЯТНОСТНЫЙ МЕТОД ИЗВЛЕЧЕНИЯ ЗНАНИЙ И ЕГО ПРИМЕНЕНИЕ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ И УПРАВЛЕНИЯ

Автореферат

Автореферат:
60*84 1/8, 1 п. л. Тираж 100 экз.
Заказ № 86. 20.11. 2008
Отпечатано ЗАО РИЦ «Прайс-курьер» ул. Кутателадзе, 4г, т. 330-7202