УДК 004.931

СИСТЕМА ФОРМИРОВАНИЯ ОБРАЗА И КЛАССИФИКАЦИИ ВРЕМЕННЫХ РЯДОВ ПО ХАРАКТЕРНЫМ ПОСЛЕДОВАТЕЛЬНОСТЯМ

Горшков А.П., Грызлова Т.П.

Представлено решение широкого класса задач классификации временных рядов. Решение ищется в автоматизированной системе формирования признакового пространства на основе статистик характерных последовательностей временных рядов. Приведены сравнительные оценки вероятности распознавания на известных тестовых задачах «Пистолет – палец», «Переходные процессы», «Полупроводниковая пластина».

Введение

Задача классификации временных рядов является одной из самых актуальных научных задач в области технической диагностики и контроля, автоматического анализа сигналов и изображений, управления и множества других задач автоматизации интеллектуальных функций человека. Ключевой проблемой при решении задачи классификации временных рядов в рамках математического подхода теории распознавания является определение эффективного признакового пространства (пространства образов). Обычно эта задача решается эвристически специалистами предметной области. Известны методы автоматизированного поиска признаковых пространств для классификации временных рядов [1, 2]. Построение подобных систем связано с решением двух задач: определением семейства признаков и разработкой алгоритма поиска эффективного пространства признаков в рамках этого семейства. Для формирования правила классификации при заданном признаковом пространстве может быть использован один из методов математического подхода теории распознавания образов [3, 4]. В настоящей работе показано, что эффективность семейства признаков, отобранных из характерных последовательностей сигналов с помощью алгоритмов и методов системы «Гиперкуб», выше, чем у систем классификации временных рядов, использующих другие типы образов.

Методы классификации временных рядов

Известные методы классификации временных рядов разрабатываются либо на основе представления сигналов как точек в признаковых пространствах, либо на основе структурного описания, либо непосредственно по сигналам, т.е. отличаются по типу образа.

Образы в признаковых пространствах, как правило, вычисляются методами ЦОС, теории динамических систем или прикладной статистики.

Распространённым типом образа является структурное описание, когда сигналы представляются в виде цепочки непроизводных элементов, каждому из которых соответствует сегмент сигнала [2]. В структурном распознавании образ характеризуется множествами непроизводных элементов и отношений между ними [3].

Во многих методах классификации временных рядов используют непосредственно сигналы, без определения признаков или структурных элементов. В частности, к ним относится классификация на основе метода k ближайших соседей и некоторого расстояния на множестве сигналов. Одним из наиболее эффективных методов сравнения сигналов является метод динамической трансформации времени (Dynamic Time Warping) [6 – 8]. Результаты исследований разработанной нами системы формирования образа и классификации временных рядов по характерным последовательностям сопоставляются с альтернативными подходами: методом динамической трансформации времени, системой Zeus и системой обобщенного формирования признаков для структурного распознавания образов.

Метод динамической трансформации времени – это процедура сравнения временных рядов на основе динамического программирования. При сравнении двух сигналов (вычислении расстояния) производится их нелинейное наложение. Классификация выполняется по правилу ближайшего соседа. Базовая процедура динамической трансформации времени заключается в следующем. Пусть заданы два временных ряда: $\mathbf{x} = \mathrm{x}(t) = x_0 x_1 ... x_{N-1}$ и $\mathbf{y} = \mathrm{y}(t) = y_0 y_1 ... y_{M-1}$. Построим матрицу D размера MxN, в каждой ячейке которой указаны квадратичные отклонения: $D_{i,j} = (x_i - y_j)^2$. Расстояние DWT – это минимальная сумма значений ячеек, соответствующих пути из ячейки $D_{0,0}$ в ячейку $D_{N-1,M-1}$. Другими словами, необходимо найти путь с минимальной суммой. Значение этой суммы и будет выражать расстояние между последовательностями. Существует несколько модификаций метода DWT с различными ограничениями на область допустимых ячеек в матрице. Например, в работе [8] область допустимых значений матрицы определяется на основе обучающей выборки, что позволяет повысить точность классификации [6 – 8].

В системе Zeus признаки для распознавания формируются при помощи генетического программирования. Структура признака описывается контекстной грамматикой. В качестве структурных элементов используются операции над сигналами: автокорреляция; определение окна заданием его центра и ширины; индекс максимального (минимального) значения; индекс точки пересечения сигналом заданного значения и другие. На основе алгоритмов генетического программирования производится поиск систем классификации в целом. То есть, одновременно с формированием признаков выполняют объединение их в признаковое пространство, обучение классификатора (например, линейного дискриминатора Фишера) и оценку эффективности полученной системы на обучающей выборке [1].

В методе обобщенного формирования признаков для структурного распознавания классификация производится на основе структурного представления временных рядов, а именно, разбиения их на части и аппроксимации полученных сегментов функциями: константы, линейной, экспоненциальной, синусоиды, кусочно-линейных в форме треугольника и трапеции. Применяется также совмещенное описание, когда разные сегменты описываются различными функциями. Несмотря на структурное представление сигналов, классификация производится в признаковом пространстве. Вектор признаков описывает ряд параметров разбиения на части и аппроксимации, включая позицию сегмента, условный номер функции аппроксимации, параметры аппроксимации, ошибку аппроксимации сегмента и другие [2].

Семейство признаков на базе характерных последовательностей

Предлагается формировать признаки на основе анализа появления похожих подпоследовательностей сигналов. Характерная последовательность (ХП) **cs** – это элементарная последовательность (ЭП), которая неоднократно встречается в сигнале (кластер «Одинаковых» ЭП). Для формализации процедуры сопоставления последовательностей вводится расстояние на множестве ЭП одинаковой длины l и максимально допустимый порог расстояния Th_ρ, при котором ЭП считаются одинаковыми. Распространенными расстояниями для сравнения последовательностей являются метрики Чебышева и СКО:

$$\rho(\mathbf{x},\mathbf{y}) = \max_{i=1..l} |x_i - y_i| \quad (1)$$

$$\rho(\mathbf{x},\mathbf{y}) = \frac{1}{l}\sum_{i=1..l} (x_i - y_i)^2 \quad (2)$$

Наряду с ними используется расстояние на основе коэффициента корреляции $\rho_{x,y}$:

$$\rho(\mathbf{x},\mathbf{y}) = 1 - \rho_{x,y} \quad (3)$$

Основная операция, производимая над сигналом – это сопоставление его ЭП с эталоном. При сопоставлении производится перебор ЭП сигнала с заданным в отсчётах шагом и их сравнение с эталоном по заданному порогу. Семейство признаков для формирования образа включает в себя частотные признаки на основе ХП. Частотный признак f на основе ХП вычисляется как отношение количества ЭП сигнала, расстояние между которыми и эталоном не превосходит порога, к общему числу перебранных ЭП.

Значение частотного признака можно интерпретировать как оценку вероятности попадания последовательностей сигнала **x** в область координатного пространства последовательностей $D_{cs} : \{\mathbf{x} \mid \rho(\mathbf{cs},\mathbf{x}) < Th_\rho\}$, где **cs** - эталонная последовательность, а Th_ρ - порог. Выбор расстояния определяет форму области D_{cs}. Порог позволяет управлять её размерами. Например, ХП с метрикой (1) в пространстве последовательностей соответствует гиперкуб с центром в эталонной последовательности, параллельными осям ребрами и стороной $2T$.

Алгоритмы формирования признаков

Пусть S - обучающая выборка, а $\mathbf{F}$ - искомое признаковое пространство. Поиск признакового пространства $\mathbf{F}$ заключается в последовательном выполнении алгоритмов формирования множества возможных признаков $\mathbf{F}_2$ (рис. 1) и алгоритма поиска эффективного признакового пространства $\mathbf{F} \subset \mathbf{F}_2$ (рис. 2).

Множество возможных признаков строится на основе множества ХП (CS)**.** Множество ХП формируется с помощью кластерного анализа последовательностей сигналов обучающей выборки S. При этом используется простой алгоритм кластерного анализа, изложенный в [3]. Для проведения кластерного анализа задаётся расстояние $\rho(x,y)$, порог Th_ρ, размер l и шаг перебора последовательностей. Формирование множества ХП производится для нескольких размеров последовательностей, расстояний и иногда для нескольких значений порогов. Таким образом, кластерный анализ последовательностей выполняется неоднократно, а полученные множества ХП объединяются. Длины последовательностей задаются на основе визуальной оценки графиков сигналов.

При настройке порогов для кластерного анализа используется его вероятностное представление. Порогу в $n\%$ соответствует такое расстояние d, для которого вероятность появления в сигналах обучающей выборки двух последовательностей с расстоянием, не превосходящим d, составляет $n\%$. Обычно используется порог в 7% или 10%. Иногда дополнительно рассматривается порог в 25% или 30%. Параметры кластерного анализа для трёх задач классификации временных рядов, рассмотренных ниже, представлены в таблице 1.

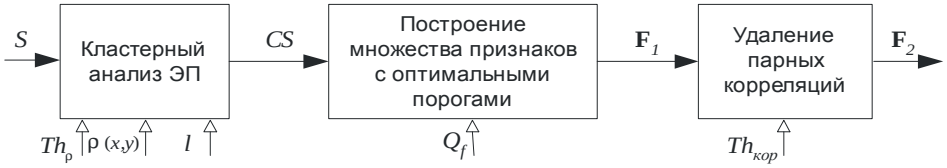

Рис. 1. Структурная схема алгоритма формирования множества возможных признаков для семейства частотных признаков на основе ХП

Таблица 1.

Параметры кластерного анализа последовательностей

Задача	Размеры последовательностей	Порог расстояния
«Переходные процессы»	5, 10, 25, 50, 75	10%, 30%
«Полупроводниковая пластина»	5, 10, 20, 30, 40, 50, 75, 100	10%, 30%
«Пистолет - палец»	10, 20, 30, 40, 50, 75, 100	7%, 30%

На основе множества ХП формируется множество признаков с оптимальными порогами за счёт оптимальной (по критерию максимума заданной меры информативности признаков Q_f) установки порогов для каждой из ХП. После этого из множества полученных признаков удаляется по одному из парно коррелированных по пороговому значению коэффициента корреляции $Th_{кор}$ признаку.

При автоматизированном формировании множество возможных признаков $\mathbf{F}_2$ имеет большие размеры (больше 500), поэтому выполняется отбор заданного числа наиболее информативных признаков. При этом формируется множество признаков $\mathbf{F}_3 \subset \mathbf{F}_2$. Далее выполняется случайный поиск с адаптацией (метод СПА), для которого необходимо задать меру информативности признаковых пространств $Q_{\mathbf{F}}$ [4]. Информативность признакового пространства определяется кластерными свойствами расположения множеств точек, соответствующих разным классам.

При использовании в системе преобразования кластеризации, оно включается в алгоритм $Q_{\mathbf{F}}$, и информативность оценивается для преобразованного пространства. В отличие от алгоритмов оценки информативности пространств признаков $Q_{\mathbf{F}}$ алгоритмы вычисления информативности признаков Q_f удаётся сделать более быстрыми: с линейной оценкой вычисления относительно размера обучающей выборки, что и определяет необходимость их отдельного использования.

налов. Время синтеза зависит от таких параметров задачи, как количество классов, размер обучающей выборки и длина реализации сигналов. Программный комплекс «Гиперкуб» позволяет решать задачу повышения качества уже разработанной системы диагностики, если она работает по принципу вычисления признаков. Для этого должна быть предоставлена таблица значений признаков существующей системы классификации для сигналов обучающей выборки. Помимо технической диагностики, система «Гиперкуб» позволяет решать и другие задачи классификации временных рядов. Готовятся публикации с успешными результатами решения трех задач классификации временных рядов, представляющих собой модельные данные, сигналы из областей технической диагностики и распознавания изображений.

Сравнительная оценка эффективности систем классификации

Испытания проводились на трёх наборах временных рядов, доступных в англоязычном Интернете: «Переходные процессы» («trace»), «Пистолет - палец» («gun-point») и «Полупроводниковая пластина» («wafer») [5 – 8, 2].

Множество сигналов «Переходные процессы» является подмножеством набора данных для распознавания переходных процессов (the Transient Classification Benchmark), впервые представленного в [9]. Оно сформировано для воспроизведения недостатков оборудования ядерных электростанций. Полный набор данных включает 16 клас-

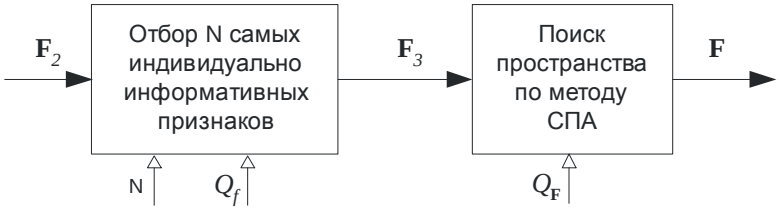

Рис. 2. Структурная схема алгоритма поиска информативного пространства признаков

Таблица 2.

Параметры формирования признаков

Параметр	Значение
Максимально допустимое значение коэффициента корреляции между признаками при формировании множества с оптимальными порогами	0,9
Количество N наиболее индивидуально информативных признаков, передаваемых на СПА	500
Количество эпох в методе СПА	1000
Количество шагов в эпохе	50

В таблице 2 представлены значения управляющих параметров алгоритмов формирования признаков, использованные при решении тестовых задач классификации временных рядов.

Программная система классификации временных рядов «Гиперкуб» создана на платформе C++ Builder 6 и позволяет за одну–две недели получить алгоритм автоматической диагностики, приложение и библиотеку с функциями вычисления признаков и принятия решения. Для синтеза процедуры классификации необходима обучающая выборка – множество классифицированных сигналов по 50 экземпляров каждого. Каждый экземпляр характеризуется 4 сигналами. Как и в ряде работ [5 – 8, 1] используется уменьшенный набор данных, включающий только второй сигнал для классов 2 и 6, а также третий сигнал для классов 3 и 7. Таким образом, получаются 4 класса, каждый из которых характеризуется одним сигналом с длиной реализации 275 отсчётов. Уменьшение количества рассматриваемых процессов, характеризующих один экземпляр, связано с избыточностью такого описания. На рис. 3 представлены примеры процессов набора данных «Переходные процессы».

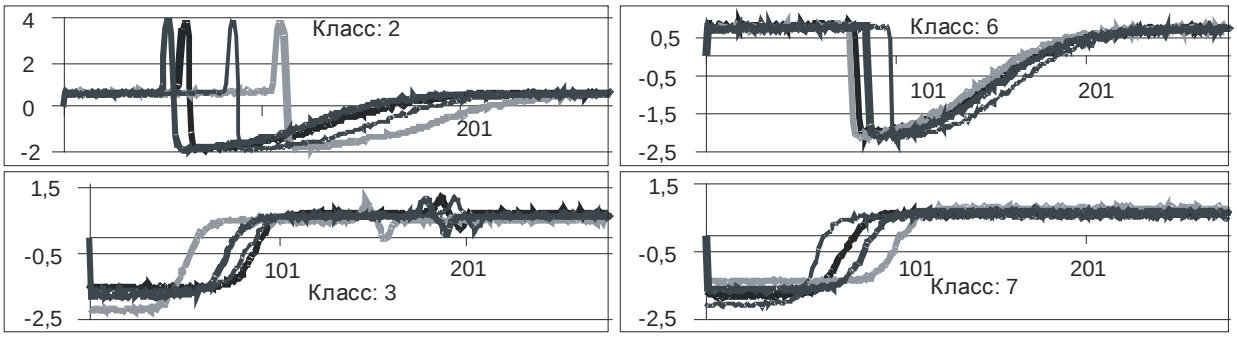

Рис. 3. Примеры процессов набора данных «Переходные процессы»

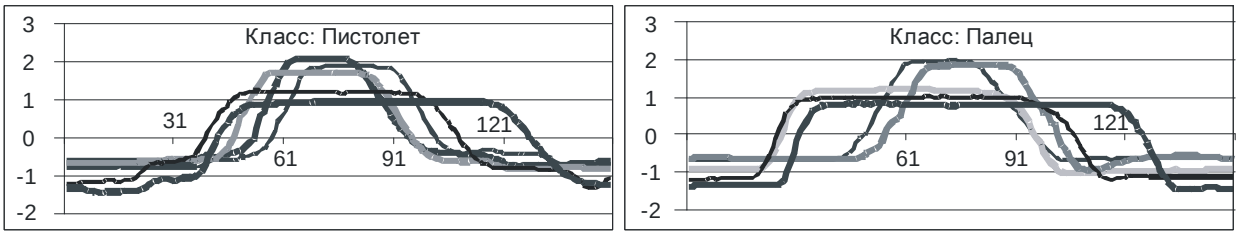

Рис. 4. Примеры процессов набора данных «Пистолет - палец»

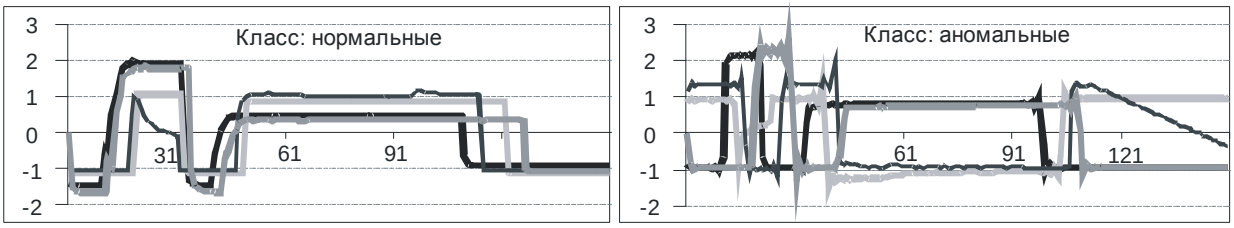

Рис. 5. Примеры процессов набора данных «Полупроводниковая пластина»

Набор данных «Пистолет - палец» относится к области видеонаблюдения. Имеются процессы двух классов, каждый из которых представлен сотней экземпляров. Процессам класса «пистолет» соответствует следующая последовательность действий:

- актёр держит руки по швам;
- достаёт пистолет из кобуры на бедре и, примерно в течение одной секунды, наводит его на цель;
- возвращает пистолет обратно в кобуру.
- Порядок действий при формировании процессов класса «палец»:
- актёр держит руки по швам;
- показывает указательным пальцем на цель в течение приблизительно одной секунды;
- возвращает руку на место.

Сигналы каждого класса построены с участием двух актеров: мужчины и женщины. Процессы отражают изменение координаты *x* центра тяжести правой руки актёра. Длина реализации составляет 150 отсчётов [6 – 8]. Примеры процессов представлены на рис. 4.

Набор данных «Полупроводниковая пластина» является коллекцией временных рядов, содержащих последовательности показаний интенсивности излучений двух различных длин волн (405 и 520 нанометров), регистрируемых в процессе напыления при производстве полупроводниковых устройств [2]. Используется набор данных, где отсутствует информация о том, какой длине волны соответствует тот или иной сигнал [6 – 8].

Каждый сигнал классифицируется как нормальный или аномальный. Аномальные сигналы представляют ряд проблем, возникающих при производстве полупроводниковых устройств. Всего имеется 7164 сигналов. Среди них 762 – аномальных. Обучающая выборка: 717 сигналов (10%), из них 77 аномальных. Тестовая выборка: 6447 сигналов, из них аномальных – 685. Примеры сигналов показаны на рис. 5.

В работах [4 – 7] представлены результаты экспериментальной проверки метода динамической трансформации времени (Dynamic time warping, DWT) для классификации рассмотренных временных процессов. Для оценки вероятности правильного распознавания (**p**) в [5] набор данных разбивается на тестовую и обучающую выборки. Оценка **p** вычисляется как отношение правильно распознанных процессов тестовой выборки к общему её объему. Такой критерий обозначим как n-TT, где n – процентное выражение доли тестовой выборки. В работах [6 – 8] оценка **p** производится при помощи более оптимистичного критерия. При этом перебираются все экземпляры набора данных. Проверка правильности распознавания каждого из них производится по ближайшему соседу на основе оставшихся. Такой критерий обозначим 1-NN (leav one out).

В [1] получена оценка **p** для классификации наборов данных «Переходные процессы» и «Пистолет - палец» на основе системы Zeus. Оценка **p** производилась по методу скользящего контроля по 10 блокам. Обозначим этот критерий: 10-CV.

Таблица 3.

Оценка вероятности правильного распознавания – **p**

Метод		«Переходные процессы»		«Полупроводниковая пластина»		«Пистолет - палец»	
		Критерий	p, %	Критерий	p, %	Критерий	p, %
На основе ХП с расстоянием	(1)	76-TT	100	90-TT	96.13	75-TT	99.3
	(2)	76-TT	100	90-TT	95.18	75-TT	94.6
	(3)	76-TT	100	90-TT	99.94	75-TT	99.3
	(1), (2) и (3)	76-TT	100	90-TT	99.98	75-TT	100
Метод динамической трансформации времени	[5]	50-TT	100	86-TT	99.5	75-TT	91.3
	[6 - 7]	1-NN	100	-	-	1-NN	99
	[8]	1-NN	100	-	-	1-NN	99.5
Система Zeus [1]		10-CV	100	-	-	10-CV	98.5
Обобщенное формирование структурных признаков [2]		-	-	10-CV	98.6	-	-

При вычислении критерия обучающая выборка случайным образом разбивается на 10 непересекающихся подмножеств одинакового или почти одинакового размера. Обучение и тестирование производится 10 раз. Каждый из блоков по очереди используется в качестве тестовой выборки, а обучение производится по 9-и оставшимся блокам. В качестве оценки **p** используется среднее из полученных 10-и значений. Таким образом, размер тестовой выборки при каждом тестировании составляет 10% размера выборки.

В [2] представлены результаты классификации процессов набора данных «Полупроводниковая пластина» при помощи метода обобщенного формирования признаков для структурного распознавания. Оценка **p** производится по процедуре скользящего контроля 10-TT. В работе представлены результаты для всех функций аппроксимации. Результаты получены отдельно для сигналов разной длины волны 405 и 520. Кроме того, вместо общей вероятности **p** даны отдельно вероятности правильного распознавания для сигналов каждого из классов: $\mathbf{p_a}$ и $\mathbf{p_n}$. Для сопоставления значения **p** мы взяли наибольшие показатели по разным функциям аппроксимации. Значение **p** вычисляется на основе $\mathbf{p_a}$ и $\mathbf{p_n}$ по следующей формуле: $\mathbf{p} = (C_a \cdot \mathbf{p}_a + C_n \cdot \mathbf{p}_n) / (C_a + C_n)$, где C_a и C_n – количества сигналов каждого из двух классов.

В таблице 3 представлены оценки вероятности правильного распознавания для всех трёх наборов данных в сравнении с другими подходами. На основе ХП системы распознавания строятся для каждого расстояния (1), (2) и (3) в отдельности и на объединенном множестве признаков для всех трёх расстояний. Решение, основанное на характерных последовательностях, оказывается более эффективным. Результаты тестирования на независимой выборке являются в данном случае более пессимистичными, чем оценки 1-NN и 10-CV. Прежде всего, это связано с уменьшением объема обучающей выборки. Как показано в [6, 7], для набора данных «Пистолет - палец», даже незначительное уменьшение объема обучающей выборки приводит к заметному ухудшению результатов.

Заключение

Анализ трёх различных задач классификации временных рядов показал, что формирование образа на основе ХП позволяет получить более эффективные системы распознавания, чем системы на основе динамической трансформации времени (DWT) [5 – 8], а также структурного описания при помощи системы Zeus [1] и обобщенного формирования структурных признаков [2]. Построенные системы легко поддаются интерпретации и имеют достаточно простую структурную схему.

В [10] представлено решение задачи диагностики состояния подшипников трансмиссии ГТД по вибрационным сигналам, основанное на представленном семействе признаков. Успешное применение частотных признаков на основе ХП при работе с различными типами сигналов позволяет сделать вывод о гибкости и, возможно, универсальности предложенной системы классификации временных рядов.

Применение расстояния на основе коэффициента корреляции для рассмотренных задач является более эффективным, чем использование метрик Чебышева и СКО. В работе [10], напротив, показана преимущественная эффективность метрик СКО и Чебышева. Это не позволяет отказаться от использования ни одной из рассматриваемых метрик в общем случае. Кроме того, наилучшие результаты получаются именно при объединении признаков, полученных для всех трёх расстояний.

Предложенная система применима для синтеза процедур классификации временных рядов различных типов. Малые временные затраты на синтез делают применение этой технологии экономически целесообразным.

Литература

1. Eads D., Glocer K., Perkins S., Theiler J. Grammar-guided feature extraction for time series classification. Neural Information Processing Systems, 2005.
2. Olszewski R. T. Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2001.
3. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М: Мир, 1978.
4. Загоруйко Н.Г. Методы распознавания и их применение. - М: Советское радио, 1972.
5. Keogh E. UCR Time Series Archive: www.cs.ucr.edu/~eamonn/TSDMA/, 2007.
6. Xiaopeng Xi, Keogh E., Shelton C., Li Wei & Ratanamahatana C.A. Fast Time Series Classification Using Numerosity Reduction. International Conference on Machine Learning, 2006.
7. Ratanamahatana C. A., Keogh E. Everything you know about dynamic time warping is wrong. In 10th ACM SIGKDD Interna-

tional Conference on Knowledge Discovery and Data Mining Workshop on Temporal Data Mining, 2004.
8. Ratanamahatana, C. A. and Keogh. E. Making Time-series Classification More Accurate Using Learned Constraints. SIAM International Conference on Data Mining, April 22-24, 2004.
9. Roverso D. Multivariate temporal classification by windowed wavelet decomposition and recurrent neural networks. In 3rd ANS International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface, 2000.
10. Горшков А.П., Грызлова Т.П., Комаров Б.И., Шепель В.Т. Диагностика состояния подшипников трансмиссии газотурбинных двигателей в пространствах статистик характерных последовательностей вибраций. // Авиационно–космическая техника и технология 10 / 36. - Харьков, ХАИ, 2006.

ПОЗДРАВЛЕНИЯ ЮБИЛЯРУ

Заместителю Главного редактора нашего журнала, лауреату премии правительства РФ в области науки и техники, доктору технических наук, профессору ДВОРКОВИЧУ Виктору Павловичу – 70 лет!

Свой юбилей Виктор Павлович встречает на новом ответственном государственном посту – заместителя директора ФГУП «Главный радиочастотный центр», проработав до этого 45 лет в ФГУП «НИИ радио». Вся творческая судьба Виктора Павловича была связана с этим ведущим российским институтом, начальником отдела в котором он был последние годы.

Понимая жизненную важность и научно-техническую значимость продвижения в России новых информационных технологий, опирающихся на цифровую обработку сигналов и изображений в реальном времени, Виктор Павлович стал одним из инициаторов организации и проведения в нашей стране международной научно-технической конференции: «Цифровая обработка сигналов и ее применение», которая уже в 10-й раз прошла в марте этого года. И с самого первого дня профессор Дворкович В.П. является бессменным руководителем секции: «Обработка изображений».

С созданием научно-технического журнала: «Цифровая обработка сигналов» в 1999 году, став заместителем Главного редактора, Виктор Павлович проводит огромную редакционную работу. Фактически ежегодно под его руководством на страницах нашего журнала формируется тематический выпуск «Цифровая обработка и передача изображений».

Виктор Павлович известен как один из ведущих российских ученых, плодотворно работающих в области современных телекоммуникаций, цифрового телевидения, обработки изображений. Он автор более 200 научных работ (в том числе более 10 книг), имеет более 70 авторских свидетельств и патентов.

Отмечая активную научную и организаторскую деятельность Дворковича В.П., как ученого и руководителя творческого коллектива, члена экспертного Совета ВАК РФ, члена оргкомитета МНТК «Цифровая обработка сигналов и ее применение», заместителя Главного редактора нашего журнала, редакционная коллегия журнала «Цифровая обработка сигналов» поздравляет Виктора Павловича с 70-летием! Крепкого Вам здоровья, новых научных достижений, верных друзей и семейного благополучия.

Главный редактор Ю.Б. Зубарев
Зам. Главного редактора В.В. Витязев