

Запорожец Дмитрий Юрьевич – Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Южный федеральный университет»; e-mail: elpilasgsm@gmail.com; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; аспирант.

Курейчик Владимир Викторович – e-mail: vkur@tgn.sfedu.ru; кафедра систем автоматизированного проектирования; зав. кафедрой; д.т.н.; профессор.

Zaporozhets Dmitry Yurevich – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”; e-mail: elpilasgsm@gmail.com; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; postgraduate student.

Kureichik Vladimir Victorovich – e-mail: vkur@tgn.sfedu.ru; the department of computer aided design; head of department; dr. of eng. sc.; professor.

УДК 519.816.

С.М. Ковалев

МЕТОДЫ МНОГОШАГОВОГО ПРЕДСКАЗАНИЯ АНОМАЛИЙ В ТЕМПОРАЛЬНЫХ ДАННЫХ*

Предлагается новая методология обнаружения аномалий в темпоральных данных на основе методов многошагового предсказания. Предлагаемая методология базируется на анализе динамики развития вероятностных значений аномалии с поступлением каждого нового отсчета паттерна в потоке данных. Приводится оригинальная модель задачи многошагового предсказания. В качестве средства вычисления вероятностных значений аномалий предлагается использовать функции истинности стохастических Марковских моделей с доходами. Приводится теоретическое обоснование корректности предлагаемого подхода. Для преобразования реального процесса к Марковскому предлагается использовать нечеткую адаптивную систему, осуществляющую отображение совокупностей точечных состояний исходного процесса на гранулированную шкалу нечеткого интегрального признака.

Темпоральные данные; обнаружение аномалий; Марковская модель; темпорально-разностное обучение; адаптивная нечеткая система.

S.M. Kovalev

THE METHOD OF MULTISTEP PREDICTION OF ANOMALIES IN TEMPORAL DATA

In this paper, we develop a new method for the detection of anomalies in the temporal data on the basis of a multi-step methods for prediction. The proposed method is based on the analysis of the dynamics of the probability values of the anomaly on the input model. As a means of calculating the probability values of the anomaly, we suggest using truth functions of stochastic Markov models with income. We prove a theorem which is the rationale for the use of the method. To convert to a real process of Markov, we propose to use an adaptive fuzzy system that can show a set of point states of the initial process on a granular scale fuzzy integral feature. To convert to a real process of Markov, we propose to use an adaptive fuzzy system that can show a set of point states of the initial process on a granular scale fuzzy integral feature.

Temporal data; anomaly detection; Markov model; temporal-difference learning; adaptive fuzzy system.

Введение. Обнаружение аномальных событий и аномальных паттернов в темпоральных данных является важной областью Data Mining, имеющей отношение к различным приложениям. Традиционно, под обнаружением аномалий пони-

* Работа выполнена при поддержке РФФИ, грант 13-07-00183-а, 12-07-00012-а.

мается поиск “новинок”, “сюрпризов”, “выбросов” и иных особых событий в темпоральных данных, существенно отличающихся от паттернов нормального поведения. В большинстве известных приложений задача детектирования аномалий и методы их обнаружения рассматриваются в статическом варианте как распознавание полностью сформировавшихся паттернов на входе классификатора, не учитывая особенности развития подлежащего распознаванию темпорального паттерна в потоке данных. Однако существует широкий круг задач, в которых требуется максимально быстрое распознавание аномалий на основе непрерывного анализа поступающего на вход системы потока данных и упреждающего вынесения гипотез о наличии в нем аномальных событий.

В настоящей статье речь пойдет об одном из приложений задачи поиска аномалий, представленной в виде задачи многошагового предсказания нештатных ситуаций, возникающих в ходе развития технологических процессов. Для таких задач являются характерными три особых признака:

- ◆ полезная информация о состоянии процесса поступает на вход системы последовательными порциями в виде развитий темпоральных паттернов;
- ◆ каждая новая порция данных, являясь предвестником потенциальной аномалии, уточняет информацию о возможности ее появления, содержащаяся в предшествующих данных;
- ◆ принятие решений должно осуществляться в упреждающем режиме, не дожидаясь появления аномального события или момента окончания развития аномального паттерна на входе классификатора.

Характерным примером задачи многошагового предсказания является прогнозирование сбоев в движении поездов на основе анализа развития поездной обстановки в течение определенного временного интервала. Входной информацией в этой задаче являются данные о временах запаздывания поездов по каждой станции, категории поездов, особенности путевого развития станций и ряд других параметров поездных ситуаций. Поездной диспетчер должен заблаговременно предсказать такую ситуацию с целью принятия превентивных мер по нормализации техпроцесса. Данная и другие аналогичные задачи обуславливают новую постановочную модель задачи многошагового предсказания.

Модель многошагового предсказания. Модель многошагового предсказания включает три следующих основных компонента.

1. *Входная информация* представляется в виде векторов наблюдений за развитием процесса

$$X(t) = x(t_1), \dots, x(t_i), \dots, x(t_k), z,$$

где $x(t_i)$ характеризует состояние процесса в i -й момент времени. Здесь особо выделены конечные значения z в качестве результатов, “1” или “0”, характеризующих факты принадлежности вектора $X(t)$ к классу аномальных паттернов.

2. *Шейп предсказания*, полученный на основе вычислений вектор

$$P(X) = p(x(t_1)), \dots, p(x(t_i)), \dots, p(x(t_k)),$$

элементы которого являются оценками результата z . Каждое из значений $p(x(t_i))$ характеризует вероятность того, что i -е значение паттерна $x(t_i)$ является предвестником аномального события в конце паттерна.

3. *Способ вычисления шейпа предсказания* на основе векторов наблюдений и конечного множества модифицируемых параметров W . Этим способом может быть алгоритм, оператор или некоторая функциональная зависимость:

$$p(x(t_i)) = F(x_i, W).$$

Очевидно, что в представленном варианте модель многошагового предсказания является обучающейся и ее построение сводится к подбору значений вектора управляющих параметров W так, чтобы при поступлении на вход модели аномальных паттернов вероятностные значения вектора предсказаний $p(x(t_i))$ вели себя неким должным образом.

Методология многошагового предсказания. В идеальном случае вероятностные значения $p(x(t_i))$ в векторе предсказаний при поступлении на вход модели аномальных паттернов должны стремиться к "1", причем желательно при минимальных индексах i с тем, чтобы на как можно более ранних стадиях обнаружить аномалии. В этом случае шейп изменения вероятностных значений в векторе предсказаний должен выглядеть так, как показано на рис. 1,б).

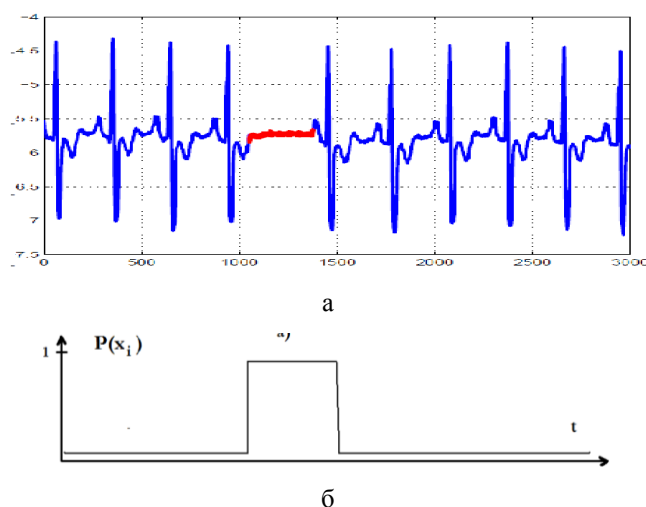


Рис. 1. График наблюдаемого процесса (а), идеальный шейп предсказаний (б)

Получить идеальный шейп предсказания в результате обучения на реальных данных не представляется возможным, поэтому для реализации упреждающего детектирования предлагается использовать иной тип шейпа, приведенный на рис. 2.

Данный шейп имеет пилообразный характер, обладающий двумя достоинствами. Во-первых, он достижим на реальных данных обучения, а, во-вторых, позволяет эффективно осуществлять упреждающее детектирование аномальных паттернов на основе анализа характера поведения шейпа предсказания в начальной стадии развития аномального паттерна. Особенность заключается в затухании колебаний вероятностных значений в преддверии аномалии и переходе к монотонному возрастанию шейпа предсказаний с поступлением новых отсчетов аномального паттерна.

Доходные Марковские модели. Ключевым элементом в методологии многошагового предсказания на основе анализа динамики развития шейпов предсказаний является выбор способа вычисления вероятностных значений аномалий в векторах предсказаний. В качестве такого средства предлагается использовать особый тип функций, имеющих отношение к специальному классу стохастических моделей Марковских процессов с доходами [1, 2] (MR-модели).

В общем виде MR-модель представляет собой триплет:

$$MR = \langle X, \| P \|, W \rangle,$$

X – множество состояний наблюдаемого процесса, $\|P\|$ – матрица переходных вероятностей между состояниями, $W: X \rightarrow R$ – доходная функция.

Введем для MR-модели ряд вспомогательных определений.

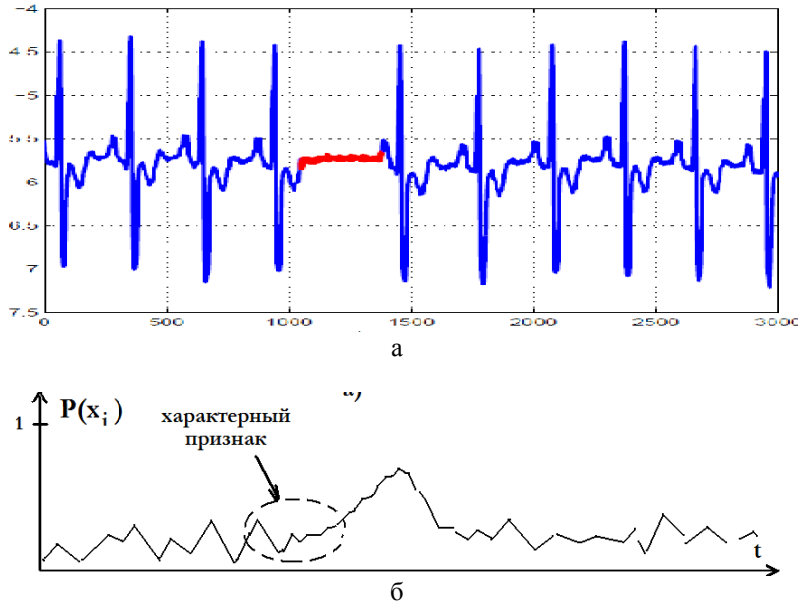


Рис. 2. График наблюдаемого процесса (а), реальный шейп предсказаний (б)

Определение 1. Маршрутом MR-модели называется любая последовательность состояний $X = x_{i_1}, x_{i_2}, \dots, x_{i_k}$.

Заметим, что любому темпоральному паттерну $X = x_{i_1}, x_{i_2}, \dots, x_{i_k}$ наблюдаемого процесса в MR-модели соответствует определенный маршрут.

Определение 2. Вероятностью маршрута $X = x_1, x_2, \dots, x_k$ MR-модели называется величина:

$$P(X) = p_{12} \cdot p_{23} \cdot \dots \cdot p_{k-1k},$$

где p_{ij} – переходные вероятности матрицы $\|P\|$.

Определение 3. Доходом маршрута $X = x_1, x_2, \dots, x_k$ MR-модели называется величина:

$$W(X) = \omega(x_1) + \omega(x_2) + \dots + \omega(x_k),$$

где $\omega(x_i)$ – значения доходной функции.

Базовым понятием, относящимся к MR-модели, является понятие истинностной функции.

Определение 4. Функцией истинности MR-модели называется вещественная функция $E: X \rightarrow R$, сопоставляющая каждому состоянию $x \in X$ величину $E(x)$, равную математическому ожиданию дохода по всем маршрутам, исходящим из x :

$$E(x) = \sum_{X=x_1, x_2, \dots, x_k \in \Theta / x_1=x} P(X) \cdot W(X),$$

где Θ – множество всех маршрутов в MR-модели.

Фактически, функция истинности сопоставляет каждому состоянию средний доход по всем маршрутам, исходящими из состояния x . или, иначе, прогнозирует доход, полученный в результате реализации Марковского процесса, стартующего из состояния x .

Приведенное общее определение MR-модели допускает различные конкретизации в разных приложениях в зависимости от свойств доходной функции. Применительно к рассматриваемой задаче поиска аномалий предполагается заданными два множества примеров аномальных A и нормальных N темпоральных паттернов, и два типа связанных с ними ограничений на доходную функцию:

$$\begin{aligned} \forall X = (x_1, x_2, \dots, x_k) \in A \quad \sum_{i=1}^k \omega(x_i) &= 1, \\ \forall X = (x_1, x_2, \dots, x_k) \in N \quad \sum_{i=1}^k \omega(x_i) &= 0. \end{aligned} \quad (1)$$

Определение 5. Целевой вероятностью $P_g(x)$ состояния $x \in X$ называется вероятность появления аномального паттерна в множестве всех темпоральных паттернов $\Theta^O = A \cup N$, начинающегося с состояния x :

$$P_g(x) = P\{(x_{i_1}, x_{i_2}, \dots, x_{i_k}) \in A \mid x_{i_1} = x\}.$$

Следующая теорема устанавливает связь между значениями функции истинности доходной Марковской модели и вероятностями появления аномальных паттернов.

Теорема. Для любого состояния $x \in X$ MR-модели, удовлетворяющей ограничениям (1), значение функции истинности $E(x)$ равно целевой вероятности состояния x :

$$E(x) = P_g(x).$$

На основе данной Теоремы можно предложить конструктивный способ вычисления вероятностных значений вектора предсказаний через вычисления функции истинности MR-модели. Истинностная функция MR-модели с достаточно высокой точностью аппроксимируется линейным выражением:

$$E(x_i) = \sum_j p_{ij} \cdot \omega(x_j). \quad (2)$$

Выражение (2) представляет эффективное средство вычисления вероятностных значений предсказывающего вектора в рассматриваемой методологии многошагового предсказания аномалий в темпоральных данных.

Другим важным выводом из данной теоремы является возможность использования нового подхода к обучению модели многошагового предсказания, через обучение предсказанию значений функции истинности MR-модели. Для этого можно использовать методы эволюционного моделирования [3]. Однако более эффективной является специальная методология темпорально-разностного обучения (TD-обучение) [4].

Нечеткая система преобразования процесса. При практической реализации методологии многошагового предсказания на основе MR-моделей возникает проблема, обусловленная Марковскими ограничениями на переходные вероятности между состояниями наблюдаемого процесса:

$$P\{x_{t+1} \mid x_t, x_{t-1}, \dots, x_{t-k}\} = P\{x_{t+1} \mid x_t\}.$$

В реальных задачах это ограничение далеко не всегда выполняется. И, тем не менее, данную проблему можно решить путем преобразования исходного процесса к квази-Марковскому путем соответствующего выбора пространства состояний MR-модели, опираясь на положения известной теории “русел и джокеров” [5].

Теория “русел и джокеров”, основывается на классе динамических моделей с “джокерами”, в которых “джокеры” представляют собой особые области фазового пространства, где поведение системы становится непредсказуемым. В этих областях “джокер” может “забросить” траекторию процесса в любую точку фазового пространства. Естественно предположить, что вероятность этого скачкообразного перехода зависит исключительно от феномена “джокера” и не зависит от истории перехода процесса к “джокеру”. Тогда, исходя из данной гипотезы, можно в качестве состояний MR-модели выбирать не точечные состояния исходного наблюдаемого процесса $x \in X$, а совокупности этих состояний $Q(x)$, агрегированные на новой шкале Q и определенные на временных интервалах длительностью, перекрывающей длительность “русел”. Наличие “джокера” в каждом из гранулированных состояний нового процесса будет обеспечивать независимость вероятностных переходов, придавая преобразованному процессу характер Марковского процесса.

Для преобразования исходного процесса в квази-Марковский процесс предлагается использовать нечеткую систему, осуществляющую отображение совокупностей точечных состояний исходного процесса X на новую шкалу нечеткого интегрального признака, индуцируя на ней новый процесс в виде нечеткого временного ряда. В задачах предсказания сбоев в расписании движения поездов такой шкалой может быть виртуальная шкала “Напряженности”, на которую отображаются точечные состояния исходного процесса, такие как, время задержки поезда, категория поезда, особенности путевого развития станции и др. Для того, чтобы полученный новый временной процесс приближался к Марковскому необходимо обеспечить адаптируемость нечеткой системы по ширине захвата области определения гранулированных состояний в исходном процессе. Для этого используется специальный класс параметрических нечетких систем, с группой модифицируемых параметров V , обеспечивающих такую настройку. Включение такой адаптивной нечеткой системы в MR-модель, приводит к гибридной системе многошагового предсказания с двумя группами модифицируемых параметров W и V , приведенной на рис. 3.



Рис. 3. Гибридная система многошагового предсказания с двумя группами модифицируемых параметров

Первая группа параметров W обеспечивает адаптацию системы под “пилообразный” характер шейпа предсказаний, а вторая – преобразование исходного процесса к квази-Марковскому. Поскольку оба критерия являются разносортными и не конфликтующими, процесс адаптации гибридной системы можно проводить независимо по обеим группам параметров в едином цикле обучения.

Заключение. В статье предложена новая методология упреждающего детектирования аномальных паттернов в темпоральных данных на основе метода многошагового предсказания с использованием адаптивных Марковских моделей. Предлагаемый подход может найти широкое применение в задачах распознавания аномалий во ВР, автоматического извлечения типовых паттернов в потоках данных, формирования темпоральных баз знаний и др.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

- 1 *Yeung D.Y., Ding Y.X.* Host-based intrusion detection using dynamic and static behavioral models, *Pattern Recognition*. – 2003. – № 36. – С. 229-243.
- 2 *Ковалев С.М.* Упреждающее распознавание нечетких темпоральных паттернов в потоковых данных // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012: Тр. конференции. Т. 2. – М.: Физматлит, 2012. – С. 313-322.
- 3 *Курейчик В.В., Курейчик В.М., Сороколетов П.В.* Анализ и обзор моделей эволюции. *Известия РАН // Теория и системы управления*. – 2007. – № 5.
- 4 *Sutton R.* Learning to predict by the method of temporal differences // *Machine Learning*. – 1988. – № 3 (1). – С. 9-44.
- 5 *Малинейкий Г.Г., Потапов А.Б.* Руслы и джокеры: о новых методах прогноза поведения сложных систем // Препринт ИМП ИИ. им. М.В. Келдыша РАН. 2001.

Статью рекомендовал к опубликованию д.т.н., профессор А.Н. Шабельников.

Ковалев Сергей Михайлович – Ростовский государственный университет путей сообщения; e-mail: ksm@real36.com; 344038, г. Ростов-на-Дону, пл. Ростовского Стрелкового Полка Народного Ополчения, 2, г; тел.: 89612687722; д.т.н.; профессор кафедры автоматизации и телемеханики на ж. д. транспорте.

Kovalev Sergey Mihailovich – Rostov State University of Transport; e-mail: ksm@real36.com; 2, g, pl. Rostov Rifle Regiment of the People's Militia, Rostov-on-Don, 344038, Russia; phone: +79612687722; dr. of eng. sc.; professor the department of automation and remote control at the railway on transport.

УДК 681.3

Ю.О. Чернышев, Н.Н. Венцов, С.А. Мухтаров

**ПРИМЕНЕНИЕ ЛОГИК ЛУКАСЕВИЧА И ЗАДЕ ПРИ РЕАЛИЗАЦИИ
МЕТОДА ОТРИЦАТЕЛЬНОГО ОТБОРА***

Описан метод отрицательного отбора. Приведен пример использования метода отрицательного отбора для анализа процесса передачи данных. Проанализирована целесообразность использования операций импликации и эквивалентности на основе логик Лукасевича и Заде при нечетком анализе информационных процессов. Приведены примеры сравнения декодера, описывающего критическое состояние процесса передачи данных, с кортежами описывающими текущее состояние процесса передачи данных при помощи операций импликации и эквивалентности. Установлено что для поставленной задачи анализа процессов передачи данных применение операции импликации является более предпочтительным. При выполнении операции импликации целесообразно использовать как логику Лукасевича так и логику Заде.

Вычислительные ресурсы; отрицательный отбор; интеллектуальные системы; оптимизация.

Yu.O. Chernyshev, N.N. Vencov, S.A. Mukhtarov

**APPLICATION LOGICS LUKASIEWICZ AND ZADEH IN THE METHOD
OF ADVERSE SELECTION***

The paper describes a method of adverse selection. An example using the method of adverse selection for the analysis of the data transfer process. Analyzed the implications of the use of operations and equivalence based on the logics of lukasiewicz and Zade when fuzzy analysis of informational processes. Examples of comparison decoder, describing the critical state of the data

* Работа выполнена при финансовой поддержке РФФИ (проекты: 12-01-00474, 13-01-00343).