

РАЗРАБОТКА И ИССЛЕДОВАНИЕ БАЗЫ ДАННЫХ И БАЗЫ ЗНАНИЙ ДЛЯ ПРИНЯТИЯ РЕШЕНИЙ В ИНФОРМАЦИОННОЙ СИСТЕМЕ ОБСЛУЖИВАНИЯ БАНКОМАТОВ

Предлагается способ извлечения скрытых знаний и построение базы знаний информационной системы в сфере сервисного обслуживания банкоматов на основе нейросетевого подхода. Описывается функциональная структура инструментального программного средства для извлечения знаний.

1. Введение

В настоящее время сети банкоматов и платежных киосков достигают огромных размеров. Тенденция развития таких сетей говорит о том, что в ближайшее время парк банкоматов и платежных киосков будет возрастать. С увеличением размеров сети самообслуживания остро возникает вопрос сопровождения. Уменьшение времени простоя, а значит обеспечение доступности банкомата – одна из главных задач сопровождения. Уменьшить время реакции можно снижением нагрузки на персонал, что возможно путем создания единого центра мониторинга, консолидирующего в своем составе данные о работе сети самообслуживания из различных автоматизированных систем.

Несмотря на то, что в последнее время появилось множество различных коммерческих систем, в той или иной мере использующих методы Data Mining, проблема применения баз данных (БД) как дополнительного источника знаний при построении баз знаний (БЗ) интегрированных экспертных систем (ИЭС) по-прежнему остается актуальной.

Это обусловлено сложностью стыковки двух различных технологий – извлечения знаний из баз данных и традиционной инженерии знаний, применяющей в качестве основного источника знаний эксперта (или экспертов). При использовании БД как дополнительного источника знаний при построении БЗ информационной системы возникает ряд проблем, начиная от формирования выборки для анализа и заканчивая верификацией полученных фрагментов БЗ.

2. Обнаружение знаний в данных

Процесс поиска полезных знаний в “сырых” данных включает в себя вопросы: подготовки данных, выбора информативных признаков, очистки данных, применения методов Data Mining (DM), постобработки данных и интерпретации полученных результатов. Безусловно, “сердцем” всего этого процесса являются методы DM, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных (деревья решений), часто встречающиеся шаблоны (ассоциативные правила), а также результаты классификации (нейронные сети) и кластеризации данных (карты Кохонена) и т.д.

Процесс извлечения знаний состоит из нескольких шагов:

Подготовка исходного набора данных. Этот этап заключается в создании набора данных, в том числе из различных источников, выбора обучающей выборки и т.д. Для этого должны существовать развитые инструменты доступа к различным источникам данных. Желательно иметь поддержку работы с хранилищами данных и наличие семантического слоя, позволяющего использовать для подготовки исходных данных не технические термины, а бизнес понятия.

Предобработка данных. Для того чтобы эффективно применять методы Data Mining, следует обратить внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т.д. Кроме того, данные могут быть избыточны, недостаточны и т.д. В некоторых задачах требуется дополнить данные некоторой априорной информацией. Наивно предполагать, что если подать данные на вход системы в

существующем виде, то на выходе получим полезные знания. Данные должны быть качественны и корректны с точки зрения используемого метода DM. Более того, иногда размерность исходного пространства может быть очень большой, и тогда желательно применять специальные алгоритмы понижения размерности. Это как отбор значимых признаков, так и отображение данных в пространство меньшей размерности.

Трансформация, нормализация данных. Трансформация подразумевает оптимизацию данных для решения определенной задачи. Обычно на этом этапе выполняется исключение незначимых факторов, снижение размерности входных данных, нормализация, обогащение и другие преобразования, позволяющие лучше «приспособить» данные к решению аналитической задачи.

Data Mining. На этом шаге применяются различные алгоритмы для нахождения знаний. Это нейронные сети, деревья решений, алгоритмы кластеризации, установление ассоциаций и т.д.

Постобработка данных. Интерпретация результатов и применение полученных знаний в бизнес-приложениях.

3. Структура БД

Исследуемая задача характеризуется следующими факторами:

- неопределенность – при этом информация, необходимая для принятия решений, по большей части носит качественный характер;
- многокритериальность, причем исследуемые критерии представлены в различных шкалах измерений (интервальная, номинальная, порядковая);
- необходимость одновременного учета как количественных, так и качественных критериев оценки альтернатив;
- необходимость согласования групповых мнений экспертов;
- многоуровневость системы частных (локальных) критериев и их неравнозначность (критерии вносят различный вклад в интегральную оценку альтернативы);
- многократность процесса выбора;
- совместимость объективных и субъективных характеристик элементов задачи.

Разработанный комплекс моделей (функциональная, процессная, информационная) позволил получить полную и наглядную информацию об исследуемой предметной области, разработать структуру базы знаний и структуру разрабатываемой системы, определить требования к программному и аппаратному обеспечению систем. При этом информационная модель является логической структурой БД, а функциональная и процессная модель – основой для проектирования структуры БЗ.

В исследованиях использовалась база данных, структура которой разрабатывалась и оптимизировалась для компании, занимающейся сервисным обслуживанием банкоматов в процессе проведения описываемых исследований. Она охватывает основные, с точки зрения экспертной оценки, блоки принимающие участие в процессе:

- описание устройств и их атрибутов – это банкоматы и их основные характеристики, имеющие значение в процессе обслуживания;
- описание клиентской базы – это клиенты компании, соглашения о сервисе, данные о поставках и инсталляциях банкоматов;
- описание инженерного состава – это инженеры компании, которые реализуют обслуживание банкоматов, их квалификация, региональное распределение;
- описание процесса обслуживания – это заявки на выполнение работ, отчеты о выполненных работах, состояние устройств.

В результате исследования БД и проведенного статистического анализа [8] были выявлены основные показатели каждого из блоков, имеющие достаточное для предмета исследования значение. К ним относятся – долгосрочность банкоматов, их работоспособность, регламент проведения работ, среда функционирования устройств, загруженность специалистов, надежность сети.

4. Подсистема извлечения знаний

Автоматизация извлечения знаний из баз данных должна учитывать следующую специфику:

1. Данные имеют практически неограниченный объём.
2. В большинстве случаев данные являются числовыми, а если они качественные или текстовые, то для их обработки все равно используются количественные методы.
3. Извлеченные знания должны быть конкретны и понятны пользователю.
4. Инструменты обнаружения знаний должны быть просты в использовании и работать при наличии «сырых» данных.

В результате проведенных исследований для извлечения знаний предлагается двухэтапный метод с использованием нейронной сети и нечеткого интерпретатора. На первом этапе «сырые» данные обрабатываются нейронной сетью, которая в неявном виде формулирует скрытые закономерности. На втором этапе полученные знания извлекает, верифицирует и формулирует в лингвистической форме нечеткий интерпретатор.

а) Существующие модели и методы извлечения знаний

Первоначально основным инструментом анализа данных были классические методы математической статистики. Но современного пользователя все реже и реже удовлетворяют результаты статистического анализа, используемого для обобщения и интерпретации искомым закономерностей в явном виде $y = f(x)$, существующих между зависимой переменной y и m -мерным вектором $x = (x_1, \dots, x_m)$ независимых переменных. В известной степени это обусловлено тем, что в рамках статистического подхода требуется, чтобы были приняты гипотезы об однородности имеющихся данных, взаимной независимости переменных и т.д. Пользователь должен знать основные понятия и положения математической статистики для того, чтобы с их помощью интерпретировать интересующие его закономерности. Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для «грубого» предварительного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP) [8].

В основу современных методов технологии Data Mining (discovery-driven data mining) [3] положена концепция шаблонов, отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, найденные путем кластеризации. Для кластеризации необходима предварительная «очистка» данных, которая проводится известными в статистике методами сглаживания. Далее проводится анализ выявленных признаков на информативность. Выявленные закономерности в числовом формате подлежат верификации стандартными статистическими методами. На последнем этапе скрытые знания формируются в виде деревьев решений и сохраняются в базе знаний в форме продукционных правил.

Анализ задачи показывает, что практически идеальным средством для извлечения знаний из баз числовых данных являются нейронные сети. Исследования аппроксимирующих свойств нейронных сетей показывают, что они способны выявлять скрытые закономерности автоматически, без предварительной обработки данных и без выяснения степени взаимной корреляции входных переменных [4].

С формальной точки зрения нейронная сеть представляет собой универсальную аппроксимирующую модель в виде графа. Моделируя реальный объект, такой граф способен путем обучения повышать свою адекватность этому объекту за счет модификации весов межэлементных связей. Именно наличие формальных методик обучения (при условии существования обучающей выборки в виде экспериментальных пар «входы-выходы») является главным преимуществом нейронных сетей.

Важным преимуществом нейронных сетей является также то, что разработка экспертных систем, основанных на правилах, требует 12...18 месяцев, а нейросетевых - от нескольких недель до месяцев [5]. Однако в качестве средства извлечения знаний нейронные сети имеют определенные недостатки:

- весовые коэффициенты межнейронных связей обученной нейронной сети не подлежат ясной и содержательной интерпретации;
- на сегодняшний день модели нейронных сетей, использующих качественные входы, развиты недостаточно, чтобы успешно моделировать сложные объекты;
- градиентный метод, который традиционно применяется для обучения нейронной сети, не всегда дает возможность достичь глобального минимума в различении модельных и экспериментальных значений выхода.

Последний недостаток устраняется использованием эволюционных методов настройки весовых коэффициентов сети.

Для извлечения знаний могут применяться также нечеткие системы логического вывода. В работе [5] предлагается метод построения нечетких БЗ путем их извлечения из данных эксперимента. В описываемом методе используется нейронечеткая сеть, которая адаптируется к исследуемому объекту путём оптимизации значений параметров функций принадлежности входных переменных и значений весовых коэффициентов, нечетких правил.

Для поиска оптимальных значений указанных параметров используется генетический алгоритм. Однако для построения нейронечеткой экспертной системы требуется предварительно заполнить базу знаний правилами, т.е. создать прототип системы, который дообучается путем параметрической оптимизации. Если априорные формулировки правил отсутствуют, то данный метод неприменим.

В книге [6] со ссылкой на работу [7] описывается простой метод извлечения лингвистических знаний из «сырых» числовых данных при помощи нечеткой интерпретации скрытых закономерностей. Применяя формализм теории нечетких множеств, метод можно описать следующим образом:

1. Для каждой входной и выходной переменной уточняется диапазон допустимых значений:

$$X_i = [x_i, \bar{x}], Y[y_i, \bar{y}_j], \quad (1)$$

В результате проведенного статистического анализа к входным переменным мы отнесли:

- долгосрочность: срок службы, наработка, новизна;
 - работоспособность: время в ремонте, время неисправности;
 - регламент проведения работ: время реакции, время выполнения;
 - среду функционирования: показатели напряжения, заземления;
 - загруженность специалистов: количество банкоматов, количество инженеров;
- к выходным:
- надежность сети: количество отказов, количество просроченных работ и количество замен запчастей.

2. Каждая переменная получает лингвистическую интерпретацию, и диапазон её изменения разбивается на нечеткие интервалы, обозначенные термами:

- терм-множество переменной X_i :

$$A_i = \{a_i^1, a_i^2, \dots, a_i^{l_i}\}, i = \overline{1, n}; \quad (2)$$

- терм-множество переменной Y :

$$D = \{d_1, d_2, \dots, d_m\}. \quad (3)$$

Здесь a_i^p – p -й лингвистический терм переменной x_i ; $p = \overline{1, l_i}, i = \overline{1, n}$, d_j – j -й лингвистический терм переменной y ; m – число различных решений в рассматриваемой области.

Мощности терм-множеств A_i , $i = \overline{1, n}$ в общем случае могут быть различны, т.е. $l_1 \neq l_2 \neq \dots \neq l_m$.

На примере переменной x_i – «срок службы» определяем терм-множество A_1 :

$$A_1 = \{a_1^1, a_1^2, \dots, a_1^{l_1}\}, i = \overline{1, n}, \quad (4)$$

где a_1^1 – бессрочный – если изготовитель не установил срок службы; a_1^2 – минимальный – до 12 месяцев; a_1^3 – средний – от 12 до 36 месяцев; a_1^4 – максимальный – от 36 месяцев.

Получаем

$$A_1 = \{\text{бессрочный, минимальный, средний, максимальный}\}. \quad (5)$$

3. Каждый нечеткий интервал снабжается функцией принадлежности. Тогда нечеткие множества a_i^p и d_j определим так:

$$a_i^p = \sum_{k=1}^{q_i} \mu^{a_i^p}(x_i^k) / x_i^k, d_j = \sum_{r=1}^{q_m} \mu^{d_j}(y^r) / y^r, \quad (6)$$

где $\mu^{a_i^p}(x_i^k)$ – степень принадлежности элемента $x_i^k \in x_i$ терму $a_i^p \in A_i, p = \overline{1, 1}, i = \overline{1, n}, k = \overline{1, q_i}, ; \mu^{d_j}(y^r)$ – степень принадлежности элемента $y^r \in Y$ терму-решению $d_j \in D, j = \overline{1, m}$.

4. Числовые данные группируются в виде пар «вход-выход»: $(X^l, y^l), l = \overline{1, M}$, где $X^l = (x^l_1, x^l_2, \dots, x^l_n)$ – входной вектор и соответствующее значение выходной переменной y^l для l -й пары «вход-выход», $y^l \in [y_i, \overline{y_j}]$.

5. Для каждой пары «вход-выход» вычисляются значения функций принадлежности $\mu^{a_i^p}(x_i^k)$ вектора X^l и $\mu^{d_j}(x_1^*, x_2^*, \dots, x_n^*)$ для всех нечетких интервалов, на которые разбит диапазон допустимых значений $[y_i, \overline{y_j}]$ выходной переменной y .

6. Для каждой входной переменной x_i^k отыскивается терм a_i^p , для которого функция принадлежности $\mu^{a_i^p}(x_i^k)$ имеет максимальное значение $\mu^{a_i^p \max}(x_i^k)$. Для выходной переменной также отыскивается терм d_j^l , для которого функция принадлежности $\mu^{d_j^l}(y^l)$ имеет максимальное значение $\mu^{d_j^l \max}(y^l)$.

7. Формируется правило типа «ЕСЛИ-ТО», которое связывает лингвистические значения $a_{i \max}^1, a_{i \max}^2, \dots, a_{i \max}^l$ входных переменных с лингвистическим значением выходной переменной $d_{j \max}^l$:

$$\left[\bigcap_{i=1}^n (x_i^l = a_{i \max}^{j^p}) \right] \rightarrow y = d_{j \max}^l, j = \overline{1, m}. \quad (7)$$

8. Для полученного правила вычисляется степень истинности:

$$R^l = \mu^{d_{j \max}^l}(y^l) \prod_i \mu^{a_{i \max}^p}(x_i^l). \quad (8)$$

9. Отбор правил. Если в накопленной БЗ уже существует правило с номером t с такими же значениями $a_{i \max}^1, a_{i \max}^2, \dots, a_{i \max}^l$ и $d_{j \max}^l$ и выполняется условие $R^l > R^t$, то правило l фиксируется в базе знаний вместо правила t .

Таким образом, накапливается БЗ в лингвистической форме на основе нечеткого анализа числовых данных. Несомненным достоинством метода является простота реализации. Однако у данного метода есть резервы повышения эффективности и точности, которые можно реализовать, если в качестве генератора правил использовать обученную нейронную сеть и применить более совершенный алгоритм отбора и верификации правил.

б) Нейронная сеть и формирование базы знаний

В изложенном выше методе числовые данные подвергаются нечеткому анализу без предварительной подготовки, а значит, содержат шумы и потенциальные противоречия. Естественно, данные можно предварительно обрабатывать статистическими методами, но гораздо удобнее использовать их для обучения нейронной сети. В таком случае сеть автоматически накапливает знания о скрытых многомерных зависимостях в числовых данных. После обучения сети ей можно предъявлять любые входные векторы из допусти-

мого диапазона и получать адекватные выводы при помощи описанного выше метода интерпретации, вообще не накапливая лингвистических знаний. В этом случае из перечисленных выше пунктов будут использоваться пункты 1...3, а степень истинности $\mu^{d_{j\max}}(y^1)$ интерпретируется как степень истинности ответа.

На данном этапе использовалась НС следующей конфигурации:

- количество узлов входного слоя – 11;
- количество узлов скрытого слоя – 5;
- количество узлов в выходном слое – 3;
- алгоритм обучения – обратное распространение;
- функция активации – сигмоидальная.

Если ставится задача составить базу знаний лингвистического характера (например, для системы нечеткого логического вывода), то встает вопрос о процедуре отбора правил. Представляется разумным не отбрасывать правила с различными, пусть и небольшими значениями степени истинности, а накопить достаточно полную совокупность экземпляров правил по всем или по значительной части пар «вход-выход» обучающей выборки нейронной сети.

Обсудим вопросы верификации и отбора правил.

Если разработчик СППР обнаруживает, что система по результатам испытаний допускает значительное количество ошибок, у него есть несколько вариантов коррекции правил.

Во-первых, можно изменить структуру нейронной сети, используя такие подходы, как МГУА [5].

Во-вторых, можно настраивать параметры функций принадлежности, используя методы, описанные в работе [5]. В-третьих, можно верифицировать и отбирать правила, полученные на выходе интерпретатора, как это предлагается в изложенном выше методе. Но замена предыдущего правила новым, с увеличенным значением степени истинности, не представляется единственным и наиболее верным подходом. На наш взгляд, следует накопить значения истинности всех найденных одинаковых правил и вычислить усредненную степень истинности по всей выборке.

Если количество найденных одинаковых правил достаточно велико, то состоятельность результирующего правила измеряется средним значением степени истинности R . Если же количество найденных одинаковых правил не превышает определённого порогового значения A , которое устанавливается по результатам частотного анализа накопленной совокупности правил, то среднюю степень истинности R для одинаковых правил вычислим как $R = k/A$, где k – количество одинаковых правил в выборке. Для отбора надежных правил установим пороговое значение степени истинности R_n . В базу знаний попадут правила, у которых $R > R_n$.

Если в накопленной совокупности правил найдены выборки противоречивых правил, имеющих одинаковые левые части, но различные выводы d_j , следует провести дополнительный анализ для выяснения вопроса: отдать ли предпочтение одному из вариантов или удалить оба варианта правил из базы знаний. Для решения предлагается следующая методика.

Для каждой выборки правил вычисляется вектор средних значений всех входных переменных $(x_1^{-1}, x_2^{-1}, \dots, x_n^{-1})$ и $(x_1^{-2}, x_2^{-2}, \dots, x_n^{-2})$.

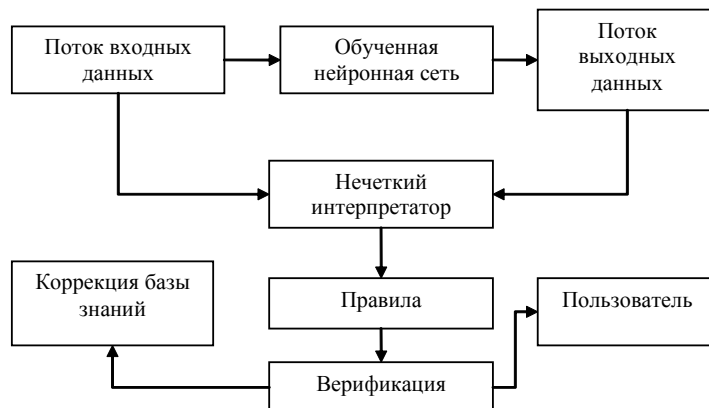
Вычисляется вектор средних значений по всем входным переменным по обоим выборкам (x_1, x_2, \dots, x_n) .

Полученный в п. 2 вектор подается на вход нейронной сети и вычисляется числовое значение выхода сети.

Вход и выход сети интерпретируются в лингвистическую форму.

Если полученное правило совпадает с одним из конкурирующих, то предпочтение отдается выигравшему правилу. Если же новое правило не совпадает ни с одним из конкурирующих, то оба конкурирующих правила удаляются из БЗ. На рисунке показана функциональ-

ная структура разработанного инструментального программного средства, реализующего предлагаемый метод извлечения и интерпретации знаний, накопленных нейронной сетью.



Функциональная структура инструментального программного средства

5. Заключение

Предложен модифицированный метод извлечения скрытых знаний из базы данных с использованием обученной нейронной сети и нечеткого интерпретатора входных и выходных данных. Метод предназначен для быстрой разработки интеллектуальных СППР на основе нейросетевого подхода и предусматривает применение специальной процедуры отбора и верификации правил. Предложена функциональная структура инструментального программного средства для извлечения знаний, позволяющая проводить эксперименты с изложенной методикой извлечения знаний.

Список литературы: 1. Рыбина Г.В. Автоматизированное построение баз знаний для интегрированных экспертных систем // Изв. РАН. Теория и системы управления. 1998. №5. С.152-166. 2. Калинина Е.А., Рыбина Г.В. Применение технологии Data Mining для автоматизированного построения баз знаний интегрированных экспертных систем // В кн.: КИИ-2000. Седьмая нац. конференция с межд. участием. Труды конференции. Том 2. М.: Физматлит, 2000. С. 119-127. 3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining / СПб.: БХВ-Петербург, 2004. 326 с. 4. Сетлак Г. Интеллектуальные системы поддержки принятия решений. К.: Логос, 2004. 251 с. 5. Митюшкин Ю.И., Мокин Б.И., Ротштейн А.П. Soft Computing: Идентификация закономерностей нечеткими базами знаний. Винница: УНИВЕРСУМ-Винница, 2002. 145 с. 6. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия, 2006. 193 с. 7. Wang L.X., Mendel J. M., Generating Fuzzy Rules by Learning from Examples, IEEE Transactions on Systems, Man, and Cybernetics, November/December 1992. Vol. 22, № 6. P. 1414-1427. 8. Кузмин А.Я., Головий Н.В., Даюб Я. Реализация модели системы поддержки принятия решений в области сервисного обслуживания банкоматов // АСУ и приборы автоматики. 2009. №143. С.57 – 64.

Поступила в редколлегию 15.08.2009

Головий Наталья Владимировна, аспирантка кафедры информатики ХНУРЭ. Научные интересы: системный анализ данных. Адрес: Украина, 61166, Харьков, пр.Ленина,14, тел. (057)702 15 15, e-mail: rica1982@mail.ru.