УДК 683.03

Т.В. Афанасьева

РЕШЕНИЕ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ВРЕМЕННЫХ РЯДОВ В РАМКАХ СТРУКТУРНО-ЛИНГВИСТИЧЕСКОГО ПОДХОДА[1]

***Афанасьева Татьяна Васильевна**, кандидат технических наук, окончила радиотехнический факультет Ульяновского политехнического института. Докторант, доцент кафедры «Прикладная математика и информатика» УлГТУ. Имеет статьи, монографию в области интеллектуального анализа временных рядов. [e-mail: tv.afanaseva@mail.ru].*

Аннотация

В статье приводится описание нового структурно-лингвистического подхода, предназначенного для реализации интеллектуального анализа временных рядов (ВР). Ориентация данного подхода на анализ временных рядов, хранимых в базах данных (БД), возможность использования для анализа временных рядов различной длины и извлечение знаний о поведении временных рядов в форме нечетких элементарных тенденций, а также представление результатов в лингвистической форме на ограниченном естественном языке позволяют расширить круг потенциальных пользователей систем, реализованных на принципах структурно-лингвистического подхода.

Ключевые слова: интеллектуальный анализ, временной ряд, нечеткая тенденция, извлечение знаний, структурно-лингвистический подход, моделирование.

Abstract

The article gives a description of a new structural and linguistic approach intended for the implementation of time-series data-mining. The orientation of this approach to the analysis of time series of different length and extraction of knowledge on behaviour of time series in the form of fuzzy elementary tendencies as well as the presentation of the results in the linguistic form in bounded natural language, allows extending the group of potential users of systems implemented using the principles of structural and linguistic approach.

Key words: data mining, time series, fuzzy tendency, knowledge extraction, structural and linguis-tic approach, modeling.

Введение

В связи с ростом хранимых, упорядоченных во времени данных о характеристиках объектов, процессов и систем в промышленности, экономике, медицине, образовании, социологии расширяются возможности анализа и прогнозирования количественных и качественных изменений этих характеристик, а также их носителей. Систематическое и комплексное исследование тенденций развития процессов на основе анализа временных рядов, извлеченных из хранилищ и баз данных, становится доступным специалистам различного профиля: проектировщикам, менеджерам среднего звена, аудиторам, специалистам в области контроля качества, экономистам, руководителям, маркетологам, аналитикам и др. Специфика временных рядов баз данных выражается в возможности наличия пропущенных значений и значений лингвистического типа, в вариабельности длины временных рядов от коротких до длинных, наличии нелинейностей, нечеткостей и нестационарностей.

Технологии баз данных расширяют круг задач, пользователей и создают новые возможности извлечения знаний из ВР баз данных при решении экспертных задач анализа процессов по:

1. Качественной оценке текущего и будущего состояний исследуемого процесса.
2. Обнаружению типичных и аномальных типов событий.
3. Выявлению существующих качественных изменений и их прогнозу.

При этом обнаружение тенденций, их качественная оценка и прогноз на основе временных рядов, извлеченных из баз данных предприятий, выступают как отдельная задача анализа, которая приобретает особую актуальность в связи со стремительным ростом и изменением хранимых данных.

Анализ временных рядов представляет собой самостоятельную, обширную и одну из наиболее интенсивно развивающихся областей исследования прикладной математики.

Решение обозначенных задач в рамках теории вероятности и математической статистики

1 Работа выполнена при финансовой поддержке ФЦП, проект № 02.740.11.5021.

сопряжено с определенными трудностями и ограничениями в силу специфики ВР БД, а также с высокими требованиями к квалификации пользователя и отсутствием методов и средств автоматической лингвистической интерпретации результатов и моделей.

Одним из новых направлений, обеспечивающих «интеллектуальную» поддержку специалистов по решению новых задач анализа ВР баз данных, является интеллектуальный анализ данных или Data Mining, в котором анализ поведения и тенденций развития процессов может быть рассмотрен как интеллектуальный анализ временных рядов или *Time Series Data Mining* (TSDM). Основными целями *Time Series Data Mining* являются, во-первых, анализ и моделирование процессов, характеризующихся высокой степенью неопределенности, в том числе «нестохастического» типа; во-вторых, повышение уровня интеллектуальной поддержки современных специалистов и, в-третьих, выявление скрытых закономерностей и извлечение новых знаний из временных рядов в лингвистической форме.

В основе новых методов *Time Series Data Mining* лежит нечеткая модель временного ряда, получившая название нечеткого временного ряда (НВР), построенная с привлечением нечетких экспертных оценок и нечетких систем. В отличие от числового временного ряда уровни нечеткого ВР образованы нечеткими значениями, сопоставимыми с экспертными оценками, в которых содержатся предметно-зависимые знания. Такая модель, принципиально являясь более грубой, тем не менее позволяет использовать дополнительные предметно-зависимые знания и моделировать поведение временного ряда в виде качественных оценок изменений и нечетких тенденций. В этом смысле также один и тот же временной ряд в различных предметных областях будет иметь разные нечеткие модели. При моделировании нечетких временных рядов необходимо определить его носитель, объект исследования и решаемые задачи. Носителем нечеткого временного ряда выступает исходный временной ряд, а объектом исследования — модель нечеткого временного ряда. Совокупность задач Data Mining применительно к нечетким временным рядам включает: сегментацию, кластеризацию, классификацию, индексирование, резюмирование, обнаружение аномалий, частотный анализ, прогнозирование, извлечение ассоциативных правил.

Данное новое направление в основном развивается в научных трудах иностранных ученых: К. Сонга, К. Хироты, В. Новака, В. Педрича, И. Перфильевой. Среди отечественных ученых данной тематике в области нечетких моделей временных рядов посвящены исследования И. Батыршина, С. Ковалева, Н. Ярушкиной.

Несомненными достоинствами нечетких моделей НВР являются: извлечение нелинейных зависимостей; моделирование коротких ВР и ВР, для которых затруднительно или невозможно построить адекватные стохастические модели, а также легко интерпретируемые результаты в виде нечетких правил; высокая степень автоматизации построения моделей и отсутствие высоких квалификационных требований к пользователям в области анализа ВР.

В целом интеллектуальный анализ нечетких ВР или Fuzzy *Time Series Data Mining* (FTSDM) — новое востребованное направление для решения задач анализа динамики развития процессов с учетом нечеткости в данных находится на этапе становления. Для него характерно отсутствие целостного, научно обоснованного подхода, методологических и теоретических положений построения моделей и методов анализа поведения ВР с учетом неопределенностей типа «нечеткость» в данных.

Методологические принципы структурно-лингвистического подхода в анализе временных рядов

Анализ методологических подходов (статистического, нейросетевого и нечеткого) к решению задачи моделирования временных рядов баз данных позволяет обобщить основные принципы, лежащие в их основе.

1. Принцип «разделяй и властвуй».

Это теоретический принцип декомпозиции общей модели временного ряда на модели, выражающие предопределенные классы поведения временного ряда. Другими словами, это экспертная декомпозиция исходного временного ряда на совокупность однородных по типу и одновременных временных рядов, поведение каждого из которых может быть описано отдельной моделью.

2. Принцип многомодельности.

Принцип является системным по своему содержанию, он основан на многомодельности и отношениях между моделями, описывающими функционирование отдельных компонент ВР. Каждая компонента может быть реализована отдельно взятыми независимыми модулями, реализующими тот или иной метод ее моделирования. Данный принцип не распространяется на многомодельное представление данных.

3. Принцип неопределенности и неточности.

Этот принцип является следствием нескольких причин: 1) использование приближенных моделей и численных алгоритмов; 2) неопределенности в исходных данных стохастического типа, продуцируемые неучтенными воздействиями внешней среды, и нечеткого типа, обусловленные нечеткостью восприятия (оценки) исходных,

промежуточных и выходных данных; 3) нечеткость оценивания моделей, порожденная экспертной деятельностью при их проектировании.

4. Принцип адаптации (обучения).

Применительно к системам моделирования временных рядов обучение может рассматриваться как важный процесс, значительно влияющий на качество создаваемых моделей. Существуют два аспекта в проблеме обучения: обучение эксперта построению моделей временных рядов и обучение модели. Системы моделирования создаются как закрытые архитектурные решения в виде лицензионных комплексов дорогостоящих программ, имеющих специфичные интерфейсы, закрытые форматы данных и критерии, нацеленные на моделирование либо отдельных компонент, либо всего ВР. Использование и настройка таких архитектур требуют высокой квалификации эксперта. Обучение моделей временных рядов рассматривается как их подгонка под новые данные путем перепостроения общей модели.

Вышеперечисленные принципы развиваются в новом структурно-лингвистическом подходе к анализу ВР БД в рамках направления FTSDM, в основе которого лежит модель нового объекта ВР — нечеткой элементарной тенденции [1].

Отличительной чертой данного подхода является тот факт, что результаты решения задач анализа ВР могут быть выражены не только в числовой форме, но и в лингвистической, выражающей тенденции развития в прошлом и будущем. Указанное свойство особенно важно, так как создает возможность представлять результаты в терминах онтологии предметной области, и актуально для задач поддержки проектных и управляющих решений в различных предметных областях, в которых человеческий фактор имеет определяющее значение.

В основе структурно-лингвистического подхода лежат следующие теоретические положения:

1. Понятия и модель нечеткого временного ряда и нечеткой тенденции, описывающей качественные изменения в нечетком временном ряду [1].

2. Формализм нечетких ACL-шкал, обеспечивающих автоматизацию качественного оценивания уровней и тенденций, а также построения и преобразования нечетких объектов временного ряда (нечетких уровней и нечетких тенденций) [2].

3. Методы и задачи направления интеллектуального анализа данных применительно к нечеткому временному ряду, позволяющие решать новые задачи [3].

4. Метод нечеткого моделирования и анализа нечетких тенденций [4].

5. Теория нечетких множеств и лингвистические информационные гранулы для достижения высокой интерпретируемости результатов анализа ВР [5].

Основная идея структурно-лингвистического подхода заключается в идентификации нечеткой грамматики языка *LANG*, задающей форму представления знаний о поведении нечеткого ВР в виде нечетких продукционных правил. При этом исследуемый временной ряд рассматривается как предложение на этом языке.

Универсальное представление модели временного ряда в форме аналитической зависимости — удобное средство, позволяющее на основе математически определенных и программно реализованных элементов языка формул (его синтаксиса и семантики) и математических методов обработки ВР определять числовые значения временного ряда в любой заданный момент времени.

В связи с этим сформулируем задачу анализа временного ряда как задачу анализа нечеткого временного ряда для идентификации модели поведения НВР, порождающую проблему определения языка представления структуры и параметров этой модели.

Так как уровни нечеткого временного ряда представлены значениями лингвистической переменной, то естественно представить структуру НВР как цепочку семантически определенных лингвистических элементов, терминальных символов V_T грамматики некоторого языка *LANG*. Отношения следования (предшествования) между этими термами, обнаруживаемые на исследуемом нечетком временном ряду, выражают синтаксические правила, правила подстановки *P*. Промежуточные структуры, составленные из последовательности термов, образуют нетерминальные символы V_N искомой грамматики, которые на основе свертки порождают аксиому грамматики S.

Параметрами рассматриваемой структурной модели могут выступать нечеткие ограничения, представляемые функциями принадлежности, а также и числовые значения.

Таким образом, при рассмотрении проблемы определения языка *LANG* для представления модели исследуемого НВР интерес представляет решение следующих задач:

1. Возможно ли свести задачу структурной идентификации НВР к задаче определения синтаксиса грамматики некоторого языка *LANG*?

2. Как определить семантические правила «вычисления» новых значений в языке *LANG*?

3. Какие информационные технологии и математические методы наиболее эффективны при реализации языка представления модели НВР?

4. Какие задачи анализа НВР могут быть определены на основе языка *LANG*?

Сформулированная выше совокупность задач позволяет обозначить структурно-лингвистический подход в решении задачи анализа ВР как задачи структурно-параметрической идентификации модели поведения НВР, задаваемой

грамматикой языка *LANG*, выраженной в форме лингвистических зависимостей, понятных прикладному пользователю.

Одной из первых задач при определении языка *LANG* и его грамматики является определение терминальных символов. Так как эта задача касается языка, предназначенного для описания модели поведения нечеткого временного ряда, то целесообразно в качестве терминальных символов языка *LANG* выбрать понятия, обозначающие нечеткие объекты НВР: нечеткие уровни, нечеткие тенденции, нечеткие временные интервалы.

Содержательно, нечеткие тенденции были введены в работе Н.Г. Ярушкиной [1].

Нечеткой тенденцией (НТ) нечеткого временного ряда будем называть нечеткую метку, выражающую характер изменения (систематическое движение) последовательности нечетких уровней НВР в заданном интервале времени. Нечеткая тенденция выражает поведение НВР в лингвистическом виде, например: «Рост», «Падение», «Стабилизация», «Колебания», «Хаос».

Изменения нечетких меток во временном пространстве порождают нечеткий временной ряд с нечеткой тенденцией.

Решение задач интеллектуального анализа временных рядов в рамках структурно-лингвистического подхода

Рассмотрим расширенную структурно-лингвистическую модель НВР при условии нечетких уровней и четких моментов времени [4]:

$$\tilde{Y} = \langle T, X, \tilde{X}, \mu_{\tilde{X}}(X), \tau \rangle,$$

где $\tau = \langle \tilde{\nu}, \mu, \tilde{\alpha}, \Delta t \rangle$,

T – атрибут времени, задаваемый упорядоченными по возрастанию моментами времени;

X – атрибут, хранящий уровни ВР;

$\tilde{X}$ – атрибут, который задает множество лингвистических термов, обозначающих нечеткие уровни ВР;

$\mu_{\tilde{X}}(X)$ – степень принадлежности X лингвистическому терму $\tilde{X}$;

τ – наименование нечеткой тенденции;

$\tilde{\nu}$ – тип нечеткой тенденции, определяемый на основе операции *TTend ACL-шкалы*. Последовательность типов нечетких тенденций моделирует структуру изменений НВР;

$\tilde{a}$ – интенсивность нечеткой тенденции, контекстное расширение тенденции, определяемое операцией RTend ACL-шкалы;.

μ – степень принадлежности нечеткой тенденции нечеткому временному ряду, при $\mu = 1$ нечеткая тенденция рассматривается как четкая тенденция;

Δt – длительность данного типа нечеткой тенденции.

Сформулируем основные положения структурно-лингвистического подхода:

1. Терминальным символам грамматики языка *LANG* сопоставим нечеткие уровни и элементарные нечеткие тенденции $v_T = \tilde{X} \cup \Im$ нечеткого ВР $\tilde{Y}$.

2. Лексика грамматики языка *LANG* определяется локальными нечеткими тенденциями $V_N = N\tau$ НВР. Аксиома грамматики есть общая тенденция $S = G\tau$.

3. Синтаксис грамматики языка *LANG* зададим в виде правил следования нечетких тенденций $P = Rule\_e\tau$, определяемых на НВР как результат зависимости нечетких тенденций от значений тенденций в предыдущие моменты времени: «ЕСЛИ тенденция(i), ТО затем тенденция (j)».

4. Семантике лингвистических выражений языка *LANG* сопоставим нечеткие множества $\tilde{X}, \tilde{V}, \tilde{A}$, задаваемые функциями принадлежности в ACL-шкале.

5. Решение задач интеллектуального анализа временных рядов (TSDM), определенных как задачи извлечения знаний из временных рядов, сведем к решению задач анализа и синтеза грамматики языка *LANG*:

- Сегментация – представление НВР в виде последовательности терминальных символов (нечетких тенденций) грамматики языка *LANG*.
- Кластеризация – построение лексики языка LANG: образование локальных нечетких тенденций.
- Поиск ассоциативных правил – определение синтаксиса языка LANG на основе извлечения синтаксических правил следования нечетких тенденций.
- Классификация – распознавание общей нечеткой тенденции НВР как вывод нечеткой аксиомы грамматики языка *LANG*.

Используем сгенерированную грамматику в режиме порождения для построения гранулярного представления НВР и решения следующих задач извлечения знаний из временных рядов:

- Частотный анализ – формирование компонент гранулярного описания НВР в виде часто встречающихся нечетких тенденций и их паттернов.
- Прогнозирование – получение многоуровневого прогноза: в терминах общей, локальной, элементарной тенденций, в терминах нечетких и числовых уровней ВР.
- Резюмирование – генерация гранулярного описания поведения НВР, определение семантики гранул и трансляция гранулярного описания НВР в виде предложения на естественном языке.

• Поиск аномалий — выявление лексических, синтаксических и семантических ошибок и нетипичных нечетких объектов и правил на основе грамматики языка *LANG*.

Исследование результативности интеллектуального анализа ВР в рамках структурно-лингвистического подхода

Для исследования результативности предложенного структурно-лингвистического подхода к решению комплекса задач анализа ВР разработана программная система *FuzzyTend* для нечеткого моделирования и анализа нечетких тенденций временных рядов [4].

Отличительными особенностями программной системы *FuzzyTend* являются:

1. Решение новых задач анализа нового объекта ВР — нечеткой тенденции. Выявление новых закономерностей поведения ВР в форме нечетких правил зависимостей между элементарными нечеткими тенденциями. Моделирование параметров элементарных нечетких тенденций.

2. Автоматизация построения моделей при высокой интерпретируемости результатов и отсутствии требований к математической подготовке конечных пользователей в области анализа ВР.

3. Решение комплекса задач Time Series Data Mining для построения нечетких моделей ВР.

4. Реализация интегрального метода нечеткого моделирования и анализа нечетких тенденций.

Результаты моделирования ВР в рамках структурно-лингвистического подхода сравнивались с результатами моделирования в нейросетевом и нечетком подходах. Исследование моделей ВР предложенного подхода показало их адекватность, полезность и конкурентоспособность для нестационарных ВР, в том числе малой длины.

Заключение

Предложенный новый структурно-лингвистический подход, развивающий методологию интеллектуального анализа ВР за счет учета и анализа неопределенностей типа «нечеткость» тенденции, основан на теоретических положениях моделирования нечетких временных рядов.

Данный подход расширяет сферу применимости систем моделирования ВР на данные, обладающие объективной нечеткостью, и нацелен, в первую очередь, не на моделирование состояний некоторого процесса, а на распознавание различных тенденций, происходящих в этом процессе. В то же время использование, автоматическая обработка и вывод формализованных нечетких лингвистических высказываний обеспечат снижение времени на разведывательный анализ данных, который значительно превосходит время, затрачиваемое на моделирование ВР, и вносит весомую долю в неточность результатов при принятии решений.

Отметим новые свойства интеллектуальных систем структурно-лингвистического моделирования ВР: снижение требований к квалификации конечных пользователей, высокая степень интерпретируемости, понятности для конечного пользователя. Эти свойства позволяют надеяться, что структурно-лингвистическое моделирование тенденций ВР найдет применение как ориентированное на массового пользователя инструментальное средство для анализа тенденций, для которых допустимы интервальные числовые оценки уровней ВР.

СПИСОК ЛИТЕРАТУРЫ

1. Ярушкина Н. Г. Современный интеллектуальный анализ нечетких временных рядов / Н. Г. Ярушкина // Труды V-й Международной научно-практической конференции «Интегрированные модели и мягкие вычисления» (Коломна, 20-30 мая 2009 г.). В 2 т. Т. 1. — М. : Физматлит, 2009. — С. 19—30.

2. Афанасьева Т. В. Модель ACL-шкалы для генерации лингвистических оценок в принятии решений / Т. В. Афанасьева // Вопросы современной науки и практики. Университет им. В. И. Вернадского. — 2008.— № 4 (2). — С. 91—96

3. Батыршин И. З. Модели и методы перцептивного дата майнинга временных рядов для систем поддержки принятия решений / И. З. Батыршин, Л. Б. Шереметов // Нечеткие системы и мягкие вычисления. — Тверь, 2007. — Т. 2, № 1.

4. Афанасьева Т. В. Нечеткое моделирование временных рядов и анализ нечетких тенденций / Т. В. Афанасьева, Н. Г. Ярушкина. — Ульяновск : УлГТУ, 2009.

5. Zadeh Lotfi A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. / Lotfi A. Zadeh // Fuzzy Sets and Systems. — 1997. — Vol. 90. — pp. 111—127.