ем метода половинного деления. Входными параметрами оптимизации являлся шаг оптимизации и начальная точка приближения. Также была реализована методика расчета стоимости обслуживания системы с определенной стратегией обслуживания.

На седьмом этапе были рассчитаны показатели надежности элементов и подсистем, результаты показали, что ресурс части оборудования еще не достиг предела, а другой части уже истек. Таким образом, на основании проведенного анализа заказчику СА было предложено решение замены части работающего оборудования, что в итоге позволило повысить уровень безопасности использования АС. Представленная ИС применялась при вероятностном анализе безопасности оборудования Билибинской и других АС. Хотя замена и профилактика оборудования часто привязана к перегрузке топлива, ИС позволила оптимизировать профилактику части подсистем АС.

Заключение. Как показывает практика, проведение системного анализа включает в себя большую совокупность разноплановых работ, требует наличие большого опыта и знаний от системного аналитика при работе с математическими методами, моделями систем, программными средствами, с методами анализа и принятия решений. В зависимости от цели системного анализа, представленных для СА ресурсов и сроков выполнения работ его реализация возможна в определенной последовательности, характерной только для данного типа СА. Вследствие многоаспектности СА и применимости его практически в любой сфере деятельности человека, подобрать стандартные процедуры, методы и этапы его проведения очень сложно, что в общем является также особым видом СА.

В ходе проделанной работы была предложена методика проведения СА, представлен пример ее реализации. Показано, что разработка математических моделей и их реализация в ИС является необходимыми компонентами СА.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Системный анализ и принятие решений. Словарь-справочник. – М: Высшая школа, 2005. – 616 с.
2. *Орловский П.Н.* Системный Анализ (основные понятия, принципы, методология). Том 1 / Учебное пособие. – Киев: Минобразования Украины, 1996. – 360 с.
3. *Антонов А.В.* Проектирование систем. – Обнинск: Изд-во ИАТЭ, 1996. – 157 с.
4. *Моисеев Н.Н.* Математические задачи системного анализа. – М.: Наука, 1981. – 488 с.
5. *Антонов А.В.* Системный анализ / Учебник для вузов. – М: Высшая школа, 2006. – 470 с.
6. *Дагаев А.В., Антонов А.В, Чепурко В.А.* Модель анализа надежности подсистем ЯЭУ со встроенным контролем // Ядерная энергетика. Известия вузов. – Обнинск: Изд-во ИАТЭ, 2001, №2. – С. 3-9.

УДК 519.007

А.Н. Шабельников, В.А. Шабельников

ПОИСК АНОМАЛИЙ В ТЕХНИЧЕСКИХ БАЗАХ ДАННЫХ ВРЕМЕННЫХ РЯДОВ*

Введение. Базы данных временных рядов (БДВР) отличаются от статических БД тем, что содержат записи, в которых некоторые из атрибутов ассоциируются с временными метками. В качестве таких записей могут выступать данные монито-

* Работа выполнена при поддержке РФФИ, проекты № 07-01-00075 и № 07-07-00010

ринга и телеметрии, биржевые данные, транзакции продаж в продовольственных магазинах и т.п. Под выявлением знаний в БДВР понимается процесс анализа темпоральных данных с целью выявления в них интерпретационно пригодных для человека, новых и полезных для целей принятия решений правил [1].

В данной статье, предлагается методология процесса поиска знаний в БДВР, полученных в ходе мониторинга состояний технологических датчиков на сортировочных горках. Задачей выявления знаний является поиск аномалий, характеризующих отклонения процесса от типового режима или свершения относительно редких событий. Выявление аномалий осуществляется на основе ассоциативных нечетких правил предсказания – правил, на основе которых оказывается возможным предсказание аномального поведения контролируемого параметра с использованием информации о текущих параметрах процесса и его поведении в прошлом.

Постановка задачи. *Определение 1.* Числовым ВР (в дальнейшем, просто ВР) называется множество упорядоченных временных отсчетов вместе с соответствующими им числовыми значениями:

$$Y = \{(y_i, t_i) / i \in N,\ y_i \in R,\ t_i \in T\},$$

где T – дискретная временная шкала; R – множество действительных чисел $y_i = y(t_i)$, характеризующих числовые значения ВР в i-е моменты времени.

Определение 2. Временным интервалом или темпором называется пара $[t_s, t_e] \in \mathrm{T}^2$, в которой $s \le e$.

Будем обозначать через $\sum = ([s,e] \in N^2 / s \le e\}$ конечное множество всех временных интервалов на шкале T, а через $Q = \{\alpha, \beta, ..., \gamma\}$ – множество признаков, характеризующих обобщенные свойства-признаки ВР Y, проявляемые на его отдельных фрагментах.

Определение 3. БДВР называется множество записей $\{r_j = (y_1, y_2, ..., y_m, t_j)\}$, в котором каждая запись содержит определенный набор числовых либо символьных атрибутов и определенное значение времени, задаваемое на временном масштабе в определенном разрешении.

Определение 4. Функцией j-го атрибута называется функция времени $y_j(t)$, значениями которой являются значения j-го атрибута в записях $r_i \in БДВР$.

Определение 5. Признаком функции атрибута $y(t)$ на интервале $[t_1, t_2]$ называется некоторая функция $Q(t)$, аппроксимирующая функцию $y(t)$ на интервале $[t_1, t_2]$, т.е.

$$y(t) \approx Q(t) \quad \forall t \in [t_1, t_2].$$

Термин “аппроксимирует” может быть интерпретирован различными способами относительно конкретной предметной области, например:

$$| y(t) - Q(t) | < \varepsilon \quad \forall t \in [t_1, t_2].$$

Иногда под признаком функции атрибута будем понимать параметры аппроксимирующей функции. Например, если на некотором интервале имеет место

$y(t) = \alpha \cdot t + \beta$, то можно сказать, что на этом интервале функция обладает уклоном α, значение которого является признаком, извлеченным из ВР $y(t)$.

Определение 6. Темпоральным паттерном (образом) ВР $\Re(Y)$ будем называть последовательность признаков, сопоставленных строго упорядоченной последовательности непересекающихся темпоров:

$$\Re(Y) = Q_{i1}([t_{i1}^n, t_{i1}^k]), Q_{i2}([t_{i2}^n, t_{i2}^k], \ldots, Q_{im}([t_{im}^n, t_{im}^k] \quad t_{ij}^k \le t_{ij+1}^n \forall j = 1,2,\ldots m .$$

Определение 7. Темпоральным правилом ассоциации называется правило вида

$$\Re_1(Y) \Rightarrow \Re_2(Y),$$

где $\Re_1(Y), \Re_2(Y)$ – темпоральные образы ВР Y.

С учетом приведенных выше определений задача поиска знаний в БДВР сводится к препроцессингу данных в БДВР; извлечению системы признаков Q и формированию правил ассоциаций.

Общий подход к выявлению знаний. Ключевую роль в контексте поиска знаний играет выбор модели представления ВР. В этом смысле весьма эффективными являются представления ВР на основе темпоральных признаков, извлеченных с использованием скользящего окна. Исходный ВР сегментируется на непересекающиеся временные интервалы с последующим представлением каждого интервала одним из нескольких базовых примитивов (шейпов). Обычно в качестве шейпов используются аппроксимирующие функции, описывающие кривизну сегмента. Так, например, в [2] шейповые признаки извлекаются с использованием многошкальной вейвлет сегментации и представления каждого интервала одним из семи примитивов. В [3] используется расширенное множество из 13 примитивов. Более простые шейповые представления в виде линейных трендов лежат в основе методов кусочно-агрегированной аппроксимации (PAA) [4].

В настоящей работе предлагается подход к представлению числовых ВР в символьном виде путем многомерной нечетко-шейповой аппроксимации ВР. Для этого ВР преобразуется в многомерный ВР, каждое измерение которого характеризует определенный числовой признак, извлеченный из исходного ВР, а затем числовым признакам сопоставляются нечеткие числовые термы. Комбинации нечетких термов, сопоставленных разным измерениям ВР, образуют нечетко-шейповые представления интервалов ВР. При таком представлении каждый интервал ВР ассоциируется с несколькими символьными признаками. Один из выделенных признаков, конкретный выбор которого определяется характером задачи, является целевым в контексте поставленной задачи. Так, например, в задачах прогнозирования ВР целевым признаком является непосредственно прогнозируемая числовое значение ВР, представленное в виде числового нечеткого терма, а в задачах автоматического (без учителя) выявления аномалий целевым признаком может быть любая величина, характеризующая резкое отклонение контролируемых параметров от нормальных значений. Строго упорядоченная последовательность из m нечетко-шейповых признаков задает темпоральный образ ВР. Далее на основе алгоритма типа Априори [5] выявляются ассоциативные нечеткие правила на m-ах нечетко-шейповых признаков. Конечные правила являются ассоциативными правилами на помеченных интервалах, которые в задачах прогнозирования интерпретируются как правила предсказания целевых атрибутов, а в задачах выявления аномалий, как правила идентификации аномалий.

Препроцессинг БДВР. Препроцессинг данных ВР включает в себя предобработку сырых данных с целью удаления в них аддитивного шума. Положим, что сырые данные $y_i = y(t_i)$ получены из долгосрочных трендов сигнала $\hat{y}(t_i)$ и аддитивного шума $e(t_i)$, т.е.

$$y(t_i) = \hat{y}(t_i) + e(t_i).$$

Целью препроцессинга является получение оценок длительного сигнала $\hat{y}(t_i)$. Для этого необходимо описать сигналы $y(t)$ и $e(t)$.

Шум является случайным по своей природе и подвергается влиянию различных факторов. В отличие от него тренд более стабилен и подвержен влиянию сравнительно малого числа факторов. Поэтому для очистки данных можно использовать оператор низкочастотного фильтра (LPF), устраняющий высокочастотные компоненты. Один из простейших типов оператора LPF является конечный фильтр (FIR), определяемый как:

$$\hat{y}(t_i) = \sum_{j=0}^{N-1} y(t_i - j + N/2 \cdot c(j),$$

где $y(t_i)$ – начальная функция атрибута; $\hat{y}(t)$ – очищенная функция атрибута; $c(j)$ – весовые коэффициенты.

Фильтр принимает на вход N сэмплов и вычисляет скалярное произведение с вектором весовых коэффициентов. Размер и значения коэффициентов формируются с учетом полосы пропускания и требуемой точности.

Извлечение признаков. В контексте выявления знаний абсолютные числовые значения ВР, описывающих течение процессов в БДВР, оказываются не так интересны, как шейпы, описывающие резкие изменения этих значений. Кроме того, информативно важным является показатель хаотичности, характеризующий вариабильность числовых значений ВР. Особо важную роль, как оказалось, данный параметр играет в поиске аномалий, поскольку одним из “предвестников” аномалии часто является хаотичность поведения ВР. В связи с этим в качестве опорных признаков ВР были выбраны три интегральных параметра: тренд (α), смещение (β), хатоичность (η).

В соответствии с выбором опорных признаков числовой ВР преобразуется в 3-хмерный ВР, каждое измерение которого представлено одним из темпоральных признаков. Последние формализованы следующим образом:

$$\alpha(t_i) = \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i};$$

$$\beta(t_i) = y(t_i);$$

$$\eta(t_{iy}) = [y(t_i) - \hat{y}(t_{t-1}, t_{i+1})] / y(t_i),$$

где $\hat{y}(t_{i-1}, t_{i+1})$ – аппроксимированное значение ВР в точке t_i, вычисленное на основе прямой, проходящей через точки $y(t_{i-1}), y(t_{i+1)})$.

Семантика признаков α и β вполне очевидна, а содержательный смысл параметра η заключается в мере нестабильности ВР (флуктуации их значений). Высокое значение η показывает, что ВР вариабилен и подвержен влиянию различных факторов, низкое значение указывает на его стабильность.

Переход от числовых значений признаков к символьным осуществляется с использованием лингвистической сегментации на основе величины переменной с использованием таких термов как «ВЫСОКОЕ», «НИЗКОЕ» и т.п. В качестве метода сегментации используется метод дискретизации числовых признаков на основе гистограмм распределения числовых значений [6]. Выделенные на гистограммах сегменты характеризуются минимальными значениями энтропии на границах и соответствуют столбцам гистограммы, имеющим примерно равные площади.

Временные сегменты, полученные в результате дискретизации числовых признаков, являются базовыми интервалами, на которых определены ФП соответствующих нечетких числовых термов, заданные в классе треугольных функций принадлежности (ТФП). Центрами ТФП являются центры соответствующих базовых интервалов, а границами – центры ТФП смежных значений.

Для окончательного извлечения интервальных признаков используется объединение всех отсчетов ВР, характеризуемых одинаковыми символьными значениями по всем измерениям, в единый интервал.

Поиск правил. В зависимости от типа выявляемых правил используются два различных подхода к их поиску.

Поиск правил предсказания основан на информационно-теоретическом подходе, впервые предложенном в [7] и адаптированном к поиску темпоральных правил. Взаимодействие между входными атрибутами (шейпами предсказания) и целевыми (классификационными) атрибутами моделируются информационно-теоретической коннекционистской сетью, состоящей из корневого узла и переменного количества скрытых слоев, соответствующих входным и целевым атрибутам. Элементы каждого скрытого уровня ассоциируются с различными нечетко-шейповыми признаками. Связи в сети устанавливаются по мере ее обучения между нейронами промежуточных слоев и целевыми нейронами. Эти связи представляют ассоциации между сочетанием входных атрибутов, соответствующих предшествующим нечетко-шейповым значениям ВР, и целевым атрибутом, соответствующим предсказанным значениям.

На рис. 1 приведен заимствованный из [7] пример двухслойной информационно-коннекционисткой сети, построенной на основе двух атрибутов.

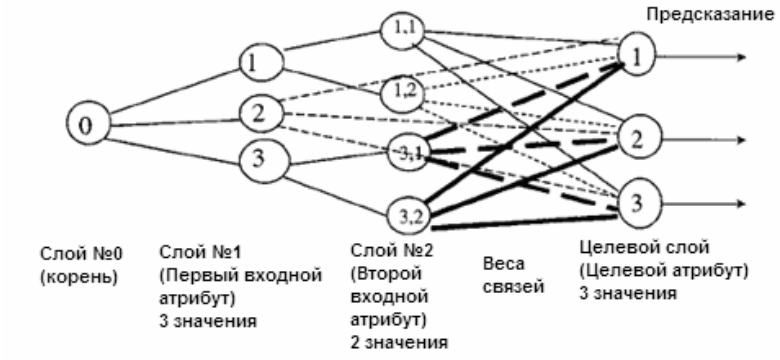

Рис. 1. Информационно-теоретическая коннекционистская сеть

В приведенной сети первый входной атрибут соответствует трем значениям, представленными элементами 1, 2 и 3 в первом слое. По результатам теста статистической значимости в качестве значимых выделены только элементы 1 и 3. Второй слой содержит четыре узла, описывающих комбинацию двух значений второго входного атрибута с двумя элементами первого слоя. Целевой атрибут имеет три значения, представленные тремя узлами в целевом слое.

Описанию алгоритма построения сети предпошлем следующие рассуждения.

Пусть в символьном ВР Y выделен темпоральный образ Ac (A – темпоральный подобраз, c – символьный признак), имеющий поддержку (количество вхождений образа во ВР Y), равную $|Ac|$. Тогда величина $\frac{|Ac|}{|*c|}$ ($|*c|$ – количество вхождений символа “c” во ВР Y) характеризует условную вероятность появления целевого атрибута “c” после образа “A”, или, иными словами, возможность наличия причинной связи $A \Rightarrow c$. С другой стороны, для любого иного целевого атрибута q величина $\frac{|Aq|}{|*q|}$ характеризует условную вероятность появления целевого атрибута “q” после образа “A”, т.е. вероятность наличия причинной связи $A \Rightarrow q$. Тогда величина $(1-\frac{|Aq|}{|*q|})$ характеризует вероятность <<не появления>> “чужого” атрибута “q” после образа “A”. Для всех иных m “чужих” атрибутов q_j величина $\prod_{j=1}^{m}(1-\frac{|Aq_i|}{|*q_j|})$ характеризует вероятность <<не появления>> ни одного из них после образа A. В результате величина $\mathrm{P}=\frac{|Ac|}{|*c|}\cdot\prod_{j=1}^{m}(1-\frac{|Aq_i|}{|*q_j|})$ характеризует условную вероятность появления целевого атрибута “c” после образа *A* при одновременном <<не появлении>> ни одного из “чужих” атрибутов. Таким образом, величина P может служить некой условной мерой полезности использования образа A в качестве причины появления целевого атрибута “c”, а, следовательно, и одновременно – неким информационно-теоретическим весом ассоциативного правила $A \Rightarrow c$.

Алгоритм построения сети сводится к последовательному установлению связей между нейронами в соответствии с правилами предсказания целевых атрибутов. Предварительно заметим, что в ниже описываемой инфонечеткой сети (ИНС) каждый путь между нейронами промежуточных слоев и целевым нейроном “c” целевого слоя характеризует некий темпоарльный образ, который, в свою очередь, порождает соответствующее ассоциативное правило предсказания вида “*<если <конъюнкция входных значений> то <целевое значение>*”.

1. На первой итерации для заданного порога ε устанавливаются все связи между атрибутами последнего внутреннего слоя a_j и целевым атрибутом “c”, если они обеспечивают информационно теоретический вес правила $a_j \Rightarrow c$ не ниже ε.

2. Пусть уже установлены связи между k-ым и (k-1)-ым слоем ИНС. Тогда устанавливаются связи между (k+1)-ым и k-ым слоем следующим образом. Добав-

ляется связь между нейронами a_{k+1} и a_k, если она увеличивает информационно-теоретический вес вновь создаваемого сетью образа, то есть, если $\mathrm{P}(a_{k+1} \cup A_k) > \mathrm{P}(A_k)$ (A_k – подсеть, уже сформированная на k-ой итерации).

3. Алгоритм завершает работу, когда не остается входных атрибутов, добавление которых увеличило бы информационно-теоретический вес целевых правил.

Правила определения аномалий часто еще называют правила поиска «новинок» или “сюрпризов” [8]. Идея выявления правил аномалий заключается в предположении, что интересующие нас образы аномальных явлений во ВР обычно проявляют себя редко, во всяком случае, по сравнению с часто встречающимися (условно нормальными) темпоральными образами. Исходя из этого, предлагаемый способ определения аномалий заключается в создании модели нормального состояния ВР в виде правил предсказания. Входные данные, не соответствующие в достаточной мере модели, отмечаются как аномальные. Для этого используются модификация правил предсказания путем введения в них операторов отрицаний с использованием техники работы с отрицаниями [9].

Заключение. Выше описанная методология пока еще не является достаточно разработанной с тем, чтобы делать какие-либо выводы. Описаны лишь принципиальные моменты метода нахождения зависимостей между паттернами во временных последовательностях. Тем не менее, предварительные результаты модельных экспериментов показали, что предлагаемая методология способна порождать правила, описывающие причинные ассоциации между часто встречающимися паттернами ВР. Следует заметить, что предложенный подход применим как к числовым, так и многомерным символьным ВР.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Батыршин И.З.* Основные операции нечеткой логики и их обобщение. – Казань: Отечество, 2001. – 102 с.
2. *A. Ultsch.* Knowledge discovery, lecture notes, 2003a. German.
3. *Bakshi B.R. and Stephanopoulos G.* Representation of process trends - IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. Computers & Chemical Engineering, 18(4):303-332, 1994.
4. *Colomer J., Melendez J., and Gamero F.* Pattern recognition based on episodes and DTW. Application to diagnosis of a level control system. In Proceedings 16th International Workshop on Qualitative Reasoning (QR'02), pages 37-43, 2002.
5. *Keogh E., Chakrabarti K., Pazzani M. J., and Mehrotra S.*. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems, 3(3): 263-286, 2001b
6. *Agrawal R. and Srikant R.*. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, pages 487 – 499, 1994.
7. *Daw C.S., Finney C. E. A, and E. R. Tracy.* A review of symbolic analysis of experimental data. Review of Scienti_c Instruments, 74(2):916-930, 2003.
8. *Last M., Klein Y., and Kandel A.* Knowledge discovery in time series databases. IEEE Transactions on Systems, Man, and Cybernetics, 31(1):160-169, 2001
9. *Keogh E., Lonardi S., and Chiu B.* Finding surprising patterns in a time series database in linear time and space. In D. Hand, D. Keim, and R. Ng, editors, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), pages 550-556. ACM Press, 2002.