
СИСТЕМНЫЙ АНАЛИЗ, МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

УДК 004.032.26 + 338.27

МЕТОД АВТОМАТИЧЕСКОГО ВОССТАНОВЛЕНИЯ ЗНАЧЕНИЙ В ПОТОКАХ ДАННЫХ НА ОСНОВЕ ВЗВЕШЕННОЙ МОДЕЛИ¹

Маркин Алексей Валерьевич, магистрант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 65, e-mail: maxim.shcherbakov@vstu.ru

Щербаков Максим Владимирович, кандидат технических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 65, e-mail: maxim.shcherbakov@vstu.ru

В статье рассматривается решение проблемы заполнения пропусков в данных временных рядов. Данная проблема встречается в процессе первичной обработки считанных данных. Игнорировать эту проблему не рекомендуется, так как дальнейший анализ данных с пропусками приведёт в итоге к неудовлетворительным результатам. Для решения этой проблемы предлагается алгоритм восстановления данных, основанный на использовании взвешенных значений, полученных в результате применения различных моделей и методов. В качестве моделей восстановления используются модифицированный метод *k*-ближайших соседей и простая средняя модель. Представлены результаты вычислительных экспериментов для различного количества последовательно пропущенных точек в данных относительно каждого из сезонов года. Даны рекомендации по выбору весов при определении восстановленных значений взвешенным усреднением результатов, полученных разными методами.

Ключевые слова: заполнение пропусков, временные ряды, *k*-ближайших соседей, простая средняя модель, гибрид, энергопотребление, прогнозирование, весовой коэффициент, средняя статистическая ошибка, автоматическое восстановление

A METHOD FOR DATA IMPUTATION BASED ON WEIGHTED BACKCAST MODELS SELECTION

Markin Aleksey V., undergraduate student, Volgograd State Technical University, 65 Lenin Avenue, Volgograd, 400005, Russian Federation, e-mail: maxim.shcherbakov@vstu.ru

Shcherbakov Maksim V., Ph.D. (Engineering), Associate Professor, Volgograd State Technical University, 65 Lenin Avenue, Volgograd, 400005, Russian Federation, e-mail: maxim.shcherbakov@vstu.ru

The article deals with the problem data imputation in time series. This problem occurs in the process of preprocessing the read data. Ignore the problem is not recommended, as further analysis of the data gaps will result in unsatisfactory results. To solve this problem, a data imputation algorithm, based on a weighted score obtained by the use of different models and methods. As the imputation models used a modified method of *k* nearest neighbor and simple mean model. The results of computational experiments for different numbers of consistently missing data points in relation to each of the seasons and recommendations on the choice of weights the importance of a particular method.

Keywords: data imputation, time series, KNN, Sams, hybrid, energy consumption, forecasting, weight coefficient, sMAPE, automatic imputation

¹ Работа выполнена при финансовой поддержки РФФИ (гранты № 12-01-00684, 12-07-31017).

Введение. Проблема первичной обработки данных, представленных в виде временных рядов, является весьма актуальной. От ее решения зависят результаты последующей обработки таких данных, например, прогнозирование временных рядов. Одной из распространенных трудностей, связанных с первичной обработкой, является наличие пропусков [1, 2, 4] в данных. Причины возникновения пропусков различны: сбой в оборудовании, устройствах передачи информации, ошибки персонала и др. При наличии пропусков в данных процесс прогнозирования потребления электроэнергии затруднен или даже невозможен.

Целью работы было создание эффективного алгоритма восполнения пропусков в данных и опробование его с помощью вычислительных экспериментов.

1. Постановка задачи и обзор существующих подходов. Постановка задачи восполнения пропусков в данных может быть сформулирована следующим образом. Пусть имеется временной ряд Y , в котором существуют пропуски в данных на определенном интервале $[t_0; t_k]$. Подобные интервалы различной длины могут встречаться на протяжении всего исследуемого временного ряда. Требуется разработать механизм M заполнения пропусков в данных, при котором критерий ошибки восстановления J стремился бы к минимуму. Разрабатываемый алгоритм направлен на решение одной из задач первичной обработки данных с пропущенными значениями во временных рядах, характеризующих потребление электроэнергии определенного класса потребителей. К нему относятся офисные здания, производства с циклическим характером работы по определенному графику (часы открытия, закрытия, выходные дни и праздники). Этот график в различные рабочие дни может варьироваться (например, в понедельник часы открытия 8-00, закрытия 18-00, четверг – открытие в 09-00, закрытие в 22-00). Потребление электроэнергии в зданиях значительно зависит от графика работы. Так, в ночные часы или в праздники оно может отличаться от дневного рабочего в 10–12 раз. Данные о потреблении электроэнергии собираются и передаются на сервер сбора данных каждые 15 минут [3]. Таким образом, временной ряд получается равномерным по времени. В данной работе эти данные без пропусков используются как «эталонные» и применяются для сравнения «восстановленных значений» с фактическими. В отечественной литературе разработчиками и исследователями было предложено много методов оценки и восстановления пропусков в данных. О. Троянская представила метод восстановления пропущенных данных на основе алгоритма ближайшего соседа [10]. Томпсон занимался автокодерами, которые восстанавливали пропущенные данные, используя алгоритм нейронных сетей [9]. В. Цяо с соавторами предложили гибридный алгоритм определения значений пропущенных данных на основе модели нейронной сети и метода роя частиц [8], этот алгоритм был использован для отслеживания динамики возникновения пропусков в датчике. М. Абдула применил генетические алгоритмы и нейронные сети для аппроксимации пропусков в базе данных [5]. Предложенный С. Мохаммедом подход основывается на мультиагентах и нейронных сетях – он также направлен на оценку недостающих данных [6]. Эффективность применения указанных подходов сильно зависит от специфики рассматриваемых предметных областей. Следовательно, они априорно мало пригодны для восстановления пропусков в рядах, характерных для энергоснабжения. Предлагаемый алгоритм основан на комбинации двух методов, специально переработанных для анализа временных рядов, связанных с энергопотреблением. Алгоритм предназначен для восстановления пропущенных значений – одиночных отсчетов и их групп.

2. Предлагаемый алгоритм заполнения пропусков. В основе алгоритма лежит использование комбинации двух методов заполнения пропусков: первый – метод на основе алгоритма ближайшего соседа (KNN), второй – на основе простой средней модели. Эти методы достаточно просты в реализации и показывают хорошие результаты на тестовой выборке. Для каждого из них вводятся веса значимости (весовые коэффициенты), на основании которых формируется результирующее пропущенное значение (значения).

Алгоритм ближайших соседей. Выбирается число k – необходимое количество частей временного ряда, которые будут взяты для анализа и восстановления данных в той части, где присутствуют пропуски. Набор данных разбивается на две части: первая содержит месяцы, в

которых замечены пропуски, во второй – остальные месяцы. Пусть D_m – массив экземпляров, которые содержат хотя бы один пропуск, а D_c – массив экземпляров без пропусков. Для каждого вектора (экземпляра) x из D_m выделим 2 части: вектор с пропущенными данными – x_m и вектор с полными данными x_o . Для того чтобы выбрать k наиболее подходящих месяцев, необходимо рассчитать разницу между точками вектора x_o и векторов из полного набора D_c :

$$S(x) = \sum_{i=1}^n \sqrt{(y_i - y'_i)^2}, \quad (1)$$

где n – количество известных точек в одном месяце с пропущенными данными.

Рассчитав разницу относительно каждого экземпляра, из D_c выбираем k наименьших результата. Соответствующие им месяцы и будут «ближайшими соседями» месяца с пропусками. Затем для каждого пропуска вектора из массива D_m мы делаем выборку из «соседних» векторов. Зная день и точное время пропуска, из числа соседей выбираются аналогичные по дню недели и времени точки временного ряда, а затем рассчитывается их среднее значение:

$$y_o = \frac{1}{k} \sum_{i=1}^k \sqrt{(y_i - y'_i)^2}. \quad (2)$$

Метод на основе простой средней модели. На место пропущенной точки подставляется среднее арифметическое суммы точек, расположенных на определённой глубине погружения:

$$y_o = \frac{1}{n} \sum_{i=1}^n y_i(t), \quad (3)$$

где

$$y_o(t) = y_o(t - i * h). \quad (4)$$

Нами параметр h был выбран равным 96, так как это количество показаний, снимаемых за сутки для объекта, по которому нами использовались экспериментальные (контрольные) данные.

В предлагаемом алгоритме заполнения пропусков используются весовые коэффициенты для результатов, полученных этими двумя методами. Так как изначально выборка данных, равная одному году наблюдений, не содержала пропусков, то мы можем подсчитать ошибку работы каждого из двух методов:

$$E_m = \frac{1}{n} \sum_{i=1}^n e_i, \quad (5)$$

где e_i – погрешность расчета для одной точки конкретной моделью (методом).

В зависимости от этих ошибок формируются веса двух методов. Отношение весов принимается обратно пропорциональным отношениям даваемых ими погрешностей, а сумма весов составляет единицу:

$$\frac{E_{kNN}}{E_{Sams}} = \frac{w_{Sams}}{w_{kNN}}. \quad (6)$$

Зная веса методов и применяя алгоритм уже к выборке с пропусками, мы можем обоснованно задавать относительные вклады первого и второго методов при расчёте значения пропуска:

$$y_o(t) = y_1 * w_{kNN} + y_2 * w_{Sams}. \quad (7)$$

3. Результаты вычислительных экспериментов. Для проверки работы предложенного нами алгоритма использовались данные из системы EcoScada [3]. Были разработаны несколько тестовых примеров, в которые мы вручную внесли пропуски. В каждый из четырех сезонов года по очереди был внесен пропуск, содержащий определенное количество последовательных точек по времени. Таким образом, в году имелся единичный пропуск, менялась лишь его продолжительность по времени. Всего было использовано 7 тестовых примеров, содержащих соответственно 1, 2, 4, 8, 96, 192 и 384 пропущенных последовательных значения. Так как данные снимались каждые 15 минут, то мы имеем максимум четверо суток пропущенных данных в выборке длиной в 12 месяцев. На практике критическим считается отсутствие данных в течение двух суток.

Для оценки точности работы метода использовалась средняя статистическая ошибка $sMAPE$, вычисляемая как:

$$sMAPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{|y_i - f_i|}{|y_i + f_i|} \right) * 100 \%, \quad (8)$$

где y_i – реальное значение; f_i – восстановленное значение; n – число восстановленных значений.

Были проведены расчеты значений для всех пропусков. Полученные ошибки, вычисленные согласно (8), показаны в таблице и на рисунке.

Таблица

Результаты работы алгоритма восстановления							
Сезон/кол-во пропусков	1	2	4	8	96	192	384
Весна	0,017	0,75	0,96	1,07	14,8	9,6	12
Лето	20,4	19,8	18,3	10,6	14,9	8,3	11
Осень	5,7	3,2	2,4	1,6	9,7	6,2	6,9
Зима	2,3	1,3	1,9	1,7	7,9	5,7	6,8

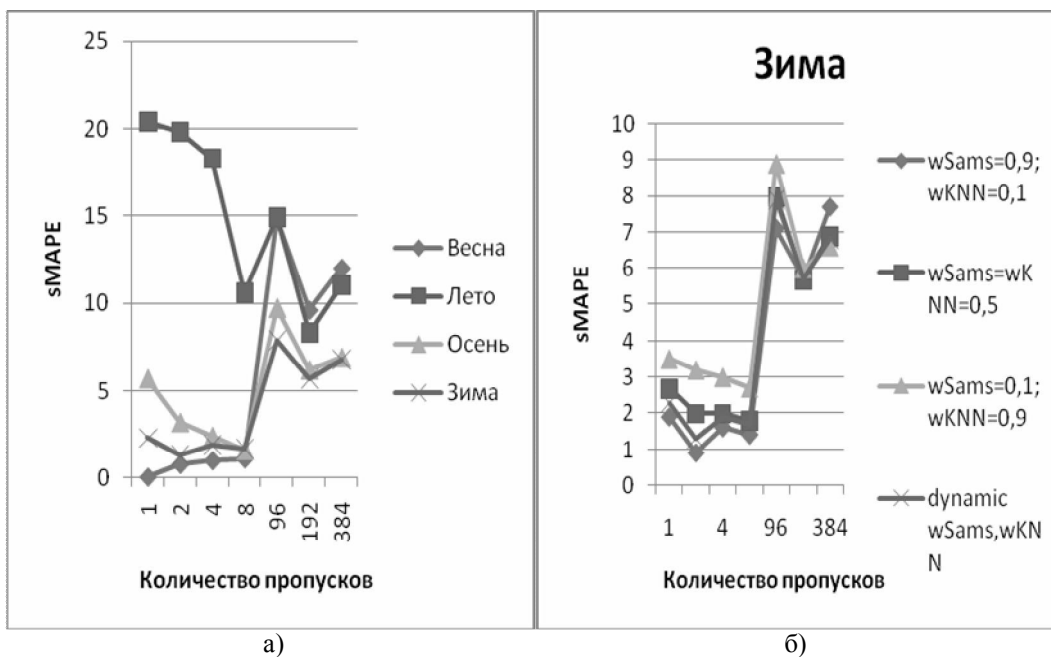


Рис. Ошибки восстановления пропусков в данных для различных выборок.
 Результаты работы алгоритма: а) относительно каждого из сезонов;
 б) в зимний период для различных пар весовых коэффициентов

4. Обсуждение результатов вычислительных экспериментов. В работе предложен подход к решению задачи заполнения пропусков в данных во временных рядах в автоматическом режиме.

Подход основан на комбинировании результатов двух методов с учетом сезона и продолжительность пропуска в данных по времени. В качестве базовых методов используются модель «ближайшего соседа» и простая средняя модель. Показана эффективность предлагаемого подхода на различных примерах (ошибка до 6,8 % на 384 пропусках). результатов вычислительных экспериментов видно, во-первых, что резкое увеличение ошибки происходит при возрастании продолжительности пропуска в данных до суток – это обусловлено тем, что показания в рабочее и нерабочее время сильно отличаются. Во-вторых, результаты сильно зависят от сезона, так как, например, потребление электроэнергии зимой и летом сильно отличаются. Можно сделать вывод, что при выборе в алгоритме весов для каждого из методов-компонент необходимо учитывать текущий сезон. Например, для зимы на малых группах пропусков модель (метод) Sams должна иметь больший вес, а для больших групп – больший вес должен иметь метод kNN.

Список литературы

1. Камаев В. А. Интеллектуальные системы автоматизации управления энергосбережением / В. А. Камаев, М. В. Щербаков, А. Бребельс // Открытое образование. – 2011. – № 2. – С. 227–231
2. Камаев В. А. Применение коннективистских систем для прогнозирования потребления электроэнергии в торговых центрах / В. А. Камаев, М. В. Щербаков, Д. П. Панченко, Н. Л. Щербакова, А. Бребельс // Управление большими системами. – Москва : ИПУ РАН, 2010. – Вып. 31. – С. 92–109.
3. Система сбора информации о потреблении электроэнергии EcoSCADA. – Режим доступа: <http://ecoscada.com> (дата обращения: 16.10.2012), свободный. – Заглавие с экрана. – Яз. рус.
4. Щербаков М. В. Методика выявления потенциала энергосбережения на основе интеллектуального анализа данных / М. В. Щербаков, Т. А. Яновский, А. Бребельс, Н. Л. Щербакова // Прикаспийский журнал: управление и высокие технологии. – 2011. – № 2. – С. 51–55.
5. Abdella M. I. The use of genetic algorithms and neural networks to approximate missing data in database / M. I. Abdella, T. Marwala // Computing and Informatics. – 2006. – Vol. 24. – P. 1001–1013,
6. Mohamed S. Neural network based techniques for estimating missing data in databases / S. Mohamed, T. Marwala // The 16th Annual Symposium of the Pattern Recognition Association of South Africa. – Langebaan, 2005. – P. 27–32,
7. Qiao W. Continuous on-line identification of nonlinear plants in powersystems with missing sensor measurements / W. Qiao, Z. Gao, R. G. Harley // Proc. of Intel. Joint Conf. on Neural Networks. – 2005. – P. 1729–1734.
8. Thompson B. B. On the contractive nature of autoencoders: Application to sensor restoration / B. B. Thompson, R. J. Marks, M. A. El-Sharkawi // Proc. of the International Joint Conference on Neural Networks. – 2003. – Vol. 4. – P. 3011–3016.
9. Troyanskaya O. Missing value estimation methods for DNA microarrays / O. Troyanskaya, M. Cantor, G. Sherlock et al. // Bioinformatics. – 2001. – vol. 17. – P. 520–525.

References

1. Kamaev V. A., Shcherbakov M. V., Brebels A. Intellektualnye sistemy avtomatizatsii upravleniya energosberezheniem [Intelligent power saving control automation systems]. *Otkrytoe obrazovanie* [Open Education] 2011, no. 2, pp. 227–231.
2. Kamaev V. A., Shcherbakov M. V., Panchenko D. P., Shcherbakova N. L., Brebels A. Primenenie konnektivistskikh sistem dlya prognozirovaniya potrebleniya elektroenergii v torgovykh tsentrakh [Application connective systems for forecasting energy consumption in shopping centers]. *Upravlenie bolshimi sistemami* [Large System Control]. Moscow, Institute of Control Sciences of Russian Academy of Sciences, 2010, issue 31, pp. 92–109.
3. Energy consumption information collection system Ecoscada. Available at: <http://ecoscada.com> (accessed 16 October 2012). (In Russ.)

4. Shcherbakov M. V., Yanovskiy T. A., Brebels A., Shcherbakova N. L. Metodika vyyavleniya potentsiala energosberezheniya na osnove intellektualnogo analiza dannykh [Method of identifying power saving potential on the basis of intelligent data mining]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2011, no. 2, pp. 51–55.
5. Abdella M. I., Marwala T. The use of genetic algorithms and neural networks to approximate missing data in database. *Computing and Informatics*, 2006, vol. 24, pp. 1001–1013.
6. Mohamed S., Marwala T. Neural network based techniques for estimating missing data in databases. *The 16th Annual Symposium of the Pattern Recognition Association of South Africa*. Langebaan, 2005, pp. 27–32.
7. Qiao W., Gao Z., Harley R. G. Continuous on-line identification of nonlinear plants in powersystems with missing sensor measurements. *Proc. of Intel. Joint Conf. on Neural Networks*, 2005, pp. 1729–1734.
8. Thompson B. B., Marks R. J., El-Sharkawi M. A. On the contractive nature of autoencoders: Application to sensor restoration. *Proc. of the International Joint Conference on Neural Networks*, 2003, vol. 4, pp. 3011–3016.
9. Troyanskaya O., Cantor M., Sherlock G. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 2001, vol. 17, pp. 520–525.

УДК 004.8

АЛГОРИТМ ПОВЕДЕНИЯ ГРУПП АВТОНОМНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ АГЕНТОВ ДЛЯ БИМЕДИЦИНСКИХ СИСТЕМ НА ОСНОВЕ ТЕОРИИ РОЕВОГО ИНТЕЛЛЕКТА

Истомин Виктор Владимирович, аспирант, Пензенский государственный технологический университет, 440039, Российская Федерация, г. Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11, e-mail: v.ist@mail.ru

Данная работа является одним из этапов решения задачи синтеза информационной модели поведения групп автономных интеллектуальных агентов для биомедицинских систем. Предварительно была произведена постановка задачи и ее обоснование, затем был разработан математический аппарат, удовлетворяющий требованиям моделирования поведения групп автономных интеллектуальных агентов.

На данном этапе работы был создан алгоритм поведения, описывающий последовательные целенаправленные действия автономных интеллектуальных агентов в ходе выполнения поставленной перед ними задачи. Важной особенностью разработанного алгоритма является то, что он учитывает специфику биомедицинских систем.

В дальнейшем предполагается провести экспериментальное исследование алгоритма и оценить адекватность прогнозирования на его основе поведения групп агентов.

Ключевые слова: информационная модель, многоагентные системы, интеллектуальные агенты, моделирование поведения, роевой интеллект, групповая робототехника, принятие решений

ALGORITHM OF BEHAVIOR OF GROUPS OF AUTONOMOUS INTELLIGENT AGENTS FOR BIOMEDICAL SYSTEMS BASED ON THE THEORY OF SWARM INTELLIGENCE

Istomin Viktor V., post-graduate student, Penza State Technological University, 1a/11 pr. Baydukova/Gagarin St., Penza, 440039, Russian Federation, e-mail: v.ist@mail.ru

This work is one of the stages of solution of the problem of synthesis of the information model of behavior of groups of autonomous intelligent agents for biomedical systems. Before was made a formulation