ния в разных ситуациях. Представлен генетический алгоритм с динамическим выбором генетических операторов в группе, решающий задачу выбора оптимального управляющего воздействия для перевода изучаемой системы в устойчивое, более эффективное, с точки зрения функционирования, состояние.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Френкель М.Б., Квятковская И.Ю.* Моделирование сложных социально-экономических систем с учетом влияния внешней среды // Вестник АГТУ. Сер. Управление, вычислительная техника и информатика. – 2009. – № 2.
2. *Петров Ю.Ю.* Управляемые генетические алгоритмы, основанные на статистике // Вторая Всероссийская научная конференция «Нечеткие системы и мягкие вычисления». – Ульяновск, 2008.

Статью рекомендовал к опубликованию д.т.н., профессор В.В. Тютиков.

Бородулина Екатерина Николаевна – Южный федеральный университет; e-mail: kaf_sau@mail.ru; 344090, г. Ростов-на-Дону, ул. Мильчакова, 10, каб. 505; тел.: 88632696991, 89044434417; кафедра системного анализа и управления; аспирантка; преподаватель.

Borodulina Ekaterina Nikolaevna – Southern Federal University; e-mail: stervyshka@mail.ru; 10, Milchakova street, of. 505, Rostov-on-Don, 344090 Russia; phones: +78632696991, +79044434417; the department of systems analysis and control; postgraduate student; instructor.

УДК 656.2 + 06

С.М. Ковалев, А.В. Суханов

ОБНАРУЖЕНИЕ ОСОБЫХ ТИПОВ ПАТТЕРНОВ ВО ВРЕМЕННЫХ РЯДАХ НА ОСНОВЕ ГИБРИДНОЙ СТОХАСТИЧЕСКОЙ МОДЕЛИ*

В настоящее время наблюдается широкое внедрение автоматизированных информационно-управляющих систем, основанных на базах данных и знаний. В связи с этим появляется необходимость компьютерного анализа больших объемов информации, полученной в результате наблюдений за работой технических устройств и напольного оборудования. Здесь для выявления и обобщения полезной информации, а также для формирования баз знаний используются различные методы обработки темпоральных данных, в частности методы классификации и кластеризации временных рядов. В статье рассматривается одна из наиболее важных задач в области интеллектуального анализа данных, связанная с обнаружением особых типов темпоральных паттернов во временных рядах. Предлагаемый метод основан на обучении без учителя Марковской модели исследуемой системы с продукционными правилами, описывающими ее «немарковские» ситуации. Представленный подход к классификации применим для решения широкого круга задач, так как не требует знаний обо всех линиях поведения. Приведенные эксперименты на одной из реализаций типового образца временного ряда доказывают актуальность применения метода для выявления особых типов темпоральных паттернов.

Обнаружение аномалий; обучение без учителя; временной ряд; Марковская модель; продукционные модели; особые паттерны.

* Работа выполнена при поддержке грантов РФФИ, № № 13-07-00183 А, 13-08-12151 офи_м, 13-07-13108 офи_м_РЖД , 13-07-13109 офи_м_РЖД , 14-01-00259 А.

S.M. Kovalev, A.V. Sukhanov

SPECIAL TEMPORAL PATTERN RECOGNITION TECHNIQUE BASED ON HYBRID STOCHASTIC MODEL

There is a wide introduction of automated information management systems based on databases and knowledge base in our days. Therefore there is a computer need in analysis of large information volumes received as a result of technical facilities and equipment work observing. Here various temporal data processing techniques are used for the identification and compilation of useful information. In particular there is time series clustering and classification techniques. This paper presents one of important problem decision in Data Mining dedicated to specific temporal patterns detection. Proposed technique based on unsupervised training of Markov chain model with productional" non-Markov" rules. Such approach could be used for wide problem decision because it is robust forthe lack of information. Represented experiments in one of time series standard sample implementation demonstrates relevance of such techniques for special pattern detection in temporal sets.

Anomaly detection; unsupervised learning; time series; Markov chain; productional models; special pattern recognition.

Введение. В настоящее время широкое развитие получили методы обработки темпоральных данных, использующие машинное обучение [1]. Результат анализа отечественной и зарубежной литературы показал, что "Temporal data classification" результативно используется во многих областях – от информатики и вычислительной техники до экономики и финансовой математики. Методы анализа темпоральных баз данных применяются и на транспорте, где широкое внедрение автоматизированных информационно-управляющих систем, основанных на базах данных и знаний, вызывает необходимость компьютерного анализа больших объёмов экспериментальных данных, полученных в результате наблюдений за работой технических устройств и напольного оборудования [2]. Здесь методы обработки темпоральных данных, в частности методы классификации и кластеризации временных рядов, используются для обобщения полезной информации и формирования баз знаний прикладных интеллектуальных систем.

В настоящей статье рассматривается одна из наиболее важных задач в области интеллектуального анализа данных, связанная с обнаружением особых типов темпоральных паттернов во временных рядах. Под особыми типами темпоральных паттернов понимаются фрагменты временных рядов, являющиеся нетипичными для рассматриваемого класса объектов и характеризующие аномальное развитие контролируемого процесса, не удовлетворяющее некоторому типовому поведению. Описанный в статье метод применим для решения широкого круга задач в областях диагностики, информационно-технологического контроля и обеспечения компьютерной безопасности.

Описываемый метод основан на использовании модифицированной Марковской модели процесса, включающей систему продукционных правил, используемых для корректировки вероятностей перехода с учетом предыстории процесса.

Состояние проблемы и постановка задачи. Разработка методов обнаружения аномалий в темпоральных данных сопряжена с рядом сложностей [4]. Во-первых, достаточно трудно определить типичный темпоральный профиль для временного ряда, описывающий все варианты нормального поведения контролируемого процесса – линии нормального поведения. Во-вторых, далеко не всегда удается отделить линии нормального поведения от аномалий. И, в-третьих, реальные процессы всегда подвержены воздействию различного рода шумов и искажений, в результате чего наблюдаемые данные становятся схожими с аномалиями, что создает трудности для их распознавания. Для представления типовых темпоральных

профилей зашумленных временных процессов используются методы, основанные на построении стохастических моделей процессов [2, 3]. При правильном построении стохастической модели можно формировать вероятностные описания линий нормального поведения, на основе которых предсказывается исход процесса и детектируются редко встречающиеся события. Для решения второй проблемы активно разрабатываются методы одноклассовой классификации [5], являющиеся разновидностью известных методов классификации на основе обучения «частично с учителем». Одноклассовый подход к классификации актуален при полном отсутствии примеров аномального поведения, что характерно для экспериментальных данных, собранных в результате мониторинга технических устройств и технологических процессов. Устойчивыми к неопределенностям являются методы, основанные на нечеткой логике и мягких вычислениях [6].

В связи с постоянным появлением новых задач по обнаружению аномалий в темпоральных базах данных, содержащих специфические условия и требования для конечного результата, существующие методы требуют доработки и дополнений.

Стохастическая модель нормального поведения. Дискретный по времени процесс представим в виде символьного временного ряда

$$X = x(1), x(2), \ldots, x(t), \ldots, x(N),$$

где $x(t) \in S$ – состояние процесса в момент времени $t \in N$; S – множество оригинальных состояний процесса,

$$S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}.$$

Реальные процессы подвержены влиянию шумов, поэтому их состояние в моменты времени t определяется через законы распределения вероятностей. Наиболее изученными являются Марковские процессы, для которых приняты следующие допущения:

- распределение вероятностей состояния в момент времени t зависит только от состояния процесса в момент времени $t–1$, и не зависит от предыдущих состояний

$$p(x(t) \mid x(t-1)) = p(x(t) \mid x(t), x(t-1), x(t-2), \ldots, x(t-l)); \quad (1)$$

- распределение вероятностей перехода из одного состояния в другое не зависит от времени.

Для описания Марковских процессов используются Марковские модели. В [3] говорится о Марковской модели нормального поведения, представляемой кортежем:

$$MM=<S, Q, P(x(t)|x(t\text{-}1))>,$$

где S – множество оригинальных состояний процесса; Q – вектор начального распределения вероятностей; P – матрица переходных вероятностей. Вероятностные параметры Марковской модели определяются на основе вычислений с использованием следующих формул:

$$Q = \{q(s_i)\}, \quad (2)$$

$$q(s_i) = \frac{N_i}{N}.$$

где N_i – количество появления состояния s_i; N – общее количество элементов исследуемого процесса;

$$P(x(t)|x(t\text{-}1)) = \{p(x(t) = s_j \mid x(t-1) = s_i)\},$$

$$p(x(t) = s_j \mid x(t-1) = s_i) = \frac{N_{ij}}{N_i}, \quad (3)$$

где N_{ij} – количество появления цепочки состояний $s_i s_j$.

Классификация на основе Марковской модели нормального поведения состоит в вычислении уровня поддержки тестового паттерна.

Пусть X_1 – тестовый темпоральный паттерн, представленный цепочкой символов $x_1(1), x_1(2),\ldots,x_1(\theta),\ldots,x_1(N_1)$. Классификация данного паттерна согласно [2] и [3] заключается в вычислении его поддержки Марковской моделью:

$$supp = q(x_1(1)) \cdot \prod_{t=2}^{|X_1|} p(x_1(t) \,|x_1(t-1)), \tag{4}$$

где $q(x_1(1)) = q(x(1) = s_i)$ при $x_1(1) = s_i$; $p(x_1(t)|x_1(t-1)) = p(x(t) = s_j|x(t-1) = s_i)$ при $x_1(t) = s_j$ и $x_1(t-1) = s_i$.

Гибридная стохастическая модель профиля нормального поведения процесса, предложенная в [2], представляет собой четверку

$$GM=<S, Q, P(x(t)|x(t\text{-}1)), \Pi>, \tag{5}$$

где Π – система уточняющих продукционных темпоральных правил (ПТП).

ПТП формируются для состояний $x(t-1)=s_i$, не удовлетворяющих условию (1), устанавливая для них новые вероятности перехода:

$$p(x(t) = s_j \mid x(t-1) = s_i) = \\ = p(x(t) = s_j \mid x(t-1) = s_i, x(t\text{-}2) = s_k, \ldots, x(t-l) = s_z), \tag{6}$$

где $N_{z\ldots kij}$ – количество повторений в символьном ряду цепочки *состояний z, …, k, i, j*. $N_{z\ldots ki}$ – количество повторений вложенной цепочки *z, …, k, i*.

Пусть паттерн $A = [a_1,a_2,\ldots,a_n]$, а паттерн $B = [aj,a_{J+1},\ldots,a_n]$. Тогда говорят, что паттерн *B* является подпаттерном *A*, а ПТП $p(A)$ является доминирующим над ПТП $p(B)$.

При классификации паттерна X_1 на основе гибридной стохастической модели сначала проверяется «Марковость» каждого $x_1(\theta)$, после чего при отрицательном результате для них выполняется поиск ПТП. В итоге правило является уточняющим для состояния $x_1(\theta)$, если оно доминирует над остальными правилами множества ПТП.

Метод классификации на основе стохастической модели. В основу предлагаемого метода положена стохастическая модель временного ряда *X*, построенная на основе метода обучения “без учителя”. При этом принято допущение, что количество аномалий во временном ряду значительно меньше числа нормальных данных. Такая модель описывается кортежем из четырех элементов (5), где *Q* вычисляется по формуле (2), *P* – по формуле (3), а Π устанавливает правила по формуле (6) для элементов *P*, не удовлетворяющих условию (1).

Для снижения вычислительной сложности при построении Π введем уточнение для выбора максимальной длины цепочки состояний ПТП. Под максимальной длиной цепочки состояний ПТП будем понимать значение l_{max}, такое, что ПТП, устанавливающие вероятности для последовательностей состояний предыстории длиной $l > l_{max}$, будут незначительно влиять на классификацию паттернов.

Анализ формулы (4) показал, что вычисляемая на ее основе поддержка тестового темпорального паттерна “быстро” стремится к нулю при увеличении длины паттерна. В этой связи для классификации используется иная формула, предложенная в [7] для повышения устойчивости результатов умножения при малых значениях множителей:

$$supp = \frac{1}{\frac{1}{N_1} \cdot \left(\frac{1}{q(x(1))} \cdot \sum_{i=l_{\max}}^{|X_1|} \frac{1}{p(x(i) \mid x(i-1))}\right)}. \tag{7}$$

При классификации каждое состояние x(t) временного ряда X заменяется паттерном длиной $\delta = \mu \cdot l_{max}$

$$x(t) \mapsto pat(t) = [x(t), x(t+1), \ldots x(\mathrm{t}+\delta-1)].$$

Чем выше значение μ, тем легче отличить паттерны, содержащие аномалии, но тем сложнее локализовать конкретную аномалию. Поэтому μ следует выбирать в соответствии с эмпирическими соображениями о соотношении точность/качество распознавания (в экспериментах, выполненных в рамках настоящей работы, μ = 5).

Для каждого pat(t) производятся вычисления поддержки стохастической моделью, в результате чего создается вектор SUPP={supp(t)}. При наличии в SUPP элементов с существенно малыми относительно других значениями, паттерны, соответствующие этим элементам, будут являться особыми.

Вычислительные эксперименты. Для проверки эффективности предложенного метода была проведена серия экспериментов, в которых тестовые примеры были составлены из образцов известных типов темпоральных паттернов, в частности, фрагментов известного числового временного ряда Coffee [8]. Последний был разбит на 28 отрезков, после чего приведен к символьному представлению путем дискретизации на 17 состояний и разделен на 7 темпоральных паттернов одинаковой длины 4 (рис. 1).

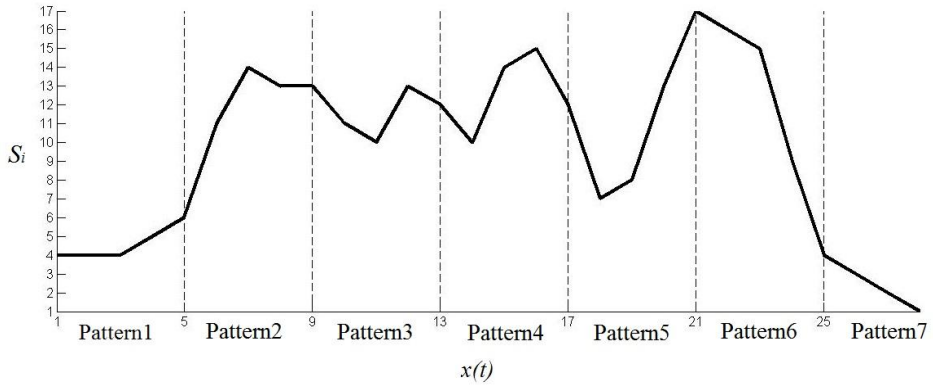


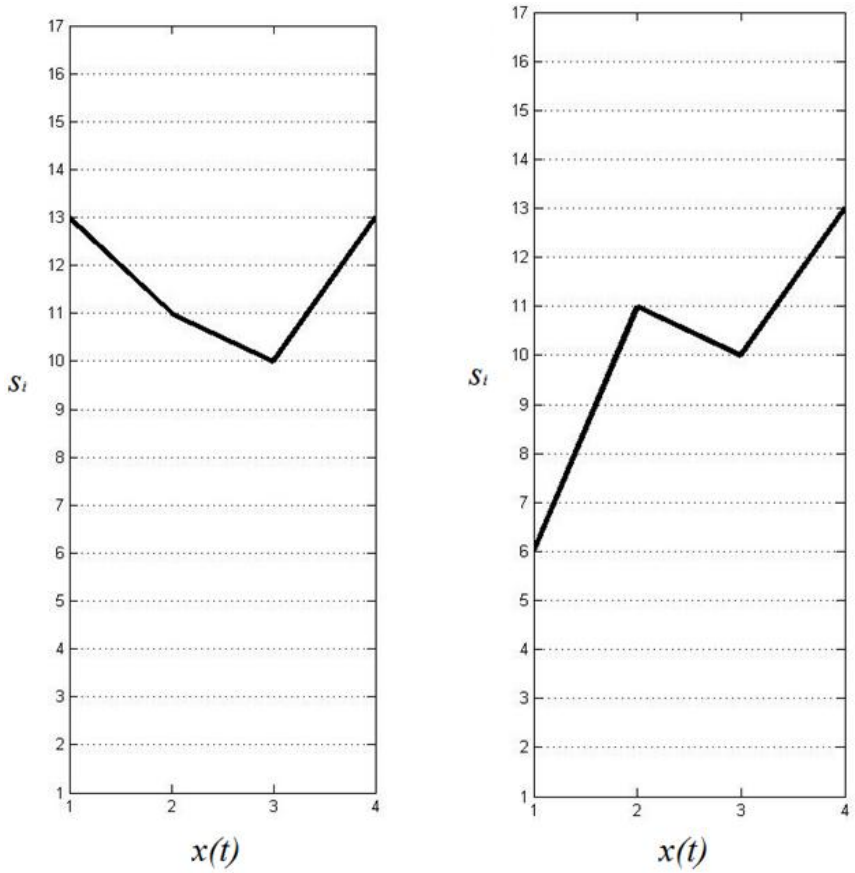
Рис. 1. Реализация Coffee (пунктирными линиями указано разделение на паттерны)

Таким образом, исходные данные были сформированы из семи типов темпоральных паттернов длины 4, принимающих 17 значений.

Алгоритм реализован в среде Matlab. Тестовый временной ряд был составлен случайным перемешиванием типовых темпоральных паттернов и включал 20 000 символов. В тестовом ряду несколько паттернов типа pattern3 = (13 11 10 13) были заменены на паттерны pattern3_wrong = (6 11 10 13) (рис. 2), принятые за особые, таким образом, что

$$\frac{N_{pattern3}}{N_{pattern3\_wrong}} = 7.$$

Следовательно, соотношение количества аномальных паттернов к количеству нормальных равнялось ~1:50. Такой вид аномального паттерна был принят из соображений наибольшей его схожести с нормальными данными (например, с паттернами pattern2 = (6 11 14 13)).



а б

Рис. 2. Нормальный паттерн Pattern3 и паттерн Pattern3_wrong, принятый за аномалию

Коэффициент l_{max} был определен из зависимости среднего количества повторения оригинальных последовательностей фиксированной длины l в исследуемом множестве от установленного значения l (рис. 3).

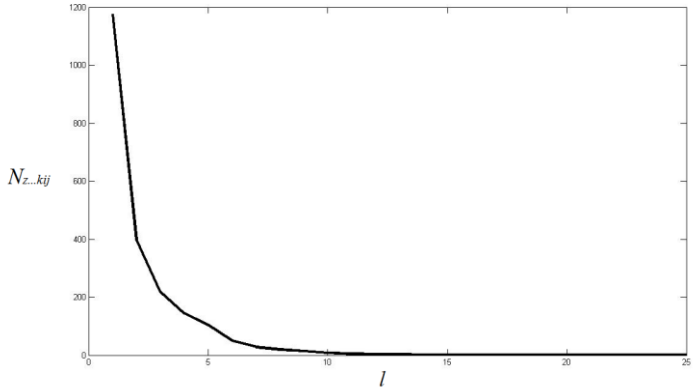


Рис. 3. Зависимость средней повторяемости последовательностей от длины

Проанализировав полученную зависимость, можно точно сказать, что при $l > 10$ ПТП окажут незначительное влияние на результаты классификации. Следовательно, $l_{max} = 10$.

На рис. 4 показаны результаты применения метода обнаружения аномалий, основанного на обычной Марковской модели и гибридной Марковской модели с продукционными правилами.

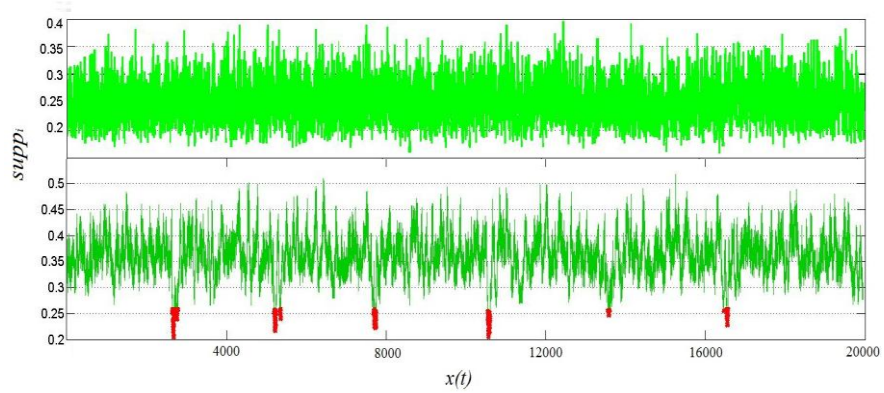


Рис. 4. График поддержки состояний временного ряда Марковской моделью (сверху) и гибридной стохастической моделью поведения (снизу цветом выделена поддержка особых паттернов)

Как видно из полученных данных, поддержка большинства нормальных паттернов заметно превосходит поддержку особых типов паттернов.

Для наглядности демонстрации метода на рис. 5 представлен фрагмент временного ряда с аномалиями и график поддержки его состояний.

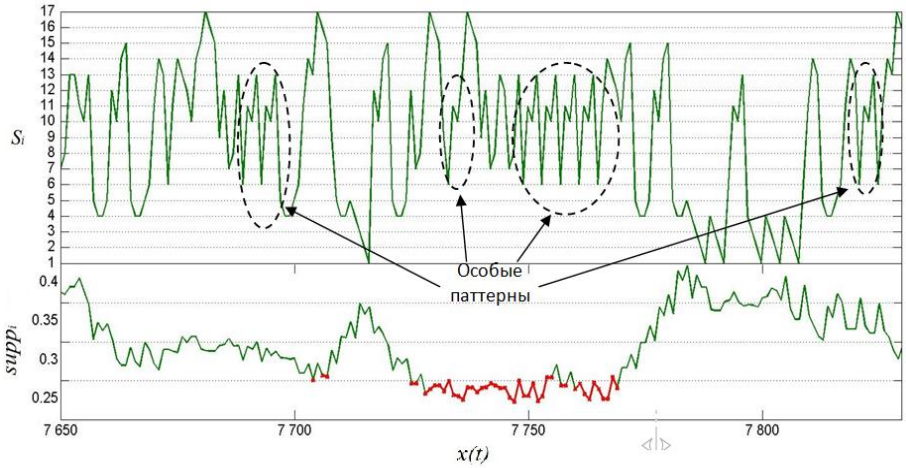


Рис. 5. Фрагмент исследуемого временного ряда и его поддержка (цветом выделена поддержка особых паттернов)

Точность классификации временного ряда составила 89,7 %.

Выводы. Предложенный гибридный подход к обнаружению особых типов темпоральных паттернов, основанный на гибридной Марковской модели временного процесса с продукционными правилами, описывающими случаи, когда про-

цесс перестает подчиняться условиям «Марковости», является одним из методов обучения “без учителя”. Описанный подход применим к немаркированным данным и устойчив к шумам. Эффективность предложенного метода подтверждается результатами экспериментов на известных примерах, используемых при тестировании методов обнаружения аномалий в темпоральных базах данных. При этом точность классификации составила 89,7 %.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Grabocka J., Nanopoulos A., Schmidt-Thieme L.* Invariant Time-Series Classification // European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). – 2012. – P. 725-740.
2. *Ковалев С.М., Гуда А.Н., Бутакова М.А.* Гибридная стохастическая модель обнаружения особых типов паттернов в темпоральных данных // Вестник РГУПС. – 2013. – № 3 (51). – С. 36-42.
3. *Суханов А.В.* Стохастическая Марковская модель поиска аномалий в темпоральных данных // Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT’13»: В 4 т. – М.: Физматлит, 2013. – Т. 1. – С. 177-181.
4. *Chandola V., Banerjee A., Kumar V.* Anomaly Detection: A Survey // ACM Computing Surveys, 2009. – Vol. 41(3). Article 15. – P. 1-72.
5. *Ma J., Perkins S.* Time-series novelty detection using one-class support vector machines // Proceedings of the International Joint Conference on Neural Networks. – July 2003. – Vol. 3. – P. 1741–1745.
6. *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 166 с.
7. *Scheuner U.* Fuzzy-Mengen Verknüpfung und Fuzzy-Arithmetik zur Sensordaten-Fusion // VDI-Verlag. – 2001. – Bd. 8.
8. *Keogh E., Xi X., Wei L., Ratanamahatana C.A.* The UCR time series classification/clustering homepage // http://www.cs.ucr.edu/ ~eamonn/time_series_data/.

Статью рекомендовал к опубликованию д.т.н., профессор Е.А. Башков.

Ковалев Сергей Михайлович – Ростовский государственный университет путей сообщения (РГУПС); e-mail: ksm@rfniias.ru; 344038, г. Ростов-на-Дону, пл. Ростовского Стрелкового Полка Народного Ополчения, 2; тел.: 88632726302; кафедра автоматики и телемеханики на железнодорожном транспорте; д.т.н.; профессор.

Суханов Андрей Валерьевич – e-mail: drewnia@rambler.ru; кафедра автоматики и телемеханики на железнодорожном транспорте; аспирант.

Kovalev Sergey Mikhailovich – Rostov State Transport University (RSTU); e-mail: ksm@rfniias.ru; 2, Rostovskogo Strelkovogo Polka Narodnogo Opolcheniya sq., Rostov-on-Don, 344038, Russia; phone: 88632726302; the department of automatics and telemechanics on railway transport; dr. of eng. sc.; professor.

Sukhanov Andrey Valerievich – e-mail: drewnia@rambler.ru; the department of automatics and telemechanics on railway transport; postgraduate student.