

УДК 004.896

НЕЙРОСЕТЕВАЯ МОДЕЛЬ АНАЛИЗА ТЕХНОЛОГИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ В РАМКАХ МЕТОДОЛОГИИ DATA MINING

А. М. Вульфин,

аспирант

А. И. Фрид,

доктор техн. наук, профессор

Уфимский государственный авиационный технический университет

Решается задача выявления и распознавания технологических событий и состояний узла инженерной сети на основе анализа временных рядов, характеризующих протекающие на объекте процессы. Поэтапно решаются задачи предобработки исходных данных, сегментации временных рядов, кластеризации и классификации имеющихся данных и построения нейросетевой модели для анализа состояния объекта в рамках методологии Data Mining.

Ключевые слова — адаптивная сегментация, распределенная нейронная сеть с временной задержкой, метод обобщенного отношения правдоподобия, временной ряд.

Введение

В данной работе технологический сигнал рассматривается как последовательность следующих друг за другом участков, обладающих на некотором временном интервале постоянными свойствами (структурная модель сигнала) [1]. Технологический временной ряд представляет собой совокупность значений какого-либо параметра исследуемого технологического процесса, последовательно измеренных через некоторые (равные) промежутки времени. Следовательно, задача обработки сигнала заключается в выделении отдельных сегментов временных рядов с последующей их кластеризацией и сопоставлении полученных кластеров с имеющимися классами событий.

Подобное построение модели сигнала является обобщением задачи выделения особенностей [1]. Выделение в технологических сигналах особенностей, повторяющихся процессов, скрытых аномалий и закономерностей в автоматическом режиме может быть выполнено с помощью технологий интеллектуального анализа данных (Data Mining) [2]. Применение таких технологий лежит в основе диагностики состояния узла инженерной сети.

Целью работы является повышение эффективности интеллектуальной автоматизированной си-

стемы распознавания ситуаций, возникающих на объектах инженерной сети, в рамках методологии Data Mining и на основе нейронных сетей. Для ее достижения поставлены три задачи, решение которых представлено в данной статье.

Анализ технологических процессов в узлах инженерной сети с помощью методов интеллектуального анализа данных

Основной проблемой, возникающей при анализе динамических систем и недетерминированных сигналов различной природы, является определение их временных и частотных характеристик, поскольку любой недетерминированный сигнал и динамическая система, как правило, являются нестационарными объектами. Большинство технологических сигналов, описывающих динамические системы, являются нестационарными [3–7].

Исследование сигналов методами Data Mining является процессом, состоящим из нескольких этапов [2, 3].

Этап I. Сбор и хранение исходных данных. Этот этап включает два синхронных процесса:

- 1) регистрацию сигналов и подготовку их цифровых записей;
- 2) регистрацию событий и их описание.

Этап II. Применение методов Data Mining. Данный этап является ключевым и состоит из нескольких процессов.

1. Использование методов выделения особенностей сигналов на основании применения методов сегментации.

2. Унификация выделенных сегментов.

3. Классификация событий и кластеризация сегментов. После получения подготовленного материала необходимо его систематизировать с выделением классов, которые соответствуют основным группам:

— сервисные события, связанные с регламентированным воздействием на систему регистрации (пропуски данных);

— внешние воздействия на дочерних узлах инженерной сети;

— внутренние события на узле инженерной сети.

Поскольку для сегментов не существует априори заданных классов, то к ним требуется применить методику кластеризации.

4. Сопоставление событий кластерам сегментов. Сопоставление производят в хронологическом порядке, основываясь на гипотезе, что в один момент времени происходит одно событие [5, 6].

5. Построение модели, способной в реальном времени обрабатывать данные технологического временного ряда и классифицировать события и состояния узла инженерной сети.

6. Тестирование и оценка работоспособности модели выявления и распознавания технологических ситуаций.

Описанная концепция Data Mining показывает, что определенную сложность представляет этап сегментации исходного сигнала. Согласно работе [1], задача сегментации сводится к построению детектора, который по данной реализации сигнала $f(t)$ устанавливает истинность одной из вероятностных гипотез.

Практически всегда можно выделить некоторый временной интервал Δt , на котором параметры объекта изменяются несущественно, т. е. в пределах некоторого значения с пороговым отклонением ε . Такой интервал называют интервалом квазистационарности, а параметры объекта на этом интервале считаются постоянными. Таким образом, задача сегментации может быть сформулирована как задача поиска границ сегментов или задача поиска моментов изменения свойств сигнала [1].

Рассмотрим временной ряд $\{x_t\}$. Требуется построить детектор, выбирающий одну из двух гипотез H_0 и H_1 . Гипотеза H_0 предполагает, что исследуемый сигнал $\{x_t\}$ соответствует модели M_1 . Гипотеза H_1 предполагает, что существует момент времени τ , в который исследуемый сигнал

$\{x_t\}$ соответствует модели M_1 при $t < \tau$ и модели M_2 при $t \geq \tau$, где τ — граница сегмента:

$$[x_t | t < \tau], H_0: f(x) \leftarrow M_1;$$

$$[x_t | t \geq \tau], H_1: f(x) \leftarrow M_2.$$

Из сформулированной задачи следует, что основой разрабатываемых алгоритмов сегментации является гипотеза о виде модели M , описывающей исследуемый сигнал. Выбор модели определяется прикладной задачей, в данной работе рассматриваются линейные авторегрессионные модели (АР-модели) и нелинейные нейросетевые модели.

По характеру смены одного квазистационарного участка другим временные ряды, характеризующие рассматриваемые технологические процессы, можно разделить на две группы:

— временные ряды с переключающейся динамикой;

— временные ряды с дрейфовой динамикой.

Для анализа таких процессов придется разрабатывать специальные методы [3–7].

Разработка нейросетевой структуры адаптивной сегментации и обработки временных рядов

В работах [3–5] описано моделирование различных типов нестационарного поведения сигналов и систем, а также их недостатки, для преодоления которых предложено использовать [3]:

— распределенную нейронную сеть (НС) прямого распространения с задержкой по времени (distributed Time lagged Feedforward Network — dTLFN) для построения модели квазистационарного сегмента;

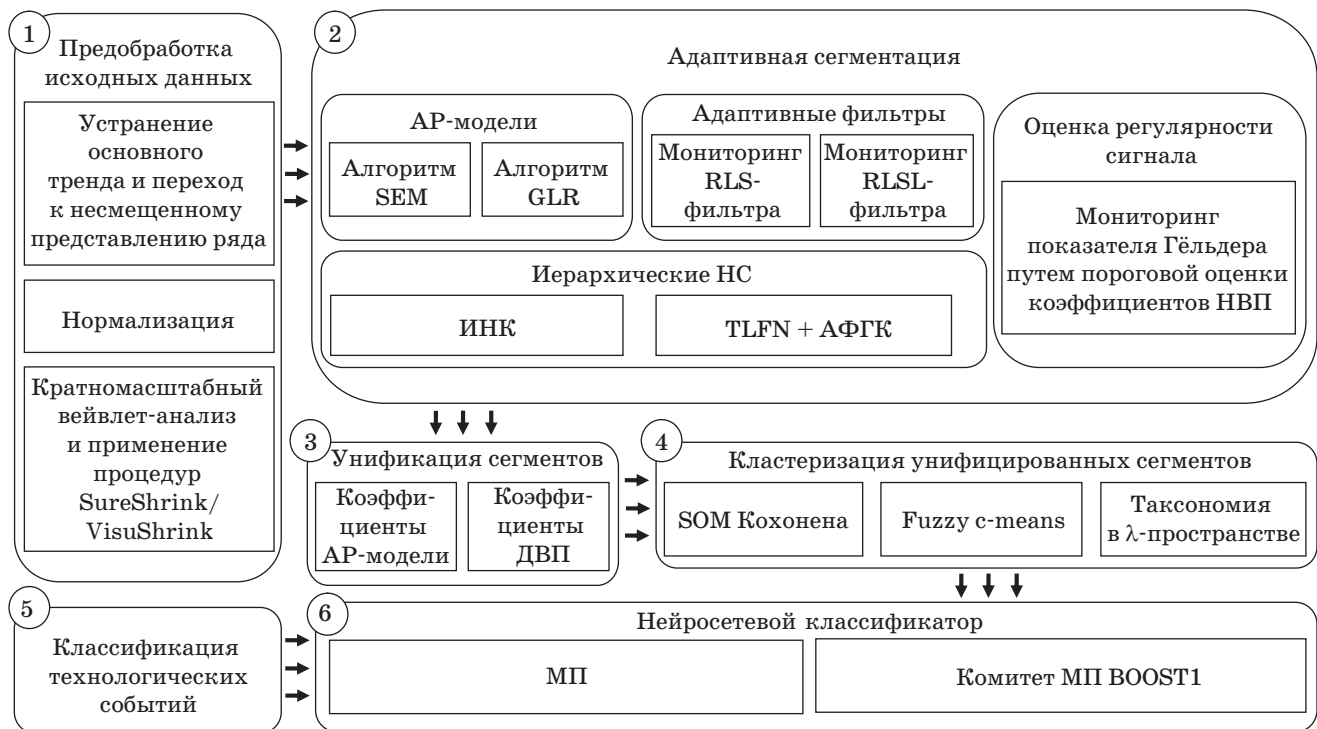
— метод обобщенного отношения правдоподобия (ООП) для адаптивной пороговой сегментации;

— модифицированный алгоритм формирования групп классов (АФГК) для кластеризации и классификации сегментов [7].

Это позволяет объединить несколько этапов анализа в рамках методологии Data Mining.

Построить модель технологических сигналов, руководствуясь технологией Data Mining, можно различными методами — как с применением предобработки исходных данных, так и с помощью разнообразных алгоритмов кластеризации на дальнейших этапах моделирования. Различные подходы к построению нейросетевой модели [3–5] представлены на рис. 1, где приняты следующие обозначения:

РНК (RLS) — адаптивный фильтр, построенный по рекурсивному методу наименьших квадратов (Recursive Least Squares); МСО (SEM) — мера спектральной ошибки (Spectral Error Mea-



■ Рис. 1. Методы построения нейросетевых систем анализа временных рядов

sure); ООП (GLR) — обобщенное отношение правдоподобия (Generalized Likelihood Ratio); ПРНК (RLSE) — адаптивный фильтр, использующий алгоритм решетчатого рекурсивного метода наименьших квадратов с задержками (Recursive Least-Squares Lattice); НВП — непрерывное вейвлет-преобразование; ИНК — иерархический нейросетевой классификатор; МП — многослойный перцептрон; SOM Кохонена — самоорганизующиеся карты (self-organizing map) Кохонена; Fuzzy c-means — алгоритм нечеткой кластеризации с-средних; BOOST1 — алгоритм обучения комитета НС.

Исходный временной ряд приводится к нулевому среднему и единичному среднеквадратическому отклонению. В ходе подбора оптимальной схемы подготовки некоторых видов технологических данных для очистки сигнала от высокочастотных шумов были использованы процедуры предобработки данных с помощью пороговой фильтрации SureShrink/VisuShrink [4] коэффициентов дискретного вейвлет-преобразования (ДВП) в целях выявления значимых компонент. Данная процедура была использована совместно с некоторыми методами адаптивной сегментации за исключением иерархических нейросетевых структур и методов оценки регулярности сигнала.

В работе [4] описаны базовые алгоритмы адаптивной сегментации с помощью АР-моделей,

а в [5] — адаптированная реализация методов для решения некоторого набора задач. Используются методы оценки МСО и метод ООП. Показано, что методы, использующие АР-модели, реагируют на «длительные» изменения и менее чувствительны к «быстрым» изменениям временного ряда, описывающего технологический процесс. Авторы [5] предлагают применять гетерогенную нейросетевую структуру, использующую dTLFN для построения модели квазистационарного сегмента, и метод ООП для адаптивной пороговой сегментации. Это позволяет расширить возможности систем адаптивной сегментации и повысить их чувствительность к «быстрым» изменениям технологического сигнала. Алгоритм адаптивной сегментации нестационарных временных рядов с помощью dTLFN и расширенного метода ООП является развитием методов, предложенных в работе [7].

Следующим этапом обработки временного ряда является этап унификации разнородных по длине сегментов и последующая их кластеризация. В качестве векторов признаков, характеризующих выделенные сегменты, могут быть использованы [3]:

- коэффициенты АР-модели, соответствующей каждому сегменту;
- k наиболее значимых коэффициентов ДВП данного сегмента.

В работе авторов [3] используются три метода кластеризации совокупности унифицированных сегментов для сравнения эффективности по критериям производительности и характеристикам полученных кластеров.

По имеющейся легенде, описывающей совокупность технологических событий, происходящих на узле инженерной сети, формируют классы, сопоставляемые с теми или иными кластерами сегментов, основываясь на временной последовательности событий.

Заключительным этапом является создание и обучение комитета НС, способного обрабатывать данные технологического временного ряда и классифицировать события и состояния узла. В работе [3] описан комитет НС, каждая из которых представляет собой МП, обучаемый согласно алгоритму BOOST1, а также проведено сравнение с одиночным МП.

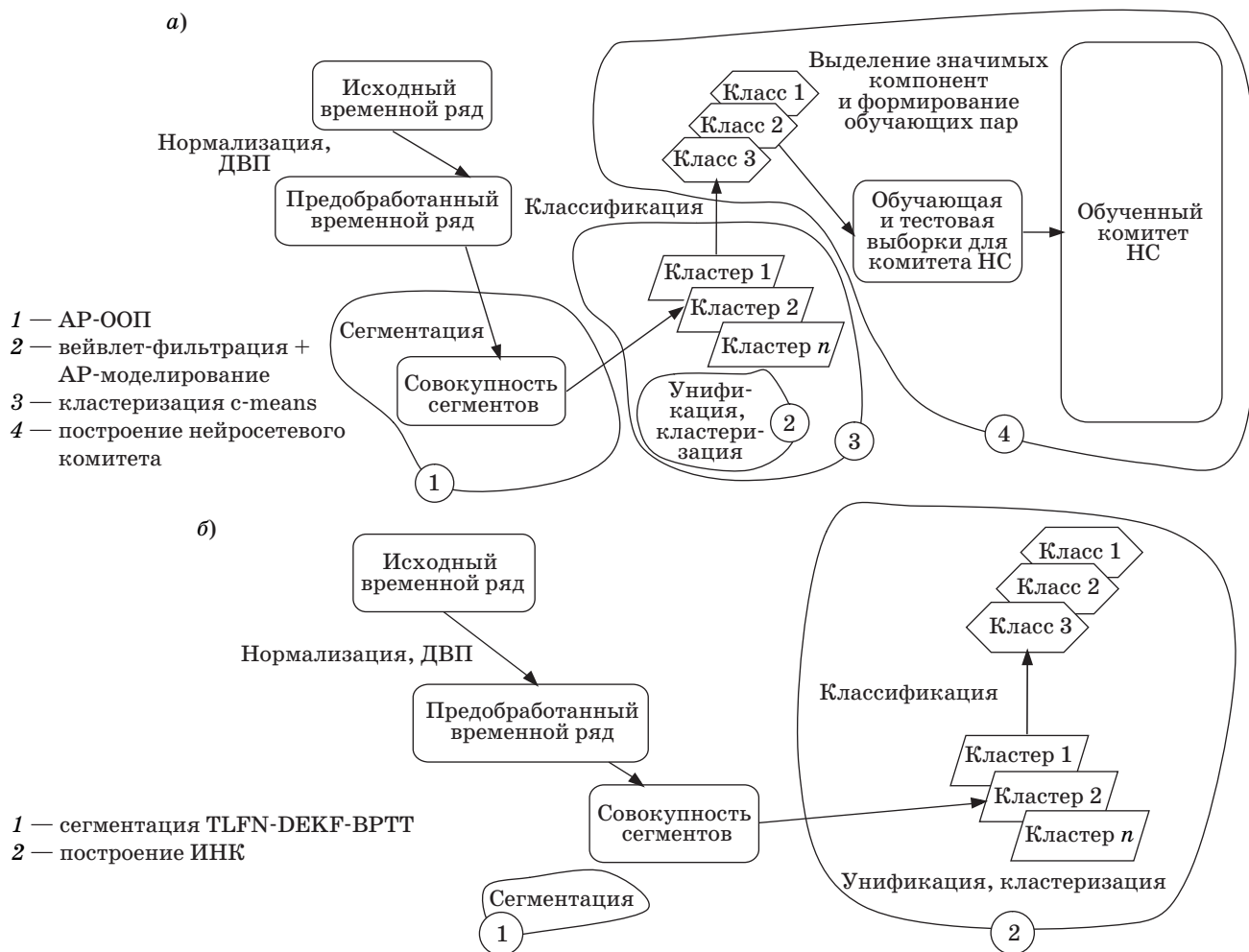
В работах [3–5] представлено несколько моделей анализа технологических временных рядов,

■ Таблица 1. Этапы и соответствующие алгоритмы анализа временного ряда

Номер метода	Используемый алгоритм			
	Предварительная сегментация	Унификация	Кластеризация	Классификация
1	AP-ООП	ДВП	Fuzzy c-means	Комитет МП BOOST1
2	ИНК			
3	dTLFN-ООП	АФГК		

комбинирующих алгоритмы различных этапов Data Mining. Этапы анализа временного ряда и соответствующие им алгоритмы приведены в табл. 1.

Для обучения dTLFN применены алгоритмы ВРТТ (Backpropagation Through Time, обратное распространение ошибки во времени) и DEKF



■ Рис. 2. Обобщенная схема работы системы выявления и идентификации событий с использованием AP-моделей (а) и dTLFN (б)

(Decoupled Extended Kalman Filter, несвязанный расширенный фильтр Калмана).

Представлена обобщенная схема (рис. 2, а) работы системы выявления и идентификации событий для первого метода (см. табл. 1).

Поскольку второй и третий методы анализа временных рядов, приведенные в табл. 1, включают помимо сегментации временного ряда этапы кластеризации-классификации согласно методологии Data Mining (рис. 2, б), то они не требуют использования дополнительных шагов и процедур, в отличие от AP-ООП, что позволяет сократить количество подстраиваемых параметров и упростить последующий анализ.

Анализ работоспособности нейросетевых систем обработки технологических временных рядов

Для оценки работоспособности гетерогенных нейросетевых систем обнаружения событий и описания текущего состояния технологического процесса рассмотрен ряд модельных задач.

В первой задаче моделирование нестационарного сигнала выполнено с помощью двухполюс-

■ **Таблица 2. Параметры методов адаптивной сегментации с помощью AP-моделей**

Параметры модели		Эксперимент 1
AP-модель	Порядок	8
Параметры алгоритма ООП	Длина опорного окна	20
	Длина тестового окна	10
Результаты сегментации	Количество выделенных/исходных сегментов	4/4
	Средневзвешенная оценка попадания	3,87
	Процент успешной классификации подвижным тестовым окном	94–98

ной AP-системы [3]. Параметры алгоритмов сегментации приведены в табл. 2, 3 (эксперимент 1) и 4 (эксперимент 1).

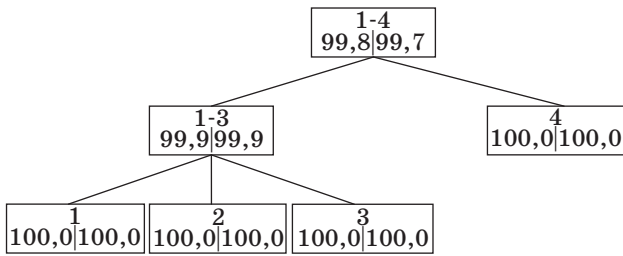
В результате работы алгоритма построения ИНК [7] сформирована иерархическая структура следующего вида (рис. 3). На каждом уровне показаны объединенные в классы сегменты и их изначальная принадлежность, процент распознавания на обучающей и тестовой выборках.

■ **Таблица 3. Параметры методов адаптивной сегментации с помощью ИНК**

Параметры модели		Номер эксперимента				
		1	2	3	4	5
Архитектура МП узла, параметры АФГК	Количество нейронов по слоям	5-2-64	5-2-40	5-20-48	5-20-128	5-20-256
	Изначальное количество сегментов	64	40	48	128	256
	Шаг анализа	5	5	200	200	200
	Скорость обучения	0,1	0,01	0,01	0,01	0,01
	Момент	0,5	0,5	0,5	0,5	0,5
Результаты сегментации	Количество выделенных/исходных сегментов	4/4	20/20	20/20	41/24	5-20-256
	Средневзвешенная оценка попадания	3,96	17,04	–	–	–
	Процент успешной классификации подвижным тестовым окном	99–100	96,1	98,0	72,1–75,0	–

■ **Таблица 4. Параметры методов адаптивной сегментации с помощью dTLFN(DEKF-BPTT)-ООП-АФГК**

Параметры модели		Номер эксперимента				
		1	2	3	4	5
Архитектура dTLFN	Количество нейронов по слоям	5-10-1	5-20-1	5-25-1	5-25-1	5-25-1
Архитектура МП узла, параметры АФГК	Количество нейронов по слоям	5-2-4	5-2-4	5-20-4	5-20-52	5-20-52
	Шаг анализа	2	2	50	50	50
	Скорость обучения	0,01	0,01	0,01	0,01	0,01
	Момент	0,5	0,5	0,5	0,5	0,5
	Результаты сегментации	Количество выделенных/исходных сегментов	4/4	20/20	20/20	35/24
Средневзвешенная оценка попадания		3,97	17,10	–	–	–
Процент успешной классификации подвижным тестовым окном		99-100	96,2	98,0	74,4–81,0	–



■ *Рис. 3. ИНК для сегментации псевдохаотического сигнала АР-системы*

В табл. 2–4 приводятся следующие данные:

- количество выделенных сегментов и количество исходных квазистационарных участков (доступно на модельных задачах);
- средневзвешенная оценка попадания — рассчитывается среднее значение отношения полученных и истинных длин сегментов;
- процент успешной классификации подвижным тестовым окном — на вход системы сегментации-кластеризации-классификации подается k последовательных отсчетов временного ряда. Принадлежность последовательности к определенному классу определяется по $(k + 1)$ -й точке. Приводится процент успешного распознавания типа текущей динамики участка временного ряда.

Как видно из табл. 2–4 (эксперимент 1), данный нестационарный ряд успешно сегментирован и выделенные сегменты распределены по классам в соответствии с типом динамики.

Следующий рассмотренный нестационарный сигнал имел переключающуюся динамику. Псевдохаотический временной ряд определяется последовательностью [7]

$$x(t + 1) = f_i(x(t)), \quad i = 1, 2, 3, 4;$$

$$f_1(x) = 4x(1 - x), \quad x \in [0, 1]; \text{ logistic map};$$

$$f_2(x) = 2x, \quad x \in (0, 0,5]; \quad 2(1 - x), \quad x \in [0,5; 1); \text{ tent map};$$

$$f_3(x) = f_1(f_1(x)); \text{ double logistic map};$$

$$f_4(x) = f_2(f_2(x)); \text{ double tent map}.$$

Эти последовательности, чередуясь, генерировали по $S_L = 100$ точек, процесс повторился 5 раз. Общая длина L псевдохаотического временного ряда составила 2000 точек. Далее ряд приводится к нулевому среднему и единичному среднеквадратичному отклонению.

Архитектура, параметры алгоритмов и результаты сегментации псевдохаотического временного ряда приведены в табл. 3 (эксперимент 2) и 4 (эксперимент 2).

Для сегментации этого временного ряда неприменим метод АР-ООП, так как ряд содержит квазистационарные сегменты, порожденные не-

линейными моделями. Результаты сегментации колебались в пределах 40–50 % успешно распознанных сегментов.

Далее был рассмотрен аналогичный сигнал с участками дрейфа между соседними сегментами. Описанные ранее псевдохаотические последовательности, чередуясь, генерировали по $S_L = 100$ точек с переходными участками шириной $D_L = 50$ точек, процесс повторился 5 раз [7], $L = 2400$ точек. Результаты адаптивной сегментации и параметры алгоритмов указаны в табл. 3 (эксперимент 3) и 4 (эксперимент 3).

Рассмотрим нестационарный временной ряд, порождаемый тремя источниками вида Makey–Glass. Исходные данные описываются следующим дифференциальным уравнением [5]:

$$\frac{dy}{dt} = -0,1y(t) + \frac{0,2y(t - t_d)}{1 + y(t - t_d)^{10}}.$$

Каждый из трех последовательно работающих источников использовал соответствующее значение параметра $t_d = 17, 23, 30$. Итоговая последовательность содержит квазистационарные сегменты длиной $S_L = 300$ точек с переходными участками длиной $D_L = 100$ точек, $L = 4800$ точек.

Данный временной ряд содержит участки с дрейфовой динамикой. Из табл. 3 (эксперимент 4) и 4 (эксперимент 4) видно, что применение dTLFN, по сравнению с исходным ИНК, оправдано. Удалось избежать излишней сегментации, тем самым повысив успешность классификации. Отметим полученный разброс результатов серии экспериментов, что связано с псевдохаотической природой порождающего соотношения.

Для анализа работоспособности систем идентификации технологических событий на узле инженерной сети были использованы данные об изменении параметров узла учета нефти СИКН (система измерения и контроля качества нефти) [3–5] инженерной сети за 9 дней. На узле СИКН каждые 5 с замеряется давление на выходе узла и расход нефти по 5 измерительным линиям. Для дальнейшего анализа использован временной ряд, описывающий давление на выходе узла. Общая длина ряда составляет 155 520 отсчетов.

Построение нейросетевой модели узла приема-сдачи нефти с помощью АР-моделей (см. табл. 1) и выявление технологических ситуаций описано в работе [3].

Архитектура нейросетевых систем, использованных для анализа технологического временного ряда, приведена в табл. 3 (эксперимент 5) и 4 (эксперимент 5).

Общие результаты работы систем с различной структурой приведены в табл. 5. Используются как гетерогенные нейросетевые модели на основе

■ Таблица 5. Результаты сегментации технологического временного ряда

Метод / Этап		Количество выделенных сегментов/ классов	Процент успешной классификации	
			подвижным окном известных событий	подвижным окном смеси известных и неизвестных событий
АР-ООП	Сегментация	548/19	78,7	61,9
wavelet-разложение/ АР-моделирование	Унификация сегментов			
Fuzzy c-means	Кластеризация			
Комитет МП BOOST1	Классификация			
ИНК	Сегментация	471/20	82	69
	Унификация сегментов			
	Кластеризация			
	Классификация			
dTLFN(DEKF- BPTT)	Сегментация	230/29	87	73
АФГК	Унификация сегментов			
	Кластеризация			
	Классификация			

АР-моделирования на этапе сегментации, так и более однородные сети ИНК и АФГК.

Нейросетевая модель на основе dTLFN (см. табл. 1, метод 3) функционирует в рамках методологии Data Mining, что позволяет осуществлять наиболее эффективный анализ технологических временных рядов. Модель объединяет сразу несколько этапов обработки данных в рамках единой нейросетевой парадигмы и использует адаптированные базовые алгоритмы обучения НС.

Анализ временных рядов с помощью однородных нейросетевых структур предпочтительнее, чем поэтапное построение многоуровневой системы обработки, что видно из сравнения результатов в табл. 5. К первой группе относятся методы, включающие различные алгоритмы поэтапной реализации схемы анализа данных, согласно концепции Data Mining. В ходе модельных и натурных экспериментов выбраны наиболее эффективные методы: АР-ООП для предварительной сегментации; коэффициенты АР-модели, описывающей каждый сегмент, для унификации; Fuzzy c-means для кластеризации унифицированных сегментов и комитет МП для сопоставления кластеров сегментов и классов событий. Недостатками метода являются сложность и неформализуемость подбора параметров каждого из ал-

горитмов, излишняя сегментация исходного сигнала, что ведет к частому переключению сигнализирующей о типе текущей динамики системы. Как ИНК, так и dTLFN позволяют сократить количество настраиваемых параметров и автоматизировать процесс построения конечной нейросетевой системы. Использование гибридной АФГК и сети dTLFN дает возможность избежать излишней сегментации и переключения сигнализирующей системы, а также расширить количество выделяемых классов событий путем обнаружения переходных состояний.

Заключение

В работе предложены некоторые подходы к анализу технологических временных рядов с помощью нейросетевых структур и в рамках методологии Data Mining.

На основе модельных испытаний и натурных данных разработана интеллектуальная модель распознавания технологических ситуаций, возникающих на объектах инженерной сети. Модель основана на нейросетевой обработке временных рядов и включает в себя нелинейный фильтр на базе распределенной сети dTLFN в совокупности с расширенным методом оценки ООП и моди-

фицированный ИНК. Применение разработанной модели позволило повысить достоверность описания технологических ситуаций.

Разработана структура адаптивной сегментации технологических временных рядов с помощью гомогенных и гетерогенных нейросетевых структур различной архитектуры, что расширяет возможность выбора наиболее подходящей по совокупности факторов сложности и производительности модели для анализа временных рядов различной природы.

Протестирована работоспособность предлагаемой модели на имеющихся исторических данных реальной задачи распознавания ситуаций на узле приема-сдачи нефти и на ряде модельных задач.

Показана эффективность данного подхода применительно к задачам разработки интеллектуальных автоматизированных систем, так как разработанная система позволяет:

— увеличить количество классов выявляемых событий на 30 % путем введения субклассов, описывающих участки смены типа динамических параметров квазистационарных сегментов;

— получить количественное описание и характеристики технологических событий, как то: временные границы приходящих событий и квазистационарных сегментов, тип и параметры динамики таких сегментов;

— увеличить на 11 % количество выявляемых технологических событий.

Литература

1. **Тристанов А. Б., Геппенер В. В., Фирстов П. П.** Применение методов сегментации к обработке геофизических данных. — СПб.: ЛЭТИ, 2007. http://www.emsd.iks.ru/konf060117lib/pdf/25_geppener_tristanov.pdf (дата обращения 12.11.2010).
2. **Барсегян А. А., Куприянов М. С., Степаненко В. В.** Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004. — С. 67–93.
3. **Вульфин А. М., Гиниятуллин В. М., Фрид А. И.** Нейросетевая модель выявления и распознавания технологических ситуаций в рамках методологии Data Mining // XII Всерос. науч.-техн. конф. «Нейроинформатика — 2010»: сб. науч. тр.: В 2 ч. Ч. 1. М.: МИФИ, 2010. С. 75–84.
4. **Vulfin A. M., Giniyatullin V. M., Frid A. I.** Neural-base Model for Detection and Recognition of Technological Situations within the Scope of Data Mining Strategy // Optical Memory and Neural Networks (Information Optics). 2010. Vol. 19. N 3. P. 207–212.
5. **Вульфин А. М., Фрид А. И.** Интеллектуальная автоматизированная система поддержки принятия решений для технологического комплекса приема-сдачи нефти // Мехатроника, автоматизация, управление. 2011. № 5. С. 29–34.
6. **Зозуля Ю. И.** Интеллектуальные нейросистемы. Науч. сер. Нейрокомпьютеры и их применение. Кн. 12. — М.: Радиотехника, 2003. — С. 14–51.
7. **Орлов Ю.** Разработка и исследование алгоритмов построения иерархических нейросетевых комплексов для классификации данных физических измерений: Автореф. дис. ... канд. физ.-мат. наук // НИИЯФ МГУ. М., 2003. С. 23.