Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

ИССЛЕДОВАНИЯ ОПЕРАЦИЙ В ЗАДАЧАХ ПРОГРАММНОЙ ИНЖЕНЕРИИ

Часть 2

Учебное пособие

Рекомендовано ученым советом федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет» для обучающихся по образовательным программам высшего образования по направлениям подготовки 09.04.01 Информатика и вычислительная техника, 09.04.04 Программная инженерия

Оренбург
2019

УДК 519.8:004.42(075.8)
ББК 22.18я73+32.973я73
И88

Рецензенты:
доктор технических наук, профессор А.М. Пищухин

Авторы: Н.А. Соловьев, Н.А. Тишина, Е.Н. Чернопрудова, Л.А. Юркевская

И 88 Исследования операций в задачах программной инженерии: учебное пособие. Часть 2 /Н.А. Соловьев [и др.]; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2019. – 119 с.

ISBN

В учебном пособии рассмотрены методы исследования операций для обоснования решений в задачах программной инженерии: цифровая обработка изображений, текстовой информации, сетевого трафика. Материал дополнен примерами задач с решениями, вопросами для проверки усвоения материала.

Учебное пособие предназначено для студентов, обучающихся по направлениям подготовки 09.04.01 Информатика и вычислительная техника, 09.04.04 Программная инженерия при изучении следующих дисциплин: «Методология научных исследований проблем информатики и вычислительной техники», «Методология научных исследований проблем программной инженерии», «Системы поддержки принятия решений», «Методы теории принятия решений», и может быть использовано для самостоятельной работы аспирантов по направлениям подготовки 09.06.01 Информатика и вычислительная техника, 10.06.01 Информационная безопасность

.

УДК 519.8:004.42(075.8)
ББК 22.18я73+32.973я73

Содержание

Введение

Методы исследования операций являются основным инструментом повышения обоснованности управленческих решений и представляют собой сложные программно-аппаратные и коммуникационные комплексы. Такие системы предназначены для сбора, хранения, обработки и распространения информации в целях управления, получившие наименование MIS (Management Information Systems) – информационно-управляющие системы, основу которых составляют системы поддержки принятия решений (СППР).

Настоящее учебное пособие является продолжением учебных пособий «Основы теории принятия решений» и «Исследование операций в задачах программной инженерии», выпущенных авторским коллективом ранее для программы бакалавриата по направлениям обучения «Информатика и вычислительная техника и «Программная инженерия», и ориентировано на магистерские программы «Информационное и программное обеспечение автоматизированных систем» и «Разработка информационно-телекоммуникационных систем».

Материал доработан на основе многолетнего опыта преподавания дисциплин «Методология научных исследований», «Системы поддержки принятия решений», «Методы теории принятия решений» на кафедре программного обеспечения вычислительной техники и автоматизированных систем ФГБОУ ВО «Оренбургский государственный университет».

Пособие включает три раздела. В первом разделе изложены методы исследования операций в задачах компьютерного зрения, как наиболее конструктивного направления практического приложения средств поддержки принятия решений в задачах идентификации и распознавания объектов на изображениях. Раздел дополнен результатами исследований авторского коллектива (Соловьев Н.А., Бугаёв Д.П., Кузьмин М.И.) по задачам контроля качества продукции станов листового металлопроката.

Второй раздел учебного пособия посвящена методам исследования операций в задачах защиты информации, как наиболее конструктивного направления практического приложения средств поддержки принятия решений в задачах обеспечения информационной безопасности. Раздел дополнен результатами исследований авторского коллектива (Соловьев Н.А., Тишина Н.А., Чернопрудова Е.Н. Юркевской Л.А.) по проблемам обеспечения информационной безопасности информационно-телекоммуникационных систем корпоративных предприятий с распределенной структурой.

В третьем разделе (авторский коллектив: Соловьев Н.А., Чернопрудова Е.Н.) изложены методы исследования операций в задачах цифровой обработки текстовой информации применительно к фильтрации контента электронной почтовой корреспонденции.

Раздел 1 Исследования операций в задачах компьютерного зрения

Любая из процедур цифровой обработки изображений опирается на модель – формализованное описание, выполненное с определенной степенью абстрагирования. Роль модели изображения в процессе извлечения информации состоит в обеспечении адекватного описания существенных свойств изображения, позволяющего дать конструктивную основу для построения эффективных вычислительных процедур.

Глава 1 Методический аппарат цифровой обработки изображений

1.1 Систематизация моделей для цифровой обработки изображений

Цифровая обработка изображений осуществляется после преобразования изображения в цифровую форму. Качество цифровой обработки зависит от адекватности модели реальному изображению. Изображение может быть представлено функцией пяти аргументов: трех пространственных координат ***x,y,z***, времени *t* и длины волны электромагнитного излучения λ .

Выбор модели изображения для задачи идентификации и распознавания поверхностных дефектов тонколистового проката (практическая задача программной инженерии) представлено на рисунке 1.1.

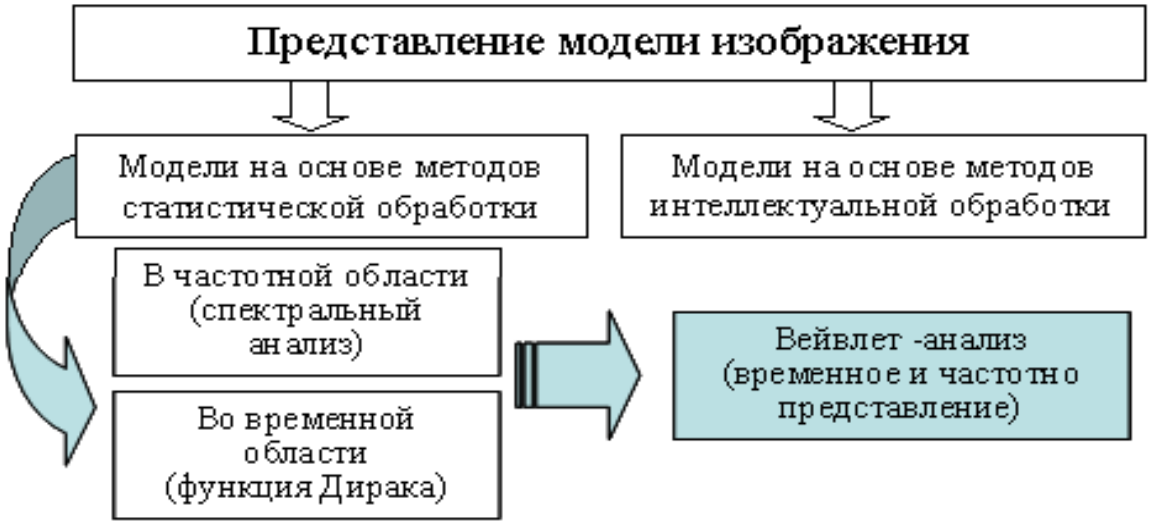


Рисунок 1.1 – Систематизация моделей изображения

Спектральные модели изображений. При цифровой обработке изображений широко применяется спектральный анализ, как способ перехода от представления модели из временной области в частотную (спектральную). Исторически первым способом спектрального анализа являлось преобразование Фурье [11]. Спектр изображения на основе двумерного прямого преобразования Фурье описывается функцией:

$$F(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\exp(-i(\omega_x x + \omega_y y))dxdy,$$

где ω_x, ω_y - пространственные частоты (гармоники);

$i = \sqrt{-1}$, мнимая единица.

Функция $\exp(-i(\omega_x x + \omega_y y))$ описывает плоскую волну изображения (x,y) при фиксированных значениях пространственных частот.

Действительная функция яркости изображения ***f(x,y)*** с комплексной функцией частоты, называемой спектром изображения, связывает преобразованием вида

$$F(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\cos(\omega_x x + \omega_y y)dxdy +$$

$$+ i\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(-f(x,y))\sin(\omega_x x + \omega_y y)dxdy = \mathrm{Re}(\omega_x, \omega_y) + i\,\mathrm{Im}(\omega_x, \omega_y),$$

где $\mathrm{Re}(\omega_x,\omega_y)$ – реальная часть спектра;

$\mathrm{Im}(\omega_x,\omega_y)$ – мнимая часть спектра.

Амплитуда и фаза спектра определяются по зависимостям:

$$A(\omega_x, \omega_y) = \sqrt{\mathrm{Re}(\omega_x, \omega_y)^2 + \mathrm{Im}(\omega_x, \omega_y)^2},$$

$$\varphi(\omega_x, \omega_y) = arctg(\mathrm{Im}(\omega_x, \omega_y) / \mathrm{Re}(\omega_x, \omega_y)).$$

Используя обратное преобразования Фурье возможно по спектру восстановить изображение:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(\omega_x,\omega_y)\exp(i(\omega_x x + \omega_y y))d\omega_x d\omega_y \ .$$

Помимо преобразования Фурье, спектр изображения может быть получен на основе вейвлет-преобразования (ВП) [6], подробнее данная модель будет рассмотрена ниже. Сравнивая два данных метода описания изображений можно отметить следующие преимущества использования ВП: возможность преобразования с различным размером окна, возможность использования различных базисных функций. Кроме того, ВП является пространственно-частотным, поэтому можно анализировать коэффициенты вейвлет-преобразования (КВП) только в идентифицированных областях изображения $R_{деф}$. Также следует отметить, что КВП рассчитываются в процессе идентификации поверхностных дефектов и, поэтому, их использование в процессе расчета значений признаков распознаваемого объекта позволяет сократить общий объем вычислений при его распознавании.

Вероятностные модели изображений и функции автокорреляции. Вероятностные модели получили широкое распространение при описании изображений. Изображение в этом случае рассматривается как случайная функция времени ***t*** и пространственных координат ***(x,y).*** Случайный процесс является стационарным в широком смысле, если он имеет постоянные значения дисперсии и математического ожидания, а его функция автокорелляции зависит от разностей координат. Случайный процесс является стационарным в узком смысле, если ***n***-мерная плотность распределения вероятностей сигнала инвариантна к сдвигу. В этом случае моменты более высокого порядка не зависят от времени (в частности, эксцесс и асимметрия). Случайный процесс для некоторого фиксированного момента времени ***t*** описывается плотностью вероятности распределения яркости в изображении по пространственным координатам.

В соответствии с определением математическое ожидание (среднее) определяется для стационарного процесса:

$$Mf = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,p(x,y)dxdy = const$$

Дисперсия:

$$Df = \sigma^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (f(x,y) - Mf)^2\,p(x,y)dxdy = const$$

Функция автокорреляции вычисляется как:

$$R(\tau_x,\tau_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,f(x-\tau_x, y-\tau_y)dxdy,$$

где τ_x, τ_y – сдвиги изображения по соответствующим осям координат.

Для действительной функции ***f*** автокорреляционная функция ***R*** является четной и действительной.

Таким образов, спектральной представление оцифрованного изображения наиболее приемлемо для решения задачи выделения структурных элементов (поверхностных дефектов) с учетом заданных требований.

1.2 Моделирование изображений для выделения структурных элементов

Описание моделей изображений с целью выделения их структурных элементов (поверхностных дефектов) базируется на математическом аппарате кратномасштабного анализа (КМА), в основе которого лежат операторы ВП ***W***.

Под ВП понимается разложение изображения по системе базисных функций, каждая из которых является сдвинутой и масштабированной (сжатой или растянутой) копией одной функции – порождающего вейвлета [2,6,11].

В процедурах разложения изображений ***I*** с помощью локальных операторов (базисных, порождающих функций) применяются так называемые локаль-

ные модели случайного поля, характеризующие статистическую зависимость интенсивности изображения ***f*** в точке ***(x,y),*** от значений интенсивности в соседних точках, представляя ***f(x,y)*** как линейную комбинацию значений ***{f(x+k,y+l), (k,l)∈F}*** и аддитивного шума ***ε***, где ***F*** – множество соседей, не включающих точку ***f(x,y)***:

$$f(x,y) = \sum_{(k,l)\in F} a_{k,l} f(x+k, y+l) + \varepsilon(x,y),$$

где $A=\{a_{k,l}, (k,l)\in F\}$ – вектор неизвестных коэффициентов модели.

Выражение ***f(x,y)*** можно переписать как ***f = f* + ε,*** где ***f*** – исходное изображение, ***f**** – поле значений пикселей, ***ε*** – поле случайного шума. После проведения ВП выполняется $W_f = W_f^* + W_\varepsilon$. Это означает, что коэффициенты двумерного ВП могут быть описаны, как показано на рисунке 3.1, посредством линейной комбинации соседних коэффициентов из четырех наборов ***W_{LL}, W_{HL}, W_{LH}, W_{HH}*** .

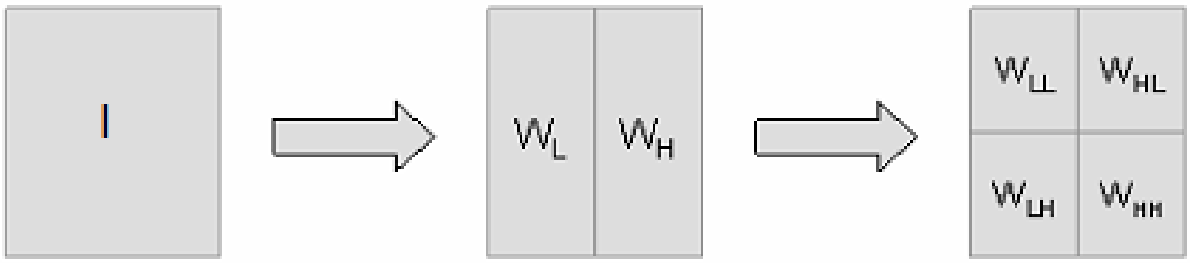


Рисунок 2.1 – Технология двумерного ВП кадра изображения ***I***

Другой подход предполагает рассмотрение поля яркости, соответствующего изображению, как статистически однородного случайного поля. В данном случае задание модели сводится к количественному описанию тех или иных характеристик случайного поля: одномерной или многомерной плотности распределения вероятностей, функции корреляции и т.п. Очевидно, что подобный

подход можно применить и для описания свойств коэффициентов $a_{i,l}$ на разных уровнях разрешения, а их наборы в W_{HL}, W_{LH}, W_{HH} рассматриваются как случайные двумерные некоррелированные поля с нулевым средним. Как показывает практика, хотя общая гистограмма всех коэффициентов является экспоненциальной ($a = exp(-\lambda k)$), гистограммы каждого отдельного шага преобразования являются симметричными, одномодальными, и, приближенно, могут быть приняты за нормальные.

Исходя из этого, математическая модель полутонового изображения может быть представлена в виде [6,11]

$$f(x,y) = \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty}(\varepsilon_{1j,k}^{(i)}\varphi\psi_{j,k}^{(i)}(x,y) + \varepsilon_{2j,k}^{(i)}\psi\varphi_{j,k}^{(i)}(x,y) + \varepsilon_{3j,k}^{(i)}\psi\psi_{j,k}^{(i)}(x,y)),$$

где $\varepsilon_{1j,k}^{(i)}, \varepsilon_{2j,k}^{(i)}, \varepsilon_{3j,k}^{(i)}$ – случайные поля ***i***-го поля изображения, с нормальным распределением, нулевым средним и дисперсиями, соответственно, равными дисперсиям W_{HL}, W_{LH}, W_{HH}.;

$\varphi\psi(x,y) = \varphi(x)\psi(y), \psi\varphi(x,y) = \psi(x)\varphi(y), \psi\psi(x,y) = \psi(x)\psi(y)$ – тензорные произведения одномерных порождающих функций вейвлета.

Очевидно, что подобные модели являются весьма удобными при синтезе алгоритмов разложения изображений, в частности, позволяют развить применительно к ВП изображений основные результаты классической теории фильтрации. Кроме того, данная модель, основанная на статистических свойствах коэффициентов разложения, позволяет описать процедуры вейвлет – анализа изображений.

Другим подходом к описанию изображения является его представление в виде совокупности областей, которые могут различаться по характеристикам. Разработка модели в рамках такого подхода заключается, во-первых, в описании пространственных признаков областей (размеров, взаимного расположения, формы границ и т.п.) и, во-вторых, в описании свойств поля внутри каж-

дой области. Очевидно, что подобные модели являются наиболее подходящими для описания алгоритмов выделения структурных элементов изображения, сегментации и т.д.

Порождающие (базисные) функции ВП, которые называются вейвлетами, могут быть самые различные функции с компактным носителем, в том числе со скачками, разрывами и перепадами значений с большой крутизной. Выбор порождающего вейвлета во многом определяется тем, какую информацию необходимо извлечь из изображения. С учетом характерных особенностей различных вейвлетов во временном и в частотном пространстве, можно выявлять в анализируемых изображениях те или иные свойства и особенности, которые незаметны на гистограммах, особенно в присутствии сильных шумов. При этом задача реконструкции сигнала может и не ставится, что расширяет семейство используемых регулярных и симметричных порождающих вейвлетов. Более того, вейвлет может конструироваться непосредственно под ту локальную особенность в изображении, которая подлежит выделению или обнаружению, если ее форма априорно известна.

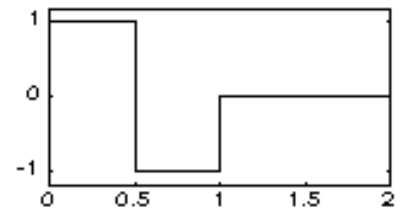
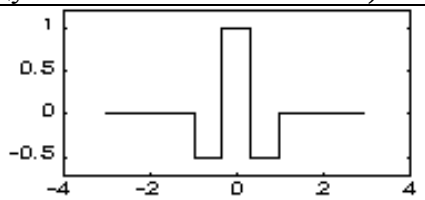
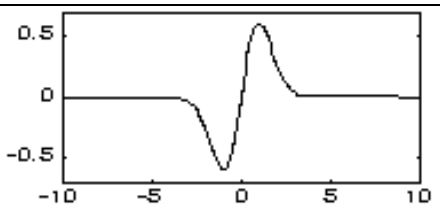
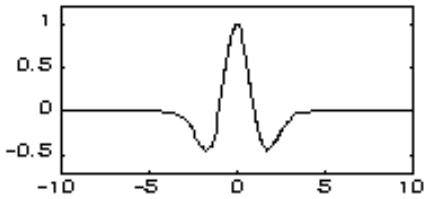
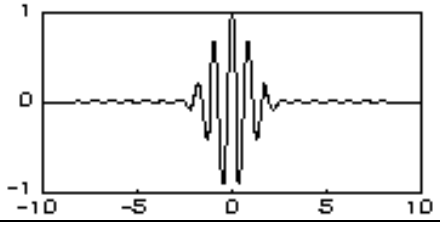
Существует большое количество вейвлет-базисов, но применительно к обработке изображений используются ортогональные и биортогональные вейвлеты. В общем случае ортогональность вейвлетов не является обязательным требованием при анализе двухмерных сигналов. Однако ортогональность базиса позволяет реализовать быстрые алгоритмы.

Другими важными свойствами вейвлет-функций являются симметричность и компактность носителя в пространственной и частотной области.

Симметричность (асимметричность) функции в некоторой степени определяет ориентацию вейвлет-базиса. Ориентация базисной функции в свою очередь определяет способность корректно анализировать ориентированные структуры, типичные для изображений. Точность определения местоположения деталей изображения на стадии анализа определяется пространственной локализацией вейвлет-функции.

В таблице 1.1 представлены основные базисные вейвлет-функции, используемые для обработки изображений [2,14].

Таблица 1.1 –Базисные вейвлет-функции

HAAR - вейвлет:	
$\psi(t)=\begin{cases}1, & 0\le t<1/2\\ -1, & 1/2\le t<1\\ 0, & t<0, t\ge 1\end{cases}$	
FHAT - вейвлет ("Французская шляпа" - French hat):	
$\psi(t)=\begin{cases}1, & \lvert t\rvert\le 1/3\\ -1/2, & 1/3<\lvert t\rvert\le 1\\ 0, & \lvert t\rvert>1\end{cases}$	
Wave - вейвлет:	
$\psi(t)=t\exp(-\frac{t^2}{2})$	
MHAT - вейвлет ("Мексиканская шляпа" - Mexican hat):	
$\psi(t)=(1-t^2)\exp(-\frac{t^2}{2})$	
Вейвлет Морле (образует комплексный базис):	
$\psi(r)=\exp(ik_0 r-\frac{r^2}{2})$	

Локальность функции в частотной области важна при подробном анализе спектра. Таким образом, степень симметричности или отклонения от симметрии для вейвлет-функции наряду с ее пространственной локализацией влияет на определение формы и ориентации деталей изображения, а также их про-

странственных координат. Для анализа изображений желательно использовать симметричные ортогональные вейвлет-базисы.

Принципиальным отличием вейвлет-преобразования от фурье-преобразования является возможность вейвлетов анализировать нестационарные сигналы с изменением компонентного содержания во времени или в пространстве. Основная область применения ВП – анализ и обработка сигналов и функций, нестационарных во времени или неоднородных в пространстве, когда результаты анализа должны содержать не только общую частотную характеристику сигнала (распределение энергии сигнала по частотным составляющим), но и сведения об определенных локальных координатах, на которых проявляют себя те или иные группы частотных составляющих, или на которых происходят быстрые изменения частотных составляющих сигнала.

По сравнению с разложением функций на ряды Фурье, вейвлеты способны с гораздо более высокой точностью представлять их локальные особенности, вплоть до разрывов 1-го рода (скачков). Кроме того, ВП одномерных сигналов обеспечивает двумерную развертку, при этом частота и координата рассматриваются как независимые переменные, что дает возможность анализа сигналов сразу в двух пространствах.

Исследования показали, что изображения могут быть описаны моделью двумерного случайного поля в виде суммы двух компонент для идентификации поверхностных дефектов листового проката

$$f(x, y) = f_1(x, y) + f_2(x, y),$$

где x, y – координаты изображения;

$f(x, y)$ – поле яркости;

$f_1(x, y)$ – яркость стационарного поля (текстурная компонента);

$f_2(x, y)$ – яркость меняющегося поля дефектов.

Такая модель позволит при принятии решений анализировать не всё изображение, а только области дефектов (вторая составляющая модели), что значительно сократит объём обрабатываемой информации.

Таким образом, наиболее приемлемой моделью описания изображения листового проката с поверхностными дефектами является модель вейвлет – разложения (преобразования) изображения в виде суммы двух компонент – текстурной компоненты и медленно меняющегося поля (области) дефектов.

Основным недостатком такого подхода является высокая сложность алгоритмической реализации, что требует развития аппарата вейвлет-анализа к задачам идентификации и распознавания поверхностных дефектов в процессе производства.

1.3 Развитие вейвлет-модели изображения для задачи идентификации поверхностных дефектов

В основе КМА [2,6] лежат иерархические свойства масштабирующих скейлинг-функций $\varphi_{m,k}(x)$ и детализирующих вейвлет-функций $\psi_{m,k}(x)$, которые позволяют представить любую одномерную функцию $f(x)$ в виде ее последовательных вейвлет-преобразований (см. рис. 1.2).

$V_{m+1}=V_m \oplus W_m = V_{m-1} \oplus W_{m-1} \oplus W_m$

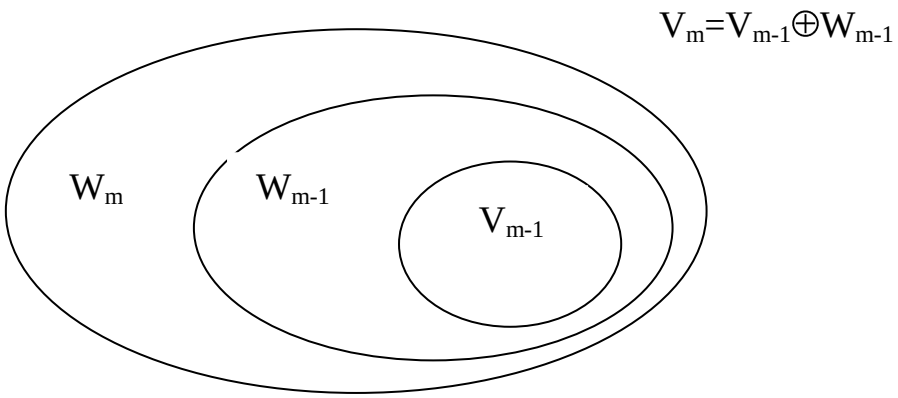


Рисунок 1.2 – Взаимосвязь функциональных пространств МРА

Пусть пространство V_m , где m – масштаб преобразования, натянуто на $\varphi_{m,k}(x)$, т.е. $V_m = span\{\varphi_{m,k}(x)\}$, и обладает свойствами иерархичности $(f(x) \in V_m \Leftrightarrow f(2x) \in V_{m+1})$ и вложенности $(V_0 \subset V_1 \subset \ldots \subset V_m \subset \ldots \subset L^2(R))$. Тогда, пространство V_{m+1} может быть определено как сумма пространства V_m и дополнения к нему W_m (базис пространства W_m образуют вейвлеты $\psi_{m,k}(x)$), т.е. $V_{m+1} = V_m \otimes W_m$ (рис 1.2).

Тогда, произвольную функцию $f(x) \in L^2(R)$ можно аппроксимировать последовательностью функций $f_m(x) \in V_m$ в соответствии с зависимостью, которая хорошо согласуется с моделью (1.11)

$$f(x) = \sum_{k=-\infty}^{\infty} c_{m,k}\, \varphi_{m,k}(x) + \sum_{m=m'}^{\infty} \sum_{k=-\infty}^{\infty} w_{m,k}\, \psi_{m,k}(x), \quad m,k \in R,$$

где $c_{m,k}$, $w_{m,k}$ – аппроксимирующие и детализирующие вейвлет-коэффициенты (ВК);

m,k – параметры масштаба и сдвига в пространстве целых чисел R.

Такое представление вейвлет-модели изображения двумерного ВП I (матрица яркости пикселей n×n) примет вид

$$I(x,y) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{j,k}\varphi\varphi(x,y) + \sum_{i=0}^{\infty}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}(w_{1j,k}^{(i)}\varphi\psi_{j,k}^{(i)}(x,y) + w_{2j,k}^{(i)}\psi\varphi_{j,k}^{(i)}(x,y) + w_{3j,k}^{(i)}\psi\psi_{j,k}^{(i)}(x,y)),$$

Используя понятия теории фильтрации $I(x,y)$ представляет собой свертку с низкочастотным $g(n)$ и высокочастотным $h(n)$ фильтрами с прореживанием результата вдвое. Тогда расчет низкочастотных c_i и высокочастотных w_i коэффициентов дискретного вейвлет-преобразования (ДВП) изображения можно реализовать по зависимостям:

$$c_i = \sum_{k=1}^{n} I_{2i+k} \cdot g_k \,, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$w_i = \sum_{k=1}^{n} I_{2i+k} \cdot h_k \,.$$

Расчет КВП обеспечивают выполнение быстрого ДВП одномерного числового ряда на основе пирамидального алгоритма вычисления вейвлет-коэффициентов (алгоритм Маллата), приведенного на рисунке 1.3.

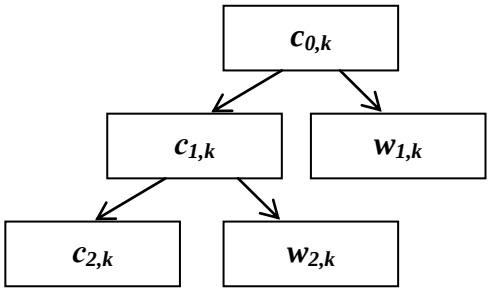


Рисунок 1.3 – Алгоритм Маллата

Сущность операций алгоритма Маллата заключается в следующем. С учетом коэффициентов h_n и g_n (рисунок 1.3), на первом этапе преобразования первый цифровой фильтр h_n из числового ряда $f_k = c_{0,k}$ выделяет низкие частоты $|\omega| \le \pi/2$, а другой (октавный) фильтр g_n выделяет верхние частоты $\pi/2 \le |\omega| \le \pi$. Поскольку на выходе фильтра h_n отсутствует верхняя половина частот, то частота дискретизации выходного изображения может быть уменьшена в 2 раза, т.е. выполнена децимация выходного массива, что и производится сдвигами $(2k+n)$ через 2 отсчета по входному массиву. Соответственно, на выходе фильтра g_n освобождается место в области низких частот, и аналогичное прореживание выходного числового массива приводит к транспонированию верхних частот на освободившееся место. Следовательно, каждый из выходных числовых массивов несет информацию о своей половине частот, при этом выходная информация представлена таким же количеством отсчетов, что и входная.

При таком описании процессов ДВП изображений особый интерес представляет вейвлет Хаара, так как в ДВП соответствующие низкочастотный (аппроксимирующих) и высокочастотный (детализирующих) фильтры состоят из коэффициентов $g = \left[\frac{1}{\sqrt{2}} \quad \frac{1}{\sqrt{2}}\right]$, $h = \left[\frac{1}{\sqrt{2}} \quad -\frac{1}{\sqrt{2}}\right]$. Отсюда, вычисление низкочастотных и высокочастотных вейвлет-коэффициентов реализуется в виде:

$$c_i = \frac{1}{\sqrt{2}}(I_{2i} + I_{2i+1}),$$

$$w_i = \frac{1}{\sqrt{2}}(I_{2i} - I_{2i+1}),$$

Поэтому в качестве базисного вейвлета в задачах идентификации и распознавания поверхностных дефектов принимается вейвлет Хаара.

При вычислении двумерного ДВП изображения сначала выполняется преобразование по строкам, а затем по столбцам [2,13]. Для преобразования каждого следующего уровня применяется схема разложения для матрицы $W_{LL.}$ (см. рис. 1.2)

Значения элементов матрицы яркости ***I*** лежат в диапазоне от 0 до 255. Тогда при ВП одного уровня разложения изображения максимальные значения высокочастотных и низкочастотных коэффициентов могут возрасти в 4 раза (при суммировании – в 2 раза при горизонтальном преобразовании и в 2 раза при вертикальном). Соответственно на 3-ем уровне преобразования максимальные значения могут возрасти в 2^8=256 раз. Поэтому для расчета целочисленных значений коэффициентов достаточно использовать двухбайтные числа.

Современные процессоры позволяют использовать команды для оперирования векторами по 128 бит, то есть по 8 двухбайтных чисел.

Тогда ВП по строкам матрицы яркости I изображения вида

$$I = \begin{bmatrix} I_{1,1} & I_{1,2} & \dots & I_{1,n} \\ I_{2,1} & I_{2,2} & \dots & I_{2,n} \\ \dots & \dots & \dots & \dots \\ I_{n,1} & I_{n,2} & \dots & I_{n,n} \end{bmatrix},$$

можно представить в виде матрицы пакетов пикселей

$$I = \begin{bmatrix} V_{1,1..8} & V_{1,9..16} & \dots & V_{1,n-7..n} \\ V_{2,1..8} & V_{1,9..16} & \dots & V_{2,n-7..n} \\ \dots & \dots & \dots & \dots \\ V_{n,9..16} & V_{1,9..16} & \dots & V_{n,n-7..n} \end{bmatrix}, \text{ где } V_{i,j..k} = [I_{i,j}, I_{i,j+1,..I_{i,k}}].$$

Отсюда расчет аппроксимирующих и детализирующих вейвлет-коэффициентов по строкам можно представить в виде

$$c_{i,\,j..j+7} = ((V_{i,2j..2j+7} + (V_{i,2j..2j+7} << 1))\downarrow) +$$
$$+ ((V_{i,2j+8..2j+15} + (V_{i,2j+8..2j+15} << 1))\downarrow >> 4)\text{'}$$

$$w_{i,\,j..j+7} = ((V_{i,2j..2j+7} - (V_{i,2j..2j+7} << 1))\downarrow) +$$
$$+ ((V_{i,2j+8..2j+15} - (V_{i,2j+8..2j+15} << 1))\downarrow >> 4)\text{'}$$

где >>, << – операции сдвига элементов вектора вправо и влево соответственно на $>> 4$ и $<<1$ элементов;

$\downarrow$ - обозначает операцию прореживания вдвое:

$V_{i,\,j..j+7}\downarrow = [V_{i,\,j}, V_{i,\,j+2}, V_{i,\,j+4}, V_{i,\,j+6}, 0,0,0,0]$ – результат прореживания.

Переход к операциям сложения, сдвига и прореживания при расчете вейвлет-коэффициентов приведет к значительному сокращению вычислительных операций при идентификации дефектов.

Матрицу ВП, полученную после преобразования по строкам, можно представить в виде матрицы векторов. Тогда расчет аппроксимирующих и детализирующих ВК ВП изображений по столбцам будет реализован в виде

$$c_{i..k,\,j} = V_{2i,\,j..k} + V_{2i+1,\,j..k},$$

$$w_{i..k,\,j} = V_{2i,\,j..k} - V_{2i+1,\,j..k},$$

Таким образом, получила развитие (модифицирована) вейвлет-модель изображения поверхностности проката, отличающаяся от известных представлением изображения в виде **пакетов** пикселей для использования **векторных команд** процессора и переходом к **операциям сложения, сдвига и прореживания** при расчете вейвлет-коэффициентов для сокращения вычислительных операций идентификации дефектов.

Вопросы для самоконтроля

1. От чего зависит качество цифровой обработки изображения?
2. Изображение может быть представлено функцией … аргументов.
3. Пояснить выбор метода цифровой обработки изображения.
4. Методы спектрального анализа изображений.
5. Пояснить смысл прямого преобразования Фурье.
6. Для чего используется обратное преобразования Фурье?
7. Смысл вейвлет-преобразования изображений?
8. Перечислите основные вероятностные модели изображений.
9. Физический смысл функции автокорреляции?

Глава 2 Моделирование изображений поверхностных дефектов для задачи распознавания

2.1 Систематизация моделей представления распознаваемого объекта

Модели представления распознаваемого объекта зависят от методов распознавания [6,7]. Можно выделить следующие основные методы распознавания образов для задачи распознавания поверхностных дефектов:

Списки признаков. Методы данной категории основаны на возможность распознавания объектов на основе некоторых характерных признаков данных объектов. Существуют два основных подхода. Первый подход основан на предположении, что простые измерения, проводимые над изображением, есть результат действия совокупности небольшого числа порождающих признаков. В этом случае разработка метода распознавания сводится к определению пространства признаков, которые необходимо измерить. Пространство признаков в этом случае имеет меньшую размерность. При данном подходе используются методы факторного анализа.

Второй подход определяет признаки как подмножества множеств простых измерений. При распознавании бинарных изображений такими признаками могут служить число черных точек на характеристической линии, наличие

черных точек в определённых областях т.д. Распознаваемые объекты представляются как различные совокупности наблюдаемых признаков.

Данный класс моделей использует в основном эвристические методы. Эффективность данных методов определяется правильностью выбора анализируемых признаков.

Структурное описание основывается на представлении объектов в виде совокупности «непроизводных элементов» и отношений между ними. Под непроизводными элементами понимаются фрагменты распознаваемых образов, которые формируют эти образы и являются простыми по собственной структуре, т.е. не содержат других непроизводных элементов, сколь-нибудь значимых для описания образа.

Как правило, системы, использующие структурное описание объектов, реализуют последовательную процедуру распознавания. В процессе решения информационной задачи обрабатывается входной образ, обходя его структуру элемент за элементом. Процесс распознавания делится на два потока: выделение структурных элементов определенного вида в образе и согласование получаемой информации о структуре с моделями для классов изображений, имеющимися в системе.

Одним из видов структурных методов являются методы грамматической классификации образов. Представление образов осуществляется в виде предложений специального языка. При этом требуется определить виды возможных структурных элементов изображений на этапе инициализации алгоритма. Построение правила классификации сводится к выводу грамматики, описывающей язык классифицируемых образов. Распознавание заключается в определении возможности вывода рассматриваемого предложения с помощью определенной грамматики.

Таким образом, наиболее приемлемым для задачи распознавания поверхностных дефектов признается структурное описание признакового пространство объектов распознавания.

2.2 Модель изображения в виде расширенного вектора признаков

Для распознавания дефектов структурными методами используется модель в виде вектора признаков. Для получения вектора признаков над первичным представлением распознаваемого образа производится серия вычислений, определяющих необходимые для классификации показатели.

Каждый такой показатель, называемой признаком, ставится в соответствие порядковый номер. Таким образом, признаки образуют вектор

$$u = (u_1, u_2, ..., u_n),$$

где n – число признаков

u_i – значение i-го признака, $i \in \overline{1,n}$.

Для решения задачи распознавания дефектов предлагается в дополнении к спектральным признакам, извлекаемым из вейвлет-модели детализирующих ВК области дефекта (ОД) w_Z^i, где i – уровень разложения, $1 \le i \le 3$, z – направление преобразования, $Z \in \{LH, HL, HH\}$, сформировать оптические и геометрические признаки.

Для расчета геометрических признаков $P_{def}, S_{def}, q_{def}, m, \sigma$ формируется матрица ОД *Map*, в которой пикселям, принадлежащим ОД, соответствуют значения равные единице, а пикселям, относящимся к текстурной области – значения, равные нулю.

Для определения оптических признаков $P_{def}, S_{def}, q_{def}, m, \sigma$ необходимо использовать матрицу яркости *I* (исходная матрица изображения).

После извлечения признаков модель изображения для задачи распознавания принимает вид вектора:

$$u = \begin{pmatrix} P_{def}, S_{def}, q_{def}, m, \sigma, mw_{HL}^1, mw_{LH}^1, mw_{HH}^1, mw_{HL}^2, mw_{LH}^2, mw_{HH}^2, mw_{HL}^3, \\ mw_{LH}^3, mw_{HH}^3, \sigma w_{HL}^1, \sigma w_{LH}^1, \sigma w_{HH}^1, \sigma w_{HL}^2, \sigma w_{LH}^2, \sigma w_{HH}^2, \sigma w_{HL}^3, \sigma w_{LH}^3, \sigma w_{HH}^3 \end{pmatrix}$$

Таким образом, для решения задачи распознавания предложено использовать модель изображения поверхностного дефекта в виде вектора признаков, отличающуюся расширенным информационным полем классификации дефектов на основе спектральных, пространственных и оптических метрик признаков, обеспечивающих повышение достоверности распознавания класса дефекта.

Вопросы для самоконтроля

1. Что представляет собой модель изображения в системах компьютерного зрения?
2. Перечислить основные классы моделей для описания изображений.
3. Пояснить принцип спектральных преобразований Фурье.
4. Физический смысл спектрального описания оцифрованного изображения?
5. Перечислить основные методы описания спектральных моделей изображения.
6. Принципы построения вероятностных моделей изображения?
7. Что отражают гармоники Фурье – пребразований?
8. Понятия вейвлет-анализа изображений.
9. Признаки (аргументы) вейвлет-модели изображения.
10. Расширенный вектор признаков модели изображения для задачи распознавания структурных элементов.

Глава 3 Методический аппарат идентификации и распознавания структурных элементов изображения

3.1 Концепция построения системы компьютерного зрения для задач идентификации и распознавания

Для использования моделей представления изображений необходимо разработать серию методик их применения в задачах идентификации и распознавания.

На рисунке 3.1 представлена реализация концепции СКЗ идентификации и распознавания структурного элемента – поверхностного дефекта листового металлопроката на основе описанных моделей изображения [6].

Рисунок 3.1 – Реализация концепции СКЗ идентификации и распознавания поверхностных дефектов тонколистового проката

3.2 Методика формирования изображения

Для формирования изображения в СКЗ можно использовать линейную камеру *Basler web ral2048* (рисунок 3.2), обладающая следующими характеристиками: количество пикселей в строке ***c = 1024***, частота съемки ***f = 30 кГц.*** Стандартная ширина прокатного листа составляет w=600 мм, погрешность Δw = 10 мм. [6].

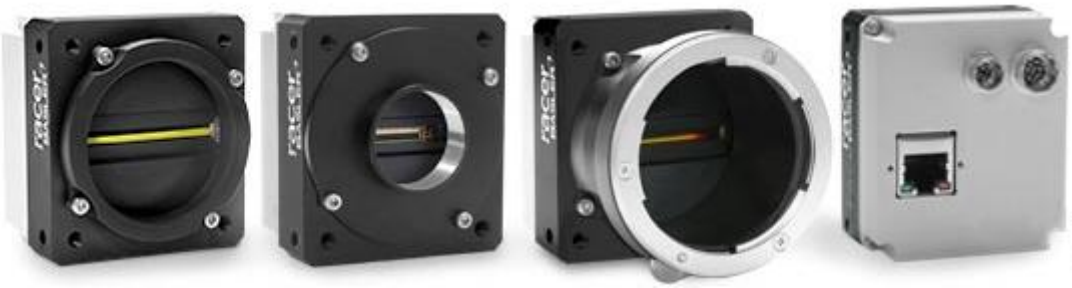


Рисунок 3.2 – Камера Basler web ral2048

Камера развернута таким образом, чтобы охватывать всю ширину листа проката. Для охвата всей площади поверхности на прокатном стане устанавливаются две камеры – фиксирующие верхнюю и нижнюю стороны прокатного листа соответственно. Схема размещения камер представлена на рисунке 3.3

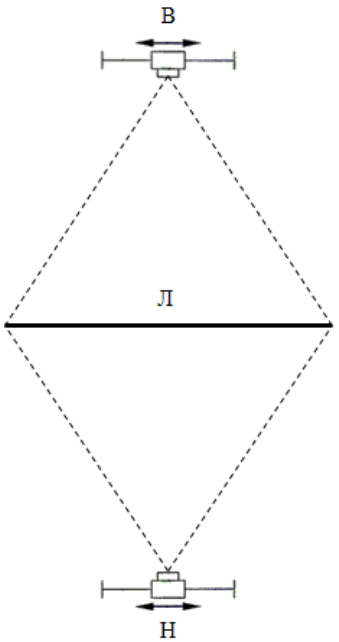


Рисунок 3.3 – Схема размещения камер (Л – лист проката, В и Н – верхняя и нижняя камеры)

Тогда стороне одного пикселя соответствует размер:

$$l_{pixw} = \frac{b}{c}, \tag{3.1}$$

где b - ширина прокатного листа;

c- число пикселей в строке.

Тогда для продукции тонколистового проката $l_{pixw} = \frac{600}{1024} = 0{,}586$ мм.

Для оценки размеров пикселя в направлении движения листа необходимо учитывать скорость движения листа. При движении листа со скоростью ***v*** смещение листа за время съемки одного кадра примет вид:

$$dl_{pixl} = \frac{v}{f} \ . \tag{3.2}$$

где ***f*** - число пикселей в строке.

Если *v* = 8 м/с = 8000 мм/с , то $dl_{pixl} = \frac{8000}{30000} = 0{,}267$ мм.

Учитывая, что длина, захватываемая одним пикселем l_{pixw}=0,586 мм, в 2 раза больше, чем смещение листа dl_{pixl} = 0,267 мм (рисунок 2.4), при формировании изображении можно пропускать ***k*** строк:

$$\mathrm{k} = \left\lfloor \frac{l_{pixw}}{dl_{pixl}} - 1 \right\rfloor \tag{3.3}$$

где $\mathrm{k}=\left\lfloor \frac{0{,}586}{0{,}267} - 1 \right\rfloor = 1$ число пропускаемых строк пикселей, соответствующих рис. 3.4 ($\lfloor x \rfloor$ обозначает округление вниз до целого).

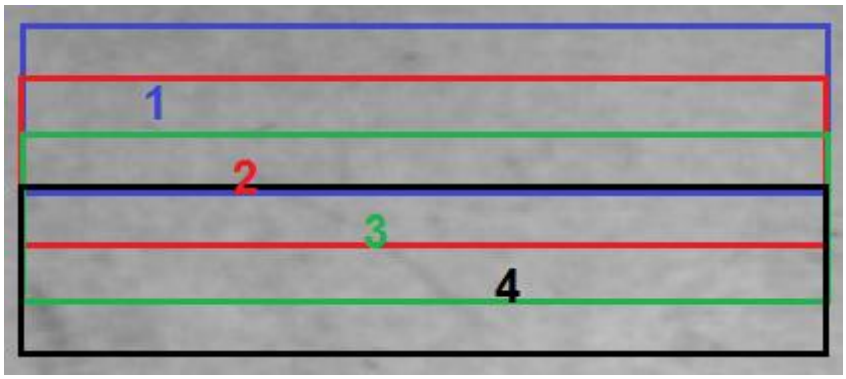


Рисунок 3.4 – Области проката, попадающие в 4 последовательных кадра

При таком способе формирования изображения, длина листа, соответствующая одному пикселю равна:

$$dl_{pixl}' = v \cdot t \cdot (k+1). \tag{3.4}$$

При этом $dl_{pixl}' \approx l_{pixl}$.

Очевидно, что количество пропускаемых кадров k должно быть неотрицательным. Это возможно только при условии $dl_{pixl} \le l_{pixl}$.

Из (3.1), (3.2) следует $v \cdot t \le \frac{b}{c}$.

Отсюда можно рассчитать максимально возможную скорость проката, при которой не будет потерь информации

$$v_{\cdot} \le \frac{b}{c \cdot t}. \tag{3.5}$$

Для используемой линейной камеры $v \le \frac{600 \cdot 30000}{1024} = 17578{,}125 мм/с \approx 17{,}6 м/с$.

Данное ограничение максимальной скорости соответствует требованиям к СКЗ распознавания дефектов. Следовательно, камера *Basler web ral2048* применима для распознавания поверхностных дефектов в требуемом временном диапазоне.

Для цифровой обработки используются кадры изображения размером 1024*1024 пикселя. Поэтому для формирования такого изображения необходимо накопить 1024 кадра линейной камеры. Укрупненный алгоритм формирования изображения представлен на рисунке 3.5 [6].

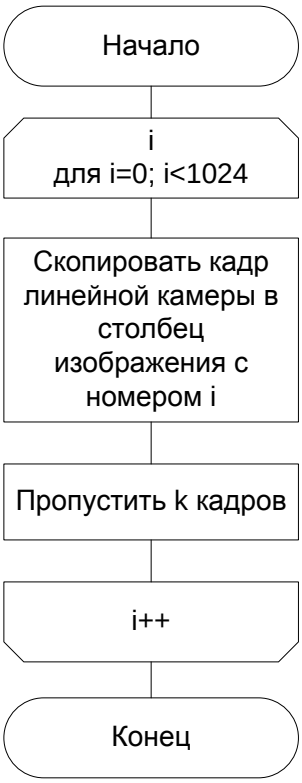


Рисунок 3.5 – Укрупненная схема алгоритма формирования изображения из кадров линейной камеры

При таком способе формирования кадра время обработки одного изображения из 1024 строк для скорости проката 8 м/с должно быть менее 0,0683 с. [6].

Таким образом, описана методика формирования изображения проката с линейной видеокамеры, отличающаяся пропуском части кадров, содержащих дублирующую информацию, что позволяет СКЗ идентифицировать дефекты на потенциально возможных скоростях проката.

Вопросы для самоконтроля

1. Что представляет собой структурный элемент изображения в СКЗ контроля качества листового металлопроката?

2. Назначение блока подготовки изображения.

3. Назначение блока идентификации дефекта.

4. Назначение блока расчета метрик.

5. Назначение анализатора класса дефектов.

6. Линейная видеокамера, принцип формирования изображения.

7. Принцип формирования кадра линейной камеры для цифровой обработка.

8. Характеристики линейной видеокамеры *Basler web ral2048*?

9. Сколь необходимо камер для контроля качества листового металлопроката?

10. Что необходимо учитывать для оценки размеров пикселя в направлении движения листа?

11. Можно ли пропускать часть кадров при цифровой обработки изображения металлопроката?

12. Какой формат кадра рекомендуется для цифровой обработки изображения?

Глава 4 Методический аппарат предварительной цифровой обработки изображений

На первом шаге цифровой обработки сформированное изображение проката подвергается двумерному вейвлет-разложению. Для идентификации поверхностных дефектов к каждой из детализирующих четвертей преобразования на каждом уровне разложения изображения целесообразно применять пороговый фильтр. Однако, вследствие влияния шумов от технологической смазки, коэффициенты вейвлет-разложения необходимо предварительно отфильтровать.

4.1 Систематизация алгоритмов фильтрации изображений

В цифровой обработке двумерных сигналов широко используется масочная фильтрация [6,9,11]. Масочная фильтрация является двумерной фильтрацией с конечной импульсной характеристикой (КИХ) фильтра. Маска представля-

ет собой множество весовых коэффициентов, которые заданы во всех точках окрестности ***S***, которые обычно симметрично окружают текущую точку кадра.

На практике чаще всего используется окрестность являющаяся квадратом 3×3, в центре которого находится текущий элемент. Возможно применение различных масок. Одним из эвристических вариантов является равномерная маска, для которой все весовые коэффициенты равны 1/9. В этом случае сохраняется средняя яркость изображения. Применение фильтрации существенно снижает уровня шума изображения.

Линейная пространственная фильтрация. Пространственная фильтрация двумерного сигнала ***f(x,y)*** позволяет применять фильтры с КИХ [6] и осуществляется как двумерная свертка импульсной характеристики фильтра ***h(s,t)*** с двумерным сигналом изображения ***f(x,y),*** где ***t*** – координата характеристики в вертикальном направлении вдоль оси ***y,*** для всех $t \in [-m/2, m/2]$, ***s*** – координата характеристики в горизонтальном направлении вдоль оси ***x,*** $s \in [-n/2, n/2]$:

$$g(x,y) = f(x,y) \otimes h(x,y) = \sum_{t=-m/2}^{m/2} \sum_{s=-n/2}^{n/2} f(s,t)h(x-s,y-t) = \sum_{t=-m/2}^{m/2} \sum_{s=-n/2}^{n/2} f(x-s,y-t)h(x,y)$$

Маской фильтра называется прямоугольная область размером ***m*** × ***n***, на которой задана импульсная характеристика.

Элементы импульсной характеристики фильтра и соответствующей области изображения для случая ***m=3, n=3*** представлены на рисунке 4.1.

	$s=-1$	$s=0$	$s=1$
$t=-1$	$f(-1,-1)$	$f(0,-1)$	$f(1,-1)$
$t=0$	$f(-1,0)$	$f(0,0)$	$f(1,0)$
$t=1$	$f(-1,1)$	$f(0,1)$	$f(1,1)$

$s=-1$	$s=0$	$s=1$
$h(1,1)$	$h(0,1)$	$h(-1,1)$
$h(1,0)$	$h(0,0)$	$h(-1,0)$
$h(1,-1)$	$h(0,-1)$	$h(-1,-1)$

Рисунок 4.1 – Положение отсчетов импульсной характеристики

Начало координат фильтра устанавливается в центр импульсной характеристики. Отклик фильтра ***g(x,y)*** рассчитывается как сумма произведений отсчетов изображения на соответствующие отсчеты повернутой импульсной характеристики. Данная операция выполняется для каждого отсчета изображения.

В случае, когда импульсная характеристика фильтра симметрична, вместо свертки можно использовать корреляционную оценку:

$$g(x,y) = f(x,y) \otimes h(x,y) = \sum_{t=-m/2}^{m/2} \sum_{s=-n/2}^{n/2} f(x+s, y+t) h(x,y) .$$

Данная операция представляет собой расчет в маске фильтра, скользящей по двумерному изображению, суммы произведений коэффициентов фильтра на соответствующие отсчеты двумерного изображения (рисунок 4.2).

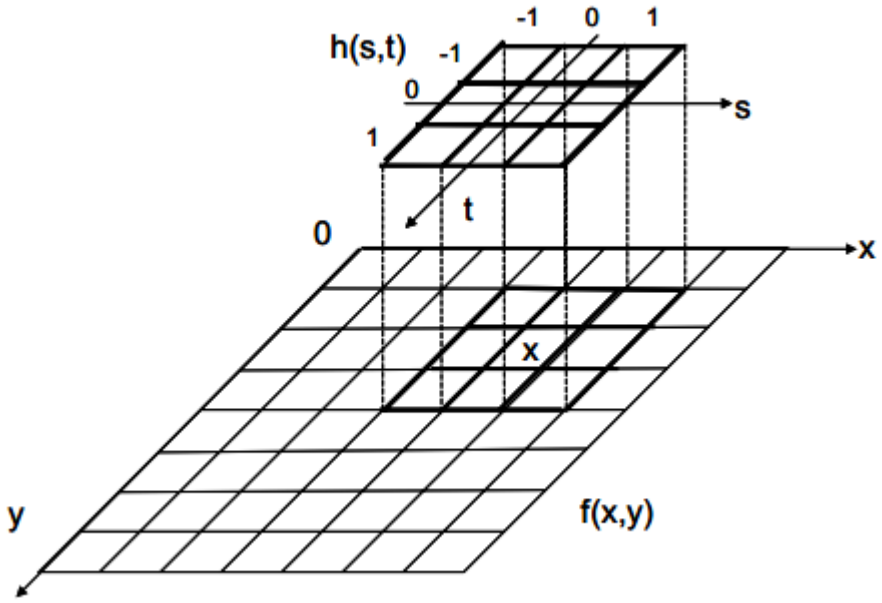


Рисунок 4.2 – Корреляционная оценка двумерного изображения ***f(x,y)*** с маской ***h(s,t)***

Нелинейная пространственная фильтрация. Сглаживающие линейные фильтры не только подавляют шумы, но также размывают границы между областями с разной амплитудой сигнала изображения [6]. Для уменьшения эффекта «смаза» границ используются нелинейные фильтры. Нелинейные КИХ фильтры, как и линейные фильтры, работают в скользящем окне. Но нелиней-

ные фильтры основаны на нелинейных преобразованиях отсчетов в окрестности элементов, определяемой маской фильтра.

Медианный фильтр. Медианный фильтр (МФ), предложенный Тьюки в 1974 г., заменяет центральный элемент маски медианой упорядоченной (по возрастанию или по убыванию амплитуды) выборки, сформированной из всех амплитуд отсчетов, покрываемых маской фильтра [6]. При медианной фильтрации используется скользящее двумерное окно. Для каждого отсчета выполняется оценка медианы в окне. Для ускорения оценки целесообразно использовать ранее выполненные вычисления алгоритмически на каждом шаге. Размер окна принимается нечетным и равным $m \times n$. Оказавшиеся в пределах окна отсчеты изображения образуют обрабатываемую выборку текущего отсчета. При упорядочивании последовательности $\{f_i,\ i=[1,mn]\}$ по возрастанию, ее медианой будет являться элемент выборки, занимающий центральное положение в этой упорядоченной последовательности. Этот элемент является результатом медианной фильтрации в текущей точке изображения. Формальное обозначение описанной процедуры $g_{med}=med(f_1,f_2,\ldots,f_n)$.

Таким образом, действие МФ состоит в игнорировании как отрицательных, так и положительных выбросов. Медианная фильтрация лучше сохраняет границы между областями двумерного сигнала, чем любая линейная фильтрация.

Импульсная помеха, размер которой меньше или равен $mn/2$, медианным фильтром подавляется, а резкие изменения амплитуды сохраняются.

Рассмотренные фильтры исследованы на эффективность подавления шумов вейвлет-коэффициентов [6]. Наилучшие результаты при этом получены с использованием медианного фильтра.

4.2 Методика фильтрации и бинаризации вейвлет-коэффициентов

Отклик медианного фильтра в окне для каждой четверти z детализирующих ВК на каждом уровне разложения вычисляется по формуле [6]:

$$w_{\Phi Z}^{i}(x,y) = med\{w_Z^i(x+m,y+n) : m=-1,0,1; n=-1,0,1\},$$

где w_z^i – матрицы ВК ***i***-м уровне, ***i=1,2,3***, до применения фильтра;

$w_{\Phi Z}^{\ i}$ – матрицы ВК после применения фильтра.

Для определения элементов матриц, лежащих на границе дефекта, предлагается выполнить бинаризацию по динамическому порогу.

Бинаризация осуществляется по принципу:

$$w_{БZ}^{\ i}(x,y)=\begin{cases}1, w_z^i(x,y)\geq treshold^i{}_Z\\ 0, w_z^i(x,y)<treshold^i{}_Z\end{cases},$$

с порогом значений вейвлет-коэффициентов

$$treshold^i{}_Z=3\sqrt{\frac{1}{nw}\sum_{x=1}^{nw}\sum_{y=1}^{nw}(w_{\Phi z}^{\ i}(x,y)-mw_{\Phi z}^{\ i})^2},$$

где $w_z^i(x,y)$ – КВП на уровне i в четверти z;

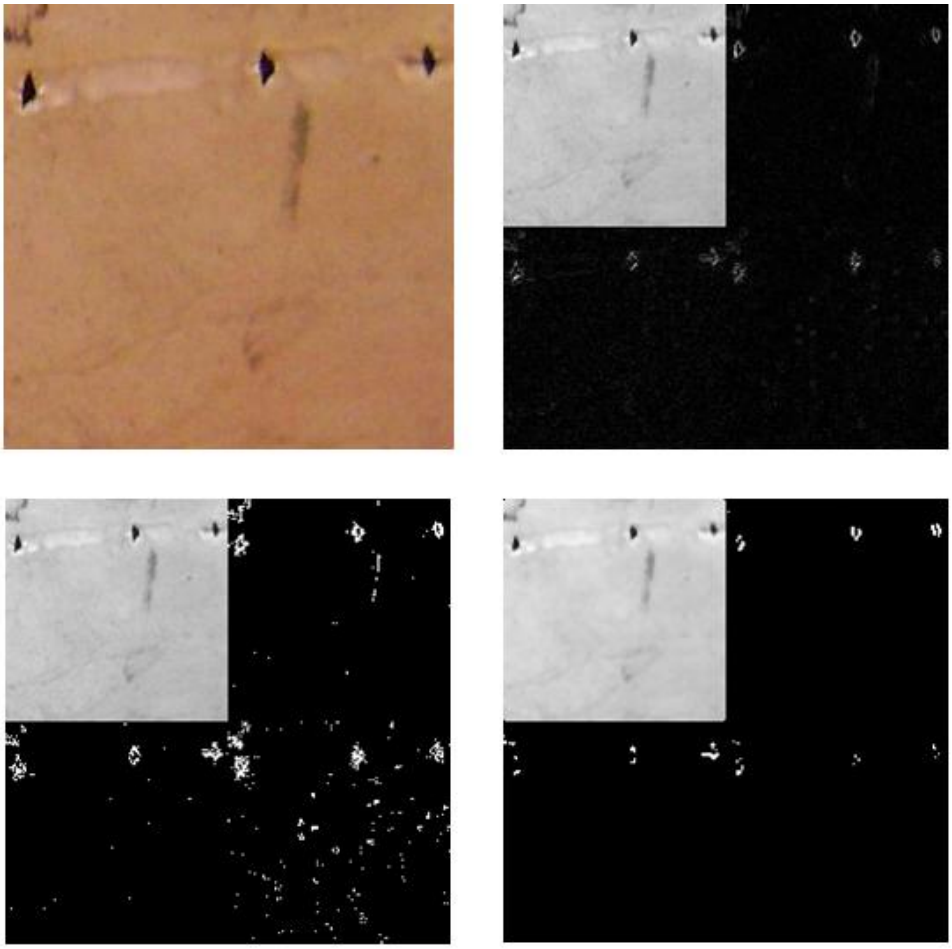
$mw_{\Phi z}^{\ i}$ – среднее значение КВП на уровне i в четверти z;

nw – размерность матриц $w_{БZ}^{\ i}$, $nw=n/2^i$;

n – размерность матрицы изображения, n = 1024.

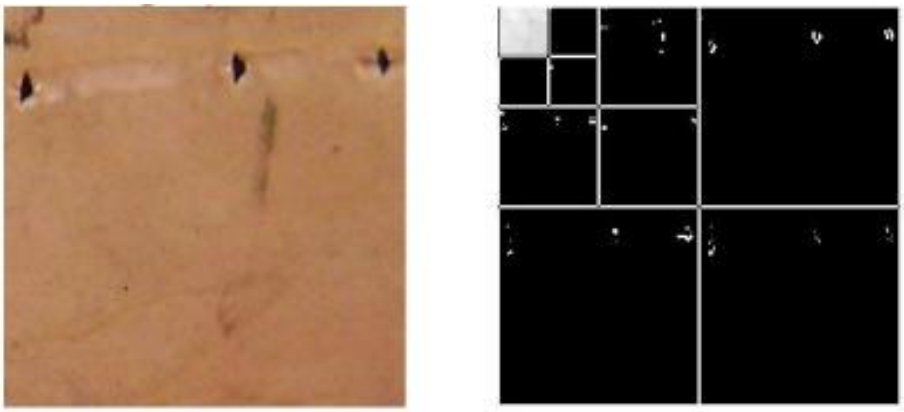
Результат идентификации поверхностного дефекта класса «Отверстие» представлен на рисунке 3.3 (а – исходное изображение, б – результат вейвлет-преобразования, в – бинаризация без фильтров, г – бинаризация с использованием медианного фильтра) [6].

Для определения минимального уровня разрешения для идентификации различных типов поверхностных дефектов исследована эффективность работы предложенного алгоритма ВР. Исследования проводились для неустранимого дефекта «Отверстие» и устранимого дефекта «Мятость». На рисунках 3.4, 3.5 представлены результаты работы алгоритма ВП (а – исходное изображение, б – результат ВП) [6].



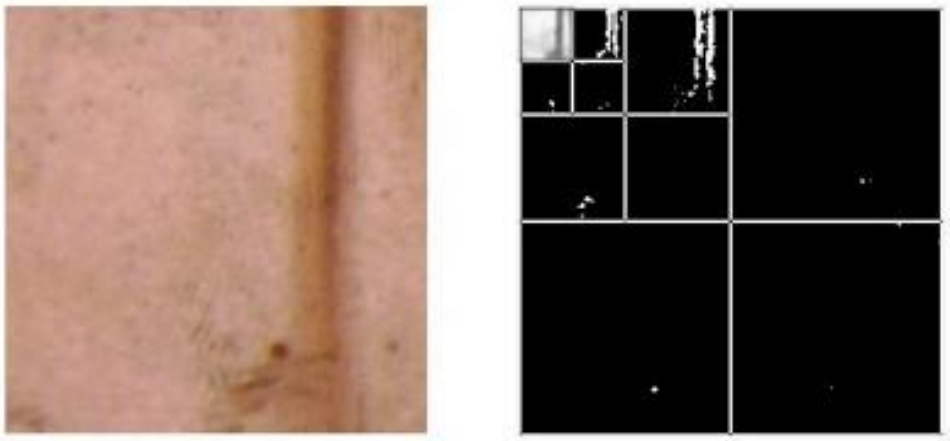
в г

Рисунок 3.3 – Идентификация поверхностных дефектов



а б

Рисунок 3.4 – Результат обнаружения для дефекта «Отверстие»



а б

Рисунок 3.5 – Результат обнаружения для дефекта «Мятость»

Анализ результатов показал, что для идентификации дефекта «Отверстие» достаточно одного уровня вейвлет-разложения изображения, а дефект «Мятость» необходимо три уровня преобразования.

Поэтому, для обнаружения поверхностных дефектов достаточно трех уровней преобразования исходного изображения, причем неустранимые дефекты выявляются после первого шага. На более высоких уровнях преобразования проявляются особенности текстурной компоненты изображения.

Таким образом, предложена методика идентификации поверхностных дефектов, отличающаяся от известных медианной фильтрацией ВК для устранения помех от технологической смазки с пороговой бинаризацией для определения числа уровней вейвлет-разложения изображения, позволяющие идентифицировать поверхностные дефекты листового проката.

Вопросы для самоконтроля

1. Что происходит на первом шаге цифровой обработки изображения в СКЗ контроля качества листового металлопроката?
2. С какой целью проводится предварительная обработка изображения?
3. Маска представляет собой ……
4. Что представляет собой линейная пространственная фильтрация?
5. Недостатки линейной пространственной фильтрации.

6. Сущность нелинейной пространственной фильтрации изображения.

7. Принцип медианной фильтрации при цифровой обработке изображения.

8. Достоинства и недостатки медианной фильтрации?

9. От чего зависит отклик медианного фильтра при вейвлет-разложении изображения?

10.Зачем выполняется бинаризация изображения?

11.Принцип бинаризации изображений.

12.Как определяется порог бинаризации?

13.На каком уровня вейвлет-разложения изображения идентифицируется дефект «Отверстие»?

14.На каком уровня вейвлет-разложения изображения идентифицируется дефект «Мятость»?

Глава 5 Методический аппарат идентификации поверхностных дефектов

Для определения количества дефектов в блоке идентификации (рис. 4.1) предлагается использовать области дефектов с целью повышения быстродействия алгоритмов идентификации (цифровой обработке подвергается не всё изображение, а только часть – область дефектов). Для устранения шумов изображения предложено использовать медианный фильтр, временные затраты на который остаются значительными. Поэтому необходимо предложить методику ускорения его работы.

5.1 Методика ускорения работы медианного фильтра

Предложенная методика вейвлет-разложения (ВР) изображений позволяет идентифицировать дефекты листового проката. Для изображения размером 1.5 мегапикселя время работы такого алгоритма на компьютере с конфигурацией Intel Core i7 3.0 ГГц, 4 GB RAM составляет в среднем 1.1 секунды, что не позволяет применять алгоритм для обработки видеоряда (25 изображений 1024x1024

пикселя в секунду) в реальном времени. Существенную часть времени при этом занимает работа медианного фильтра (1.06 с.). Поэтому необходимо ускорение алгоритма медианной фильтрации для обнаружения дефектов в реальном времени. Реализация медианного фильтра при помощи гистограмм позволяет сократить время работы медианного фильтра до 0.45 с. [6].

Для ускорения работы алгоритма фильтрации ВК предложен алгоритм обнаружения, состоящий из трех основных этапов: двумерное вейвлет-преобразование до третьего уровня, применение медианного фильтра к ВК и определение ВК, больших, чем утроенное их СКО. Если величина ВК больше половины количества элементов в окне фильтра, тогда медиана больше, чем утроенное СКО. Реализованный таким образом медианный фильтр срабатывает за 0,1 с., что является недостаточным для работы алгоритма в реальном времени.

Для дальнейшего ускорения работы медианного фильтра применяется распараллеливания алгоритма обнаружения в сочетании с использованием векторных команд процессора из набора SSE2. Технология SSE2 включает в архитектуру процессора восемь 128-битных регистров и набор инструкций, работающих с векторами значений.

На рисунке 5.1 представлен алгоритм фильтрации ВК при обнаружении дефектов с использованием векторных команд процессора [6].

Исследования проводились на компьютере с процессорами Intel Core i7 3.0 ГГц, 4 GB RAM для изображений 1.5 мегапикселя. Результаты работы различных вариантов реализаций фильтра приведены в таблице 5.1 [6].

Таблица 5.1 – Результаты работы различных вариантов фильтрации

Реализация	Результат, с
С использованием сортировки	1.04
С использование гистограммы	0.45
С использованием сравнения медианы с утроенным СКО	0.1
С использованием векторных команд процессора	0.017

Сочетание ВКП и распараллеливания на два ядра	0.01

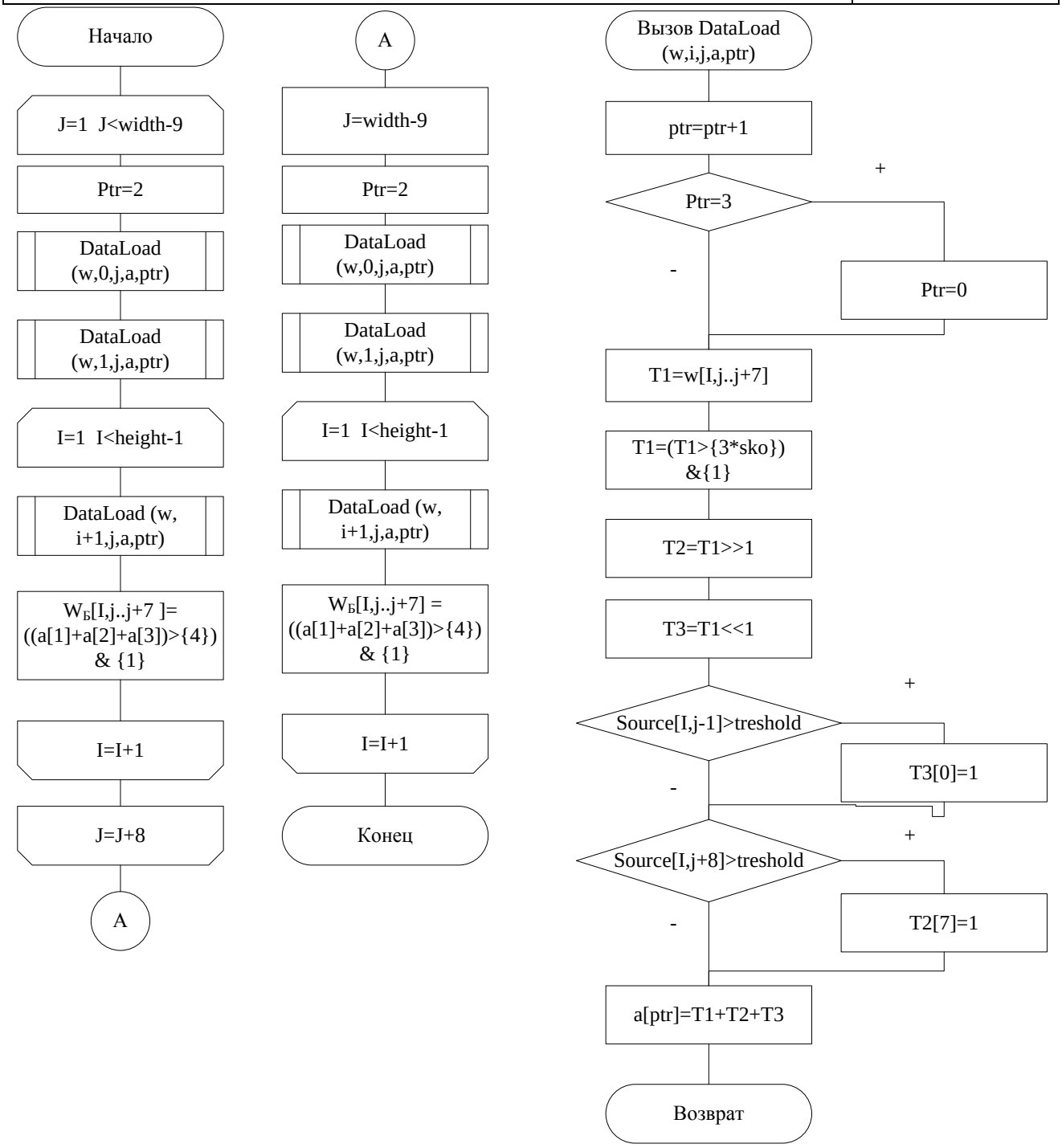


Рисунок 5.1 – Алгоритм фильтрации ВК при идентификации (обнаружении) дефектов на основе векторных команд процессора

Распараллеливание алгоритма медианной фильтрации на два ядра по технологии OpenMP позволяет осуществлять фильтрацию за 0.01с. [6], что приемлемо для обнаружение дефектов в реальном времени.

Таким образом, предложенные алгоритмы идентификации позволяют распараллелить часть цифровой обработки изображений и, в сочетании с использованием векторных команд процессора, обеспечить идентификацию поверхностных дефектов в процессе производства.

5.2 Методика выделение областей поверхностных дефектов

В полученной матрице $w_{БZ}^{i}$ единицы соответствуют границе области дефекта. Полученные таким образом области значительно меньше по отношению к исходному изображению.

Для того чтобы восстановить границы области дефекта в масштабе исходного изображения, необходимо выполнить операцию слияния [6,8]:

$$Map(x,y) = \bigcup_{i=1}^{3} \bigcup_{Z \in (LH,HL,HH)} w_{БZ}^{i}(\lfloor x / 2^{i} \rfloor, \lfloor y / 2^{i} \rfloor),$$

где $\lfloor \_ \rfloor$ – округление в меньшую сторону.

В результате формируется матрица, отражающая область дефекта, в которой границе области соответствуют значения равные 1. Результат операции слияния представлен на рисунке 5.2б.

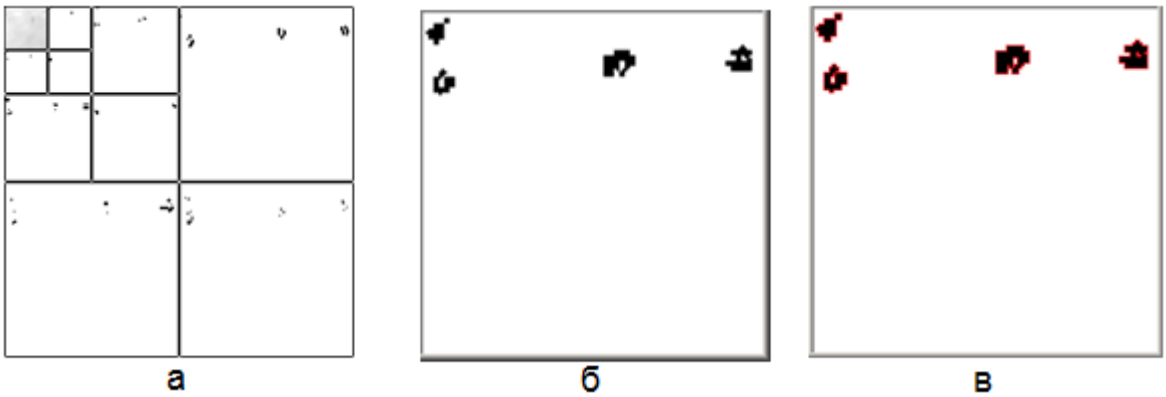


Рисунок 5.2 – Результаты операций бинаризации (а); слияния (б); утолщения границ (в)

Для объединения близкорасположенных частей границы дефекта в одну область используется операция утолщения границы по следующему правилу: если *Map(x,y)=0* и существует *Map(x+m,y+n)*=1; $m \in$ -1,0,1; $n \in$ -1,0,1, то *Map(x,y):*=2.

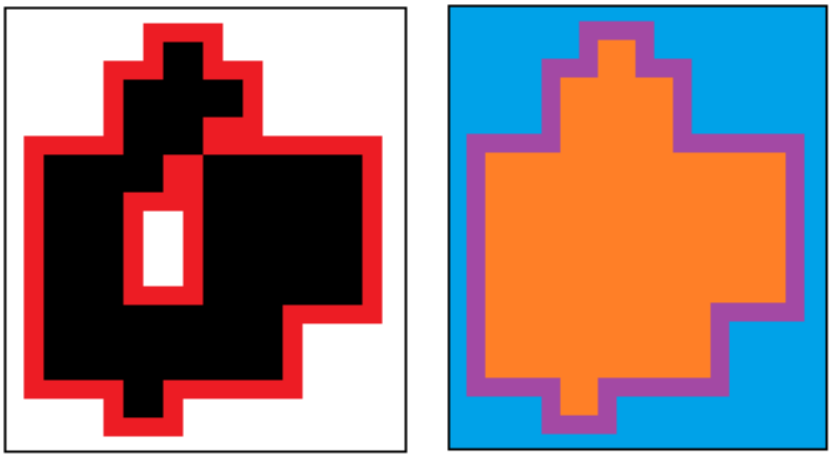
Результат операции утолщения границы представлен на рисунке 5.3в.

С целью ограничения анализируемой ОД (карта дефектов) выделяются отдельные области на изображении (рисунок 5.3а):

1 Находится Map(x,y)$\neq$0.

2 Начиная с этого элемента, выполняется поиск в ширину для копирования объекта в отдельную матрицу M. На исходной карте дефектов объект при этом заполняется нулевыми значениями (стирается). Матрица M, в которую путем копирования выделяется объект, создается таким образом, чтобы крайние значения в матрице были равны нулю.

3 Повторяются пункты 1–2, пока на карте дефектов существуют элементы Map(x,y) ≠ 0.



а б

Рисунок 5.3 – Локализованный дефект (а); выделение границы и внутренней области дефекта (б)

Далее отделяется текстурная компонента изображения от ОД. Для этого от крайнего пикселя *M(0,0)* выполняется обход в ширину для значений равных

0 и их значение переопределяются на 3. Отсюда, можно разграничить пиксели с нулевым значением, расположенные вне ОД и внутри ОД.

Последним этапом методики формирования ОД является разделение границы дефекта и внутренней ОД (рисунок 5.3б). Выделение границы осуществляется по следующему принципу: если $M(x,y) = 3$ и существует $M(x+m,y+n)=1$; m -1,0,1; n -1,0,1, то $M(x,y)$ принимается равным 4.

Тогда внутреннюю ОД можно выделить по принципу: если $M(x,y)$ 3 и $M(x,y)$ 4, то $M(x,y)$ принимается равным 5.

Предложенная методика формирования ОД позволяет исключить текстурную компоненту модели изображения $\boldsymbol{I}$ и определить пространственное положение дефекта [6].

Таким образом, предложена методика, позволяющая идентифицировать области дефектов по изображению и определить их пространственное положение.

Вопросы для самоконтроля

1. Как подсчитывается число идентифицированных поверхностных дефектов?
2. Принцип выполнения операции слияния?
3. Результат операции слияния?
4. Правило операции утолщения?
5. Результат операции утолщения?
6. Принцип формирования области дефектов (карты)?
7. Принцип отделения текстурной компоненты изображения от области дефекта?
8. Правило разделения границы дефекта и внутренней области дефекта?

Глава 6 Методический аппарат распознавания структурных объектов

Критичным фактором при выборе метода распознавания поверхностных дефектов листового проката является скорость выполнения алгоритма, так как распознавание должно производиться в процессе производства. Эффективное (в вычислительном отношении) использование распознающих алгоритмов пред-

полагает наличие некоторого аппарата оценки близости, не требующего сравнения распознаваемого объекта с каждым элементом обучающего набора. Одним из таких методов является метод окрестностей [6].

6.1 Методика использования метода окрестностей в задаче распознавания поверхностных дефектов

В методе окрестностей степень выраженности каждого признака дефекта (объекта изображения) должна иметь 2^m градаций, для чего выполняется отображение численных значений признаков на множество $\{0,1,\ldots,2^m-1\}$.

Отсюда, каждый распознаваемый объект $u=(u_1,\ldots,u_N)$ можно интерпретировать как точку в гиперпространстве $D=\{u\in R^N : -2^{-1}\le u_i\le 2^m-2^{-1}, 1\le i\le N\}$, называемом пространством признаков (предполагается, что D охватывает любые возможные в задаче сочетания значений признаков).

Распознавание, т.е. указание принадлежности объекта к одному из классов Ω_v, осуществляется на основе принципа прецедентности или частичной прецедентности, т.е. путем оценки близости объекта к элементам заданного обучающего набора объектов, принадлежность которых к классам дефектов известна.

Первым шагом методика распознавания является построение системы окрестностей в пространстве D. Для этого необходимо выполнить:

1. Вводится вспомогательный гиперкуб

$$D_0=\{u\in R^N : -2^{-1}\le u_i\le 2^{m+1}-2^{-1}, 1\le i\le N\},$$

и набор отражений

$$g_n(u)=u-h_n, h_n=(2^{n-1},\ldots,2^{n-1}),$$

порождающих набор образов

$$D_n=g_n(D_0),\ 1\le n\le m+1,$$

гиперкуба D_0 каждый из которых включает гиперкуб D, т.е.

$$D = D \cap D_n, 0 \le n \le m+1$$

2. Вспомогательный гиперкуб D_0 делится на 2^N гиперкубов «первого разбиения» гиперплоскостями, параллельными координатным и проходящими через серединные точки его ребер, ортогональных к этим гиперплоскостям. Каждый из гиперкубов первого разделения разбивается тем же приемом на 2^N гиперкубов второго разбиения. Последовательное разбиение продолжается до тех пор, пока не будут построены гиперкубы ***m+1*** разбиения с длинной ребра, равной 1.

Необходимо отметить, что общее число подкубов l-го разбиения равно $\boldsymbol{2^{lN}}$ и длины их ребер равны $\boldsymbol{2^{m+1-l}}$. Пример построенной системы окрестностей для случая ***m=2, N=3*** представлен на рисунке 6.1.

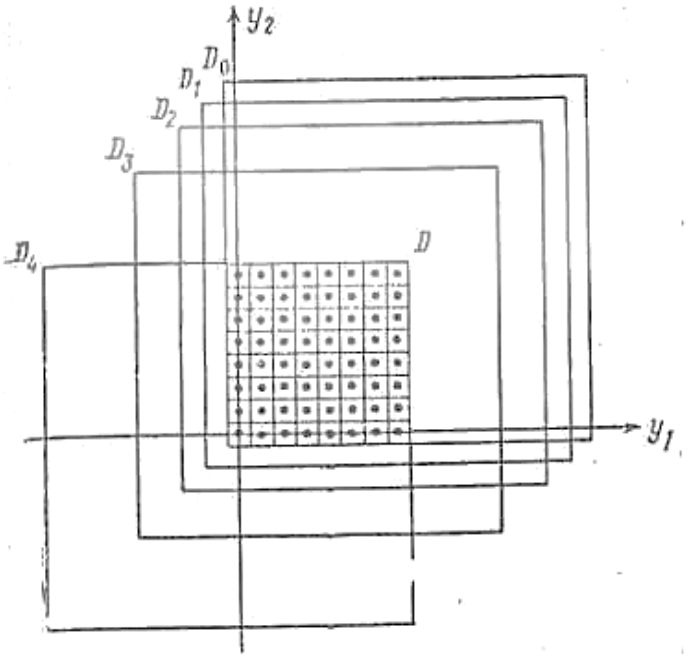


Рисунок 6.1 – Пример системы окрестностей для случая ***m=2, N=3***

6.2 Особенности индексации подкубов и окрестностей

Для указания любого подкуба l-го разбиения, где $1 \le l \le m+1$, в предлагается индексация подкубов на каждом уровне разбиения действительными числами ***x*** из интервала [0,1] на основе кривых Пеано.

Однако при большом количестве признаков и числе градаций количество окрестностей на последних уровнях разбиения будет большим, чем $\boldsymbol{2^{64}}$**.** Соот-

ветственно, в программной реализации необходимо будет использовать так называемые «длинные числа», реализуемые через массивы значений, что приводит к значительному увеличение процессорного времени, необходимого для вычисления ***x***.

В качестве альтернативы для индексации гиперкубов на каждом уровне разбиения можно использовать вектора значений $\boldsymbol{r=(r_1,\ldots,r_N)}$ [6]. Для этого необходимо выполнить деление отрезка $\boldsymbol{[-2^{-1},2^{m+1}-2^{-1}]}$ на $\boldsymbol{2^l}$ равных частей для каждого уровня разбиения ***l***. Тогда для каждого уровня разбиения ***l*** получится упорядоченное множество отрезков. Такой подход при поиске требует выполнить большее количества сравнений, но не требует дополнительных вычислений

$$p_l = \{[a_i, b_i], a_i = -2^{-1} + (i-1)\cdot 2^{(m+1)-l},$$
$$b_i = -2^{-1} + (i)\cdot 2^{(m+1)-l}, 1 \le i \le 2^l\}.$$

В этом случае каждое из ребер любого из подкубов ***l***-го разбиения соответствует одному из отрезков $\boldsymbol{p_l}$. Вектор значений ***r*** для уровня разбиения ***l*** формируется по следующему принципу: если ребро гиперкуба, лежащее на оси координат признака $\boldsymbol{y_j}$ соответствует отрезку $[a_i,b_i]$, тогда $\boldsymbol{r_j = i}$. Таким образом, подкуб l-го разбиения, соответствующий вектору r, будет обозначаться $D_0(l,r)$. Пример индексации окрестности представлен на рисунке 6.2.

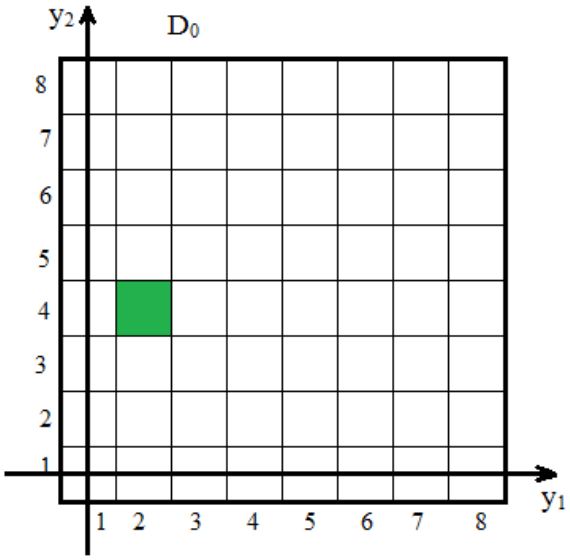


Рисунок 6.2 – Окрестность [2,4].

Отображения g_n, порождающие сдвинутые образы D_n из гиперкуба D_0, одновременно порождают и образы $D_n(l,z) = g_n(D_0(l,r))$ введенных элементов разбиения этого гиперкуба (рисунок 6.3). Принимается, что $D_0(0,0) = D_0, g_0(u) = u$ и вводится система окрестностей в пространстве D как набор непустых пересечений $D(l,r,n) = D \cap D_n(l,r)$, где $0 \le l \le m+1$, $0 \le n \le m+1$.

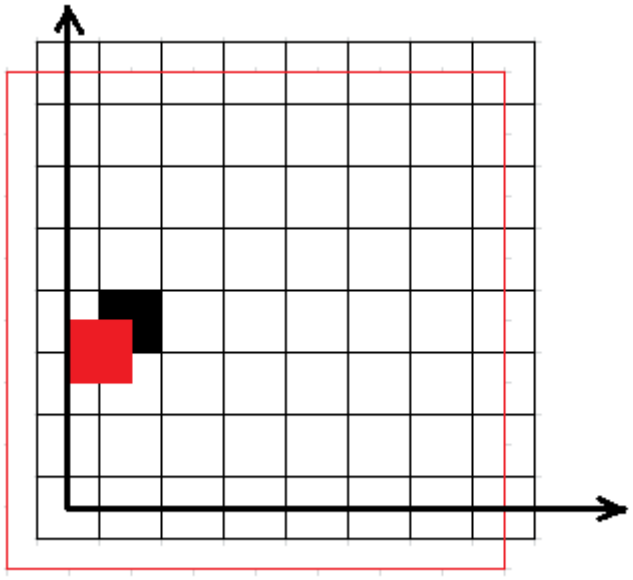


Рисунок 6.3 – Отображение окрестностей

Данная система окрестностей построена таким образом, что для любых двух точек x и y, для которых $\max(|x_i - y_i|) \le 2^{m+1-l}$ на l-м разбиении существует хотя бы одна окрестность $D(l,z,n)$, содержащая точки x и y.

Полученная система окрестностей позволяет для любого распознаваемого объекта $u = (u_1,...,u_n)$ и уровня l определить соответствующую ему окрестность $D(l,r,0)$. Для однозначного определения этой окрестности необходимо найти вектор z. Для этого в p_l осуществляется поиск отрезков $[a_i,b_i]$, $1 \le i \le 2^l$, такого, что $u_j \in [a_i, b_i]$ для каждого u_j. Вследствие того, что множество p_l является упорядоченным, нахождение соответствующего интервала можно вы-

полнить при помощи бинарного поиска. При этом вектор ***r*** формируется на основе правила: если $u_j \in [a_i, b_i]$, то $\boldsymbol{r_j = i}$.

Для определения соответствующих окрестностей $D(l,r,n)$, $1 \le n \le m+1$, необходимо выполнить отображение $\boldsymbol{g_n}$ для множества признаков $\boldsymbol{p_l}$

$$g_n(p_l) = \{[g_n(a_i), g_n(b_i), 1 \le i \le 2^l\}.$$

После этого вектор ***z*** формируется аналогично случаю $\boldsymbol{n = 0}$, за исключением того, что поиск отрезков осуществляется в отображении $\boldsymbol{g_n(p_l)}$.

Таким образом, предложена система окрестностей в пространстве признаков для однозначной классификации дефектов в виде вектора и разработана методика определения классификационного вектора для распознаваемого объекта на каждом уровне разбиения в каждом из отображений вспомогательного гиперкуба.

Вопросы для самоконтроля

1. Какой фактор критичен для распознавания дефектов?
2. Какой аппарата оценки близости распознаваемых объектов предпочтителен?
3. Как определяется степень выраженности каждого признака дефекта?
4. Что понимается под пространством признаков распознавания?
5. Пояснить принцип прецедентности?
6. Правило построения системы окрестностей в пространстве D.
7. Длина ребра границы разбиения гиперкуба?
8. Чему равно общее число подкубов *l*-го разбиения и длина их ребер?
9. Принцип индексации подкубов на каждом уровне разбиения?
10. Для индексации гиперкубов на каждом уровне разбиения можно использовать вектора значений, пояснить принцип.
11. Принцип использования бинарного поиска?

Глава 7 Алгоритмизация процессов обучения и распознавания анализатора класса поверхностных дефектов

7.1 Нормирование признаков дефектов

Для использования метода окрестностей в задаче распознавания поверхностных дефектов необходимо признаки дефектов нормализовать. Для каждого из изображений обучающей выборки рассчитываются величины признаков и определено их минимальное и максимальное значения. Для каждого из признаков отрезок *[min,max]* разделен на *2m* равных отрезков $[a_i, b_i]$, $0 \le i \le 2^m - 1$. Если соответствующее значение признака попадает в отрезок $[a_i, b_i]$, то его нормированное значение принимается равным i. Если значение признака меньше определенного минимального значения, то его нормированное значение принимается равным нулю. Если значение признака больше определенного максимального значения, то его нормированное значение принимается равным $2m-1$. Таким образом, выполняется отображение численных значений признаков на множество $\{0,1,\ldots,2m-1\}$.

Таким, образом, предложена методика отображения значений признаков на множество $\{0,1,\ldots,2^m-1\}$, что позволяет использовать метод окрестностей, адаптированный для распознавания поверхностных дефектов.

7.2 Алгоритмизация обучения анализатора класса

Разработана методика обучения анализатора класса дефектов на основе метода окрестностей, которая сводится к построению на основе набора образцов U для каждого уровня разбиения l, $0 \le l \le m+1$ и номера отображения n, $0 \le n \le m+1$ списка *List*(*l*,*n*), содержащего объекты $r_i \Omega_\nu$, упорядоченные по векторам r. При этом для векторов r^1 и r^2 принимается, что $r^1 = r^2$, если $r_i^1 = r_i^2, 1 \le i \le N$ и $r_i^1 < r_i^2$, если $\exists i : r_i^1 < r_i^2, 1 \le i \le N, \forall j < i, r_i^1 = r_i^2$.. Матрица списков List представлена в таблице 6.1.

Таблица 7.1 – Матрица списков List

	Уровень разбиения, l			
Отображение гипер-куба, n	1	2	…	m+1
1	$List_{1,1}$	$List_{1,2}$	…	$List_{1,m+1}$
2	$List_{2,1}$	$List_{2,2}$	…	$List_{2,m+1}$
…	…	…	…	…
m+1	$List_{m+1,1}$	$List_{m+1,2}$	…	$List_{m+1,m+1}$

Схема алгоритма обучения анализатора класса поверхностных дефектов представлена на рисунке 7.1.

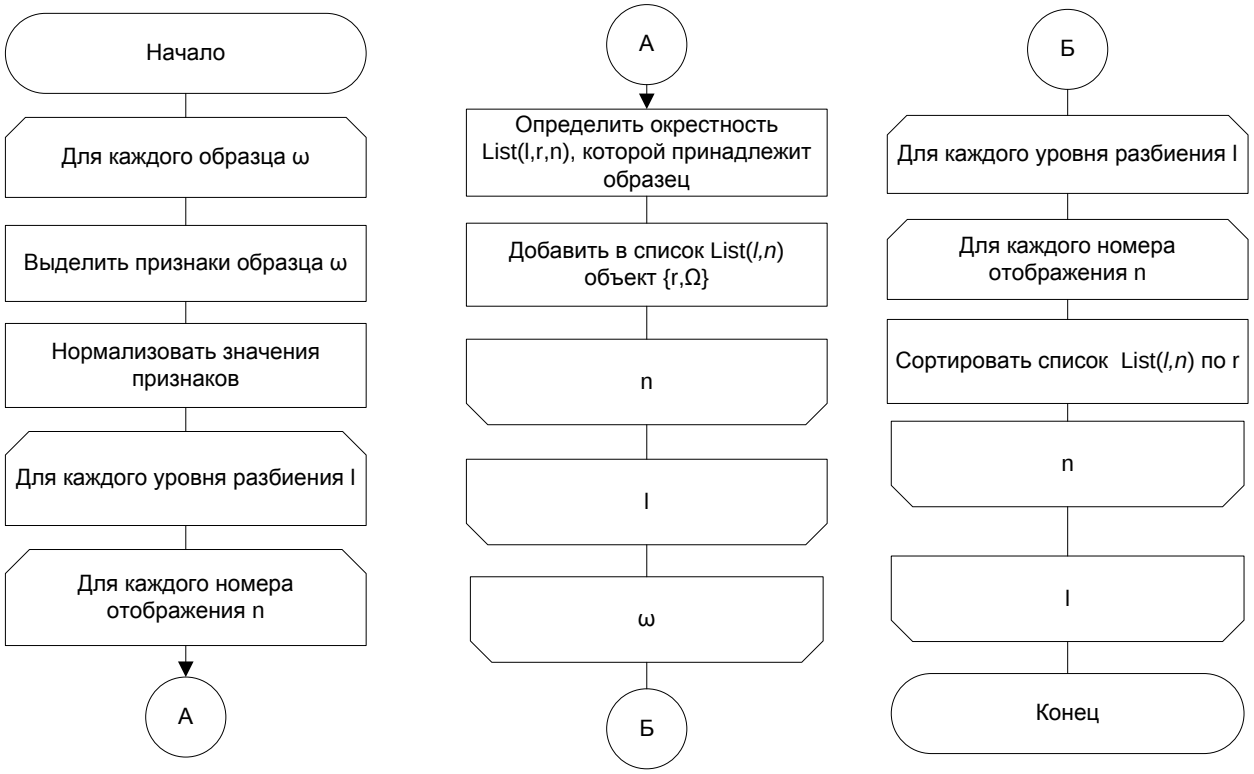


Рисунок 7.1 – Схема алгоритма обучения анализатора класс

Таким образом, разработан алгоритм обучения анализатора класс поверхностных дефектов на основе использования адаптированного метода окрестностей.

7.3 Алгоритмизация распознавания анализатора класса

Степень близости $L(u, u^j)$ объектов ***u*** и ***u^j^*** будет характеризоваться максимальным значением номера разбиения ***l***, $1 \le l \le m+1$, при котором существует некоторая окрестность ***D(l,z,n),*** охватывающая узлы соответствующие этим объектам, т.е. $L(u,u^j) = \max \{ r, n \mid u, u^j \in D(l,r,n) \}$.

Тогда для распознавания объекта необходимо [6]:

1) Найти максимальный уровень разбиения ***l***, для которого существуют окрестности ***D(l,r,n),*** охватывающие узлы, соответствующие распознаваемому объекту и не менее чем ***s*** образцов (принято ***s = 3***). Данный уровень разбиения обозначается, как $\tau(u)$.

2) Определить долю образцов γ_ν, принадлежащих тем же окрестностям *D*(*l*,*r*,*n*), которым принадлежит и распознаваемый объект, и относящихся к классу объектов Ω_ν.

$$\gamma_\nu = \frac{\mu_\nu(u)}{\mu(u)},$$

где $\mu(u)$ – количество образцов, принадлежащих тем же окрестностям ***D(l,z,n),*** которым принадлежит и распознаваемый объект;

$\mu_\nu(u), 1 \le \nu \le q$ – количество образцов, принадлежащих тем же окрестностям ***D(l,r,n),*** которым принадлежит и распознаваемый объект, и относящихся к классу объектов Ω_ν.

3) Определить класс объекта на основе порогового решающего правила. В работе использовано следующее пороговое правило: объект ***ω*** относится к классу объектов Ω_ν, если $\tau(u) \ge m-2$ и $\gamma_\nu > 0.7$.

Укрупненная схема алгоритма распознавания поверхностных дефектов СКЗ представлена на рисунке 6.2.

Так как каждый из списков ***List(l,n)*** является упорядоченным, можно выполнять поиск образцов, принадлежащих соответствующим окрестностям, при помощи известного алгоритма бинарного поиска.

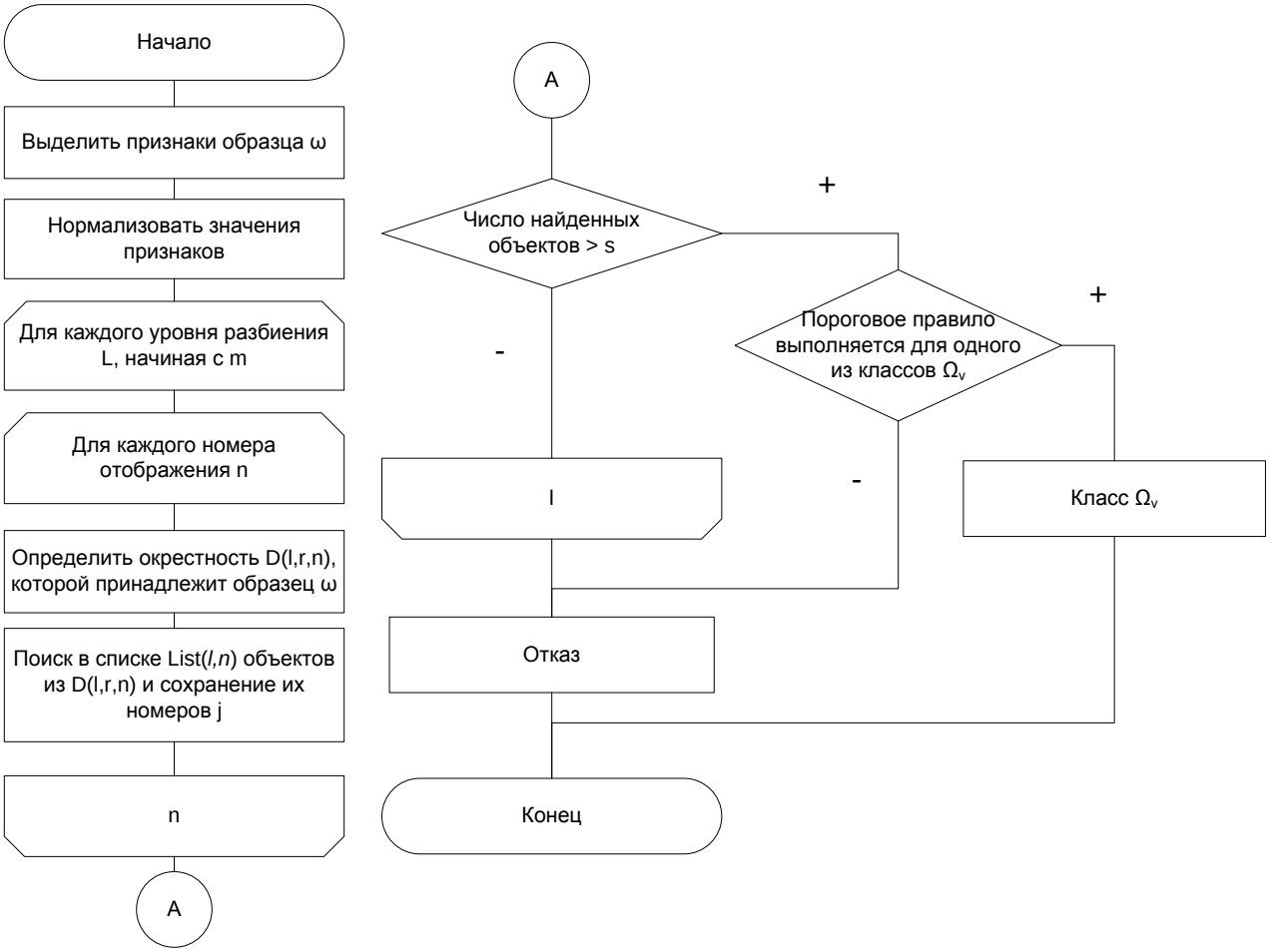


Рисунок 3.18 – Схема алгоритма распознавания дефектов

Таким образом, выделены параметры распознавания, определено пороговое решающее правило и разработан алгоритм распознавания поверхностных дефектов листового проката на основе метода окрестностей.

Вопросы для самоконтроля

1. Пояснить смысл нормализации признаков дефектов?
2. Что является основой методики обучения анализатора?
3. Как формируется набор образцов?
4. Цель построения матрицы списков List?

5. Пояснить алгоритм обучения анализатора класса поверхностных дефектов.

6. Чем характеризуется степень близости объектов распознавания?

7. Алгоритм распознавания объектов?

8. Как определить долю образцов, принадлежащих тем же окрестностям?

9. Как определить класс объекта на основе порогового решающего правила?

10. Смысл использования бинарного поиска в алгоритме распознавания?

Раздел 2 Исследование операций в задачах обеспечения информационной безопасности

Автоматизация процесса управления правилами доступа к ресурсам сети, а также его ключевой составляющей – средства поддержки принятия решений в условиях информационного противоборства, занимает центральное место в обеспечении информационной безопасности (ИБ) компьютерных корпоративных сетей (ККС) [13]. Принятие решений при реализации управленческих функций является важнейшей составной частью теории исследования операций и системного анализа [4,11], а сам процесс принятия решений характеризуется следующими основными элементами: объект управления, цель управления, лицо, принимающее решение, альтернативные варианты как средство достижения цели; внешние условия; исходы; правила выбора решений. Синтез алгоритмов выбора управляющих воздействий основывается на соответствующих математических моделях и методах.

Для разработки моделей и методов, составляющих научное обеспечение принятия решений, необходимо выбрать методический аппарат и формализовать понятие аномалий трафика ККС.

Глава 8 Методический аппарат выявления аномалий трафика корпоративной сети

Для решения задачи выявления аномалий трафика ККС предлагается использовать методы обработки трафика, характеризующего загрузку каналов передачи данных при отслеживающих и блокирующих работу сети информационных атаках (ИА). При этом объектами управления становятся средства разграничения доступа (маршрутизаторы, управляемые коммутаторы, файловые сервера FTP), а задача управления сводится к решению частной задаче управления трафиком сети на основе базы правил межсетевого экрана (МЭ) со средствами обнаружения вторжений (СОВ).

8.1 Систематизация моделей выявления аномалий компьютерной сети

Для проведения анализа трафика необходима математическая модель нормального функционирования сети и допустимые границы разброса её параметров. Математическая модель сетевого трафика удобно описывается с помощью известных методов спектрального анализа. Однако в ККС информационные процессы нестационарны и неоднородны в пространстве, что снижает адекватность используемых моделей.

Представляется перспективным для моделирования сетевого трафика и выявления угроз безопасности информационных ресурсов сети использовать теорию мультиразрешающего анализа (МРА) [2], основой которой составляют методы вейвлет – преобразований и теорию идентификации динамических систем, одной из широко применяемых параметрических моделей которой является прогнозирующая авторегрессионная модель (ARX-модель) (рисунок 8.1).

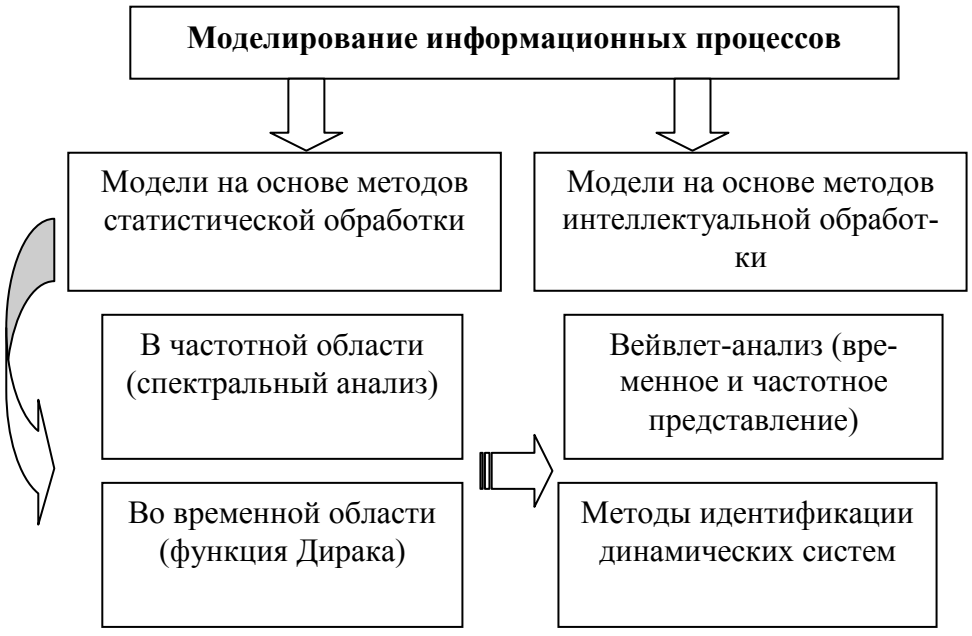


Рисунок 8.1 – Модели информационных процессов

Теоретической основой МРА является теория вейвлет – преобразований (ВП). Вейвлеты (wavelet – короткая волна) – это обобщенное название функций определенной формы, локализованных по оси аргументов (независимых переменных), инвариантных к сдвигу и линейных к операции масштабирования

(сжатия/растяжения), имеющих вид коротких волновых пакетов с нулевым интегральным значением [2]. В основе ВП специальные базовые функции, которые определяют их вид и свойства. По локализации во временном и частотном представлении вейвлеты занимают промежуточное положение между гармоническими функциями, локализованными по частоте, и функцией Дирака [11], локализованной во времени.

Основная область применения вейвлетных преобразований – анализ и обработка сигналов и функций, в том числе и дискретных в виде массивов цифровых данных, нестационарных во времени или неоднородных в пространстве, когда результаты анализа должны содержать не только общую частотную характеристику сигнала (распределение энергии сигнала по частотным составляющим), но и сведения об определенных локальных координатах, на которых проявляются те или иные группы частотных составляющих, или, на которых происходят быстрые изменения частотных составляющих функций. По сравнению с разложением сигналов на ряды Фурье, вейвлеты способны с гораздо более высокой точностью представлять локальные особенности функций, вплоть до разрывов 1-го рода (скачков). Кроме того, в отличие от преобразований Фурье, вейвлет-преобразование одномерных массивов цифровых данных обеспечивает двумерную развертку, при этом частота и координата рассматриваются как независимые переменные, что дает возможность анализа массивов сразу в двух пространствах.

Целью построения модели является оперативное управление трафиком сети, представляющим собой динамическую систему, в режиме реального времени. Прогнозирование состояния трафика сети на шаг вперёд позволит повысить оперативность принятия решения. Одним из часто применяемых видов моделей прогнозирования состояния динамических систем является класс ARIMAX-моделей, которые позволяют устанавливать зависимость состояния системы в момент времени t от предыдущих состояний в моменты времени $s \leq t-1$ и позволяют прогнозировать состояние системы, в частности, состояние трафика сети.

8.2 Модель сетевого трафика системы обнаружения аномалий

На рисунке 8.2 предложена технология выявления и блокирования аномалий трафика ККС.

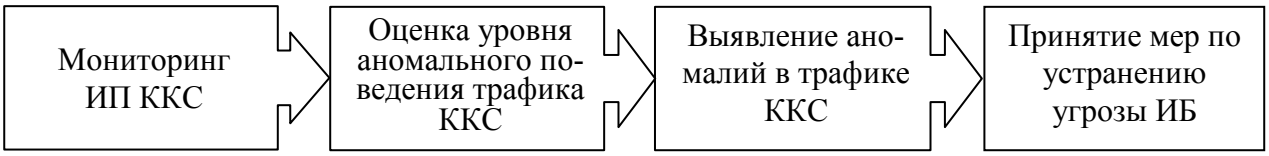


Рисунок 8.2 – Технология выявления и блокирования аномалий

Под мониторингом ИП ККС понимается анализ сетевого трафика (объем переданной информации в байтах, количество переданных пакетов, количество потоков) за определенный интервал времени. Трафик сети рассматривается в виде совокупности одномерных числовых рядов $f(t_i)$ – отсчётов характеристик трафика, заданных в дискретные моменты времени $t_i = i\Delta$, где $i = 0,1,\ldots, N\text{-}1$, Δ – интервал между отдельными наблюдениями, N – количество наблюдений.

В качестве характеристик трафика берутся его показатели, имеющие потенциал для различения нормального состояния сети от аномалии:

а) количество потоков в трафике за период времени;

б) среднее количество пакетов в потоке на интервале времени. Большинство атак происходит с увеличением количества пакетов;

в) среднее количество байт в потоке на интервале времени. Некоторые атаки используют большой размер пакета для истощения вычислительных ресурсов;

г) среднее количество байт на пакет в потоке за период времени. Описывает размер пакета более подробно, чем в предыдущем случае;

д) отношение количества потоков к среднему количеству пакетов в потоке. Характеризует начало большого количества соединений с небольшими пакетами с целью достижения максимальной производительности сканирования.

Каждый полученный ряд обрабатывается независимо от остальных. Такое описание трафика позволяет учитывать его различные характеристики, не прибегая к анализу многомерных сигналов и обрабатывать их в параллельном режиме.

Произвольная последовательность $f(t_i)$ в теории цифровых временных рядов обычно рассматривается в виде суммы разнотипных составляющих [11]:

– функции тренда $q_т(t_i)$ – средних значений по большим интервалам усреднения (медленно меняющаяся во времени функция, описывающая изменения среднесуточных загрузок ККС за интервалы времени большие, чем суточная периодичность);

– циклических компонент с определенным периодом повторения $q_ц(t_i)$ как правило, достаточно гладких по форме (периодическая составляющая, описывающая изменения среднесуточных загрузок ККС);

– локальных особенностей (аномалий) разного порядка $\varepsilon_a(t_i)$, вплоть до вторжений – резких изменений в определенные редкие моменты;

– флюктуаций значений более высокого порядка (шумов**)** $\varepsilon_ф(t_i)$ вокруг всех вышеперечисленных составляющих функции.

Таким образом, обобщенная модель сетевого трафика может быть представлена в следующем виде:

$$f(t_i) = q_т(t_i) + q_ц(t_i) + \varepsilon_a(t_i) + \varepsilon_ф(t_i). \qquad (8.1)$$

Построение модели сетевого трафика сводится к определению аналитического представления каждого слагаемого (8.1). В ранее проведенных исследованиях относительно независимой случайной последовательности $\varepsilon_ф(t_i)$ принималось допущение, что в случайные моменты времени t_i математическое ожидание $M[\varepsilon_ф(t_i)] = 0$ с дисперсией $\sigma_\xi^2.(t_i)$. Однако использование открытых каналов передачи данных привело к существенному росту шумовой составляющей $\varepsilon_ф(t_i)$ ИП ККС и допущение стало неприемлемым.

При выполнении МРА сетевого трафика *f(ti)* гильбертово пространство *L2(R)* этого одномерного числового ряда представляется в виде системы вло-

женных подпространств Vm , отличающихся друг от друга перемасштабированием независимой переменной.

Исходные условия МРА можно сформулировать следующим образом:

1 Пространство $L^2(R)$ может быть представлено в виде последовательности вложенных друг в друга замкнутых подпространств соответствующих уровней декомпозиции ряда:

$$\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \subset V_m \subset V_{m+1} \ldots.$$

«Размеры» подпространств непрерывно расширяются по мере роста значения m (масштабирующий коэффициент), а объединение всех подпространств, в пределе, дает пространство $L^2(R)$, при этом подпространства не пересекаются. Для этого требуется выполнение условий:

– условие полноты и плотности разбиения в виде $\cup_{m \in I} V_m = L^2(R)$,

– условие ортогональности подпространств – $\cap_{m \in I} V_m = \{0\}$.

2 Для любой функции (8.1) $f(t_i) \in V_m$ ее масштабное преобразование по аргументу в 2 раза перемещает функцию в соседнее подпространство:

$$f(t) \in V_m \Leftrightarrow f(2t) \in V_{m+1}, \ldots f(t) \in V_m \Leftrightarrow f(t/2) \in V_{m-1}$$

3 Для пространства V_0 существует функция $\varphi(t) \in V_0$, целочисленные сдвиги которой по аргументу образуют ортонормированный базис пространства V_0:

$$\varphi_{0,k} = \varphi(t-k), \quad k \in I \quad (k = 0, \pm 1, \pm 2, \ldots).$$

Из условий 2 и 3 непосредственно следует, что если пространство V_0 имеет ортонормированный базис $\varphi_{0,k}$, то и все остальные пространства также имеют ортонормированные базисы, которые образуются масштабным преобразованием базиса $\varphi_{0,k}$:

$$\varphi_{m,k}(t) = 2^{m/2} \varphi(2^m t-k), \quad m,k \in I. \qquad (8.2)$$

Условия 1 и 2 гарантируют, что если функция $f(t_i)$ принадлежит пространству V_m, то одновременно $f(t_i)$ входит и в пространство V_{m+1}, и вместе с ним в этом пространстве находится и массив данных $f(2t)$. Увеличение номера пространства позволяет изучать все более и более мелкие детали и особенности массива данных (трафика) с более высокочастотными компонентами (как под микроскопом).

Для того чтобы использовать МРА, достаточно знать только одно из подпространств V_m, остальные определятся уравнением (8.2). Поскольку $V_0 \subset V_1$, то функцию $\varphi_0(t)$ можно представить линейной комбинацией сдвигов функции $\varphi_1(t)$ (с учетом ее более компактного носителя) с определенными весовыми коэффициентами h_k.

В общем случае, носитель функции может иметь произвольный размер с числом отсчетов $2M$ (в единицах k), при этом уравнение линейной связи базисных функций пространств записывается в следующем виде [2]:

$$\varphi(t) = \sqrt{2} \sum_{k=0}^{2M-1} h_k \varphi(2t-k). \qquad (8.3)$$

Уравнение (8.3) называется масштабирующим. Решение этого уравнения дает функцию, которую называют «отцовским» вейвлетом. Значения h_k определяются из условия для ортонормальных базисов [2,11]:

$$h_k = \sqrt{2} \int_t \varphi(t)\varphi^*(2t-k)\, dt \qquad (8.4)$$

где $\varphi^*(t)$ – функция, комплексно сопряженная с $\varphi(t)$.

При дискретных значениях параметров сдвига масштабирующий вейвлет $\varphi(t)$ также дискретен и при задании функции $\varphi(t)$ на конечном интервале имеет конечное число коэффициентов h_k, отличных от нуля.

Из условия нормировки масштабирующих коэффициентов: $\int_{-\infty}^{\infty} \varphi(t)\, dt = 1$, следует: $\sum_{k} (h_k)^2 = 1$.

Функции $\varphi_{m,k}(t)$ образуют ортонормальный базис пространства V_m . При переходе из пространства V_{m+1} в пространство V_m от пространства V_{m+1} отделяется подпространство W_m функции $\psi_{m,k}(t)$ – подпространство «материнских» вейвлетов.

Следовательно, пространства V_{m+1} могут быть представлены в виде суммы подпространств:

$$V_{m+1} = V_m \oplus W_m, \quad m \in I.$$

В пределе, с учетом свойства ортогональности пространств, пространство $\boldsymbol{V_{m+1}}$ определяется как:

$$V_{m+1} = \bigoplus_{k=-\infty}^{m} W_k.$$

Пространства W_m образуют взаимно ортогональный набор, в котором вейвлеты $\psi_{m,k}(t)$ формируют ортонормальный базис при любом заданном уровне разрешения m.

Физический смысл процесса разложения пространств представлен на рисунке 1.2.

Исходное пространство V_{m+1} является пространством цифрового массива с определенным частотным диапазоном. При разложении дискретной функции в пространство W_m отделяются высокочастотные составляющие пространства V_{m+1}, а в пространстве V_m остаются его низкочастотные составляющие. С этих позиций функции $\psi_{m,k}(t)$ и $\varphi_{m,k}(t)$ играют роль высокочастотного и низкочастотного фильтров соответственно.

Если «отцовский» вейвлет установлен, то базисный («материнский») вейвлет определяется зависимостью [2]:

$$\psi_n(t) = \sqrt{2} \sum_{k \in I} g_k \varphi(2t - k). \qquad (8.5)$$

$$g_k = (-1)^k h_{2M-1-k}. \qquad (8.6)$$

Следовательно, используя методологию MPA модель сетевого трафика $f(t_i)$ (8.1) можно рассматривать на любом m - уровне разрешения между ее усредненными значениями и флюктуациями вокруг средних значений в виде:

$$f(t_i) = \sum_{k=-\infty}^{\infty} c_{m,k}\,\varphi_{m,k}(t) + \sum_{m=m'}^{\infty} \sum_{k=-\infty}^{\infty} d_{m,k}\,\psi_{m,k}(t), \qquad (8.7)$$
$$m, k \in I,$$

где $\varphi_{m,k}(t)$ – масштабирующая функция, с помощью которой выполняется аппроксимация сетевого трафика;

$\psi_{m,k}(t)$ – вейвлет-функция, выделяющая детали сетевого трафика и его локальные особенности;

$c_{m,k}$, $d_{m,k}$ – апроксимирующие и детализирующие коэффициенты;

m – параметр масштаба;

k – параметр сдвига;

I – пространство целых чисел $\{-\infty, \infty\}$.

Значения коэффициентов $c_{m,k}$ и $d_{m,k}$, которые называются масштабными коэффициентами (приближения) и вейвлет-коэффициентами (детали), определяются зависимостями [2]:

$$c_{m,k} = \int_t f(t)\ \varphi_{m,k}(t)\,dt\,, \qquad (8.8)$$

$$d_{m,k} = \int_t f(t)\ \psi_{m,k}(t)\,dt\,, \qquad (8.9)$$

Первая сумма в (8.7) содержит усредненные (с весовыми функциями $\varphi_{m,k}$) значения функции $f(t_i)$ по диадным интервалам $[k\cdot 2^{-m}, (k+1)\cdot 2^{-m}]$, характеризует тренд и циклические составляющие трафика (суточные и недельные) [11], т.е.

$$q_m(t_i) + q_ц(t_i) = \sum_{k=-\infty}^{\infty} c_{m,k}\,\varphi_{m,k}(t), \qquad (8.10)$$

а вторая – значения флюктуаций на данных интервалах, характеризующих аномалии трафика $\varepsilon_a(t_i)$, с учетом случайной шумовой помехи $\varepsilon_\phi(t_i)$ [11,14]:

$$\varepsilon_a(t_i) + \varepsilon_\phi(t_i) = \sum_{m=m'}^{\infty} \sum_{k=-\infty}^{\infty} d_{m,k}\, \psi_{m,k}(t). \qquad (8.11)$$

Таким образом, выражение (8.7) показывает возможность аппроксимации любой произвольной функции $f(t_i)$ набором простых локальных функций $\varphi_{m,k}(t)$ и $\psi_{m,k}(t)$, ортогональных на разных уровнях значений m и полностью покрывающих пространство $L^2(R)$ за счет смещений k. Переход от m к $m+1$ эквивалентен замене t на $2t$, т.е. перемасштабированию функций $\varphi_{m,k}(t)$ и $\psi_{m,k}(t)$.

При оценке сетевого трафика в виде конечного набора отсчетов, наилучший уровень разрешения определен интервалом, содержащим один отсчет, и суммирование выполняется в конечных пределах. Значение $m = 0$ принимается для этого наилучшего уровня разрешения. При принятой форме вейвлетов (8.2), (8.5) коэффициенты $c_{m,k}$ и $d_{m,k}$ вычисляются для $m > 0$.

МРА при последовательном увеличении значений m приводит к форме быстрых итерационных вычислений вейвлет-коэффициентов вида [11,14]:

$$c_{m+1,k} = \sum_n h_n\, c_{m,2k+n}\,, \qquad (8.12)$$

$$d_{m+1,k} = \sum_n g_n\, c_{m,2k+n}\,, \qquad (8.13)$$

$$c_{0,k} = \int_{k\Delta t}^{(k+1)\Delta t} f(t)\cdot \varphi(t-k)\, dt\,. \qquad (8.14)$$

Для массивов цифровых данных сетевого трафика в качестве значений $c_{0,k}$ принимаются исходные значения одномерного числового ряда, т.е. $c_{0,k} = f(k) = f(t_0)$.

Отсюда, явный вид вейвлета требуется только для расчета коэффициентов h_n и g_n, а при собственно быстром вейвлет-преобразовании он не используется – используются полученные значения коэффициентов h_n и g_n на соответствующем уровне детализации.

Уравнения (8.12), (8.13) обеспечивают реализацию быстрого вейвлет-преобразования одномерного числового ряда на основе пирамидального алгоритма вычисления вейвлет-коэффициентов (алгоритм Малла) [2,11,14], приведенного на рисунке 8.3 слева или алгоритма вейвлет-пакетов (ВП), приведённого на рисунке справа.

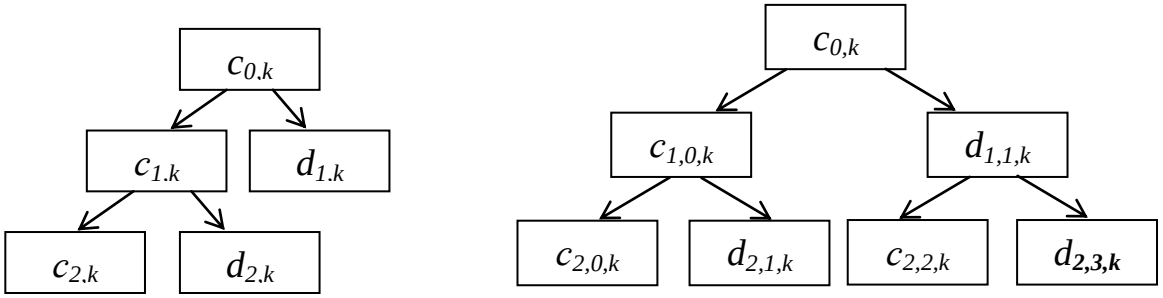


Рисунок 8.3 – Алгоритм Малла и вейвлет-пакетов для MPA

Сущность операций алгоритма Малла, выполняемых на основе (8.12) и (8.13), заключается в следующем. С учетом спектров коэффициентов h_n и g_n , на первом этапе преобразования первый цифровой фильтр h_n из числового ряда $f_k = c_{0,k}$ выделяет низкие частоты $|\omega| \le \pi/2$, а другой (октавный) фильтр g_n выделяет верхние частоты $\pi/2 \le |\omega| \le \pi$ (рисунок 8.4). Поскольку на выходе фильтра h_n отсутствует верхняя половина частот, то частота дискретизации выходного сигнала может быть уменьшена в 2 раза, т.е. выполнена децимация выходного массива, что и производится в формуле (8.12) сдвигами $(2k+n)$ через 2 отсчета по входному массиву. Соответственно, на выходе фильтра g_n освобождается место в области низких частот, и аналогичное прореживание выходного числового массива приводит к транспонированию верхних частот на освободившееся место. Таким образом, каждый из выходных числовых массивов несет информацию о своей половине частот, при этом выходная информация представлена таким же количеством отсчетов, что и входная.

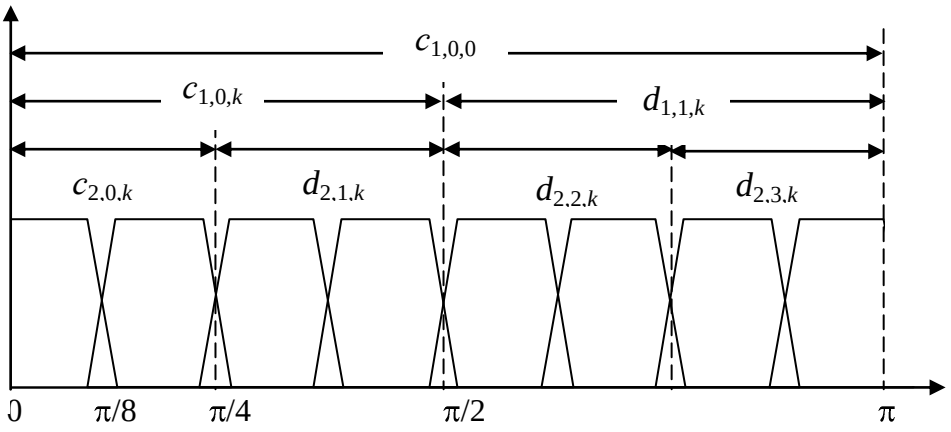


Рисунок 8.4 – Спектральные характеристики ВП

ВП позволяют наиболее полно выделить локальные особенности трафика – всплески и аномалии за счёт дополнительной декомпозиции высокочастотных составляющих спектра сигнала, представляющего трафик (рисунок 8.4 справа). Реализация алгоритма ВП аналогична пирамидальному алгоритму – свёрткой с фильтрами h_n и g_n. В алгоритме ВП при переходе с масштабного уровня m на уровень m+1 как низкочастотные (аппроксимирующие) коэффициенты $c_{m,k}$, так и высокочастотные (детализирующие) коэффициенты $d_{m,k}$ разделяются на низкочастотные ($c_{m+1,k}$) и высокочастотные ($d_{m+1,k}$) части спектрального диапазона по формулам [13]:

$$c_{m+1,2p,k} = \sum_n h_n c_{m,p,2k+n},\ d_{m+1,2p,k} = \sum_n g_n c_{m,p,2k+n},\ c_{0,k} = f(k), \qquad (8.15)$$

$$c_{m+1,2p+1,k} = \sum_n h_n d_{m,p,2k+n},\ d_{m+1,2p+1,k} = \sum_n g_n d_{m,p,2k+n},.\ k = 0..N/2^m. \qquad (8.16)$$

Итак, правая сумма (8.7), задающей временной ряд, характеризует аномалии сетевого трафика ККС на фоне флюктуаций и, следовательно, анализ данной части можно положить в основу методики определения текущего уровня отклонения от нормального поведения сетевого трафика.

Из множества возможностей определения базиса для разложения ВП – от «минимального» вейвлет-разложения (алгоритм Малла) до полного пакетного раз-

ложения на всех уровнях выбирается оптимальный, адаптированный к анализируемому трафику.

При обычном вейвлет-разложении (алгоритм Малла) вейвлет-коэффициентов достаточно для восстановления трафика, и можно раскладывать высокочастотные коэффициенты деталей или отказаться от этого. Таким образом, появляется гораздо больше возможностей выбора базиса для разложения – от «минимального» вейвлет-разложения до полного пакетного разложения на всех уровнях. Из всех представителей нужно выбрать то, которое представляет трафик наиболее эффективно. Под «эффективным» подразумевается то, что трафик может быть представлен небольшим количеством коэффициентов разложения, т.е. базис для разложения должен быть таким, что большие коэффициенты сконцентрированы на небольшом количестве элементов вейвлет-пакетного базиса и большое количество коэффициентов близко к нулю [13, 14].

В качестве критерия выбора самого эффективного или лучшего базиса для трафика используется критерий минимальности энтропии. Энтропия характеризует усреднённость трафика. Лучший базис – это тот базис, который даёт наименьшее количество энтропии. Интуитивно энтропия Шеннона дает критерий того, сколько эффективных компонент необходимо, чтобы представить трафик в определенном базисе. Чем меньше энтропия, тем меньше существенных коэффициентов нужно для представления трафика.

Критерий энтропии используется для выбора оптимального базиса следующим образом. Если при разложении коэффициентов некоторого узла сумма энтропий, полученных при разложении компонент, меньше, чем энтропия коэффициентов в исходном узле, то разложение применяется, в противном случае коэффициенты (вместе с базисными функциями) остаются без изменения.

На рисунке 8.5 приведен пример трафика и соответствующее ему вейвлет-пакетное разложение.

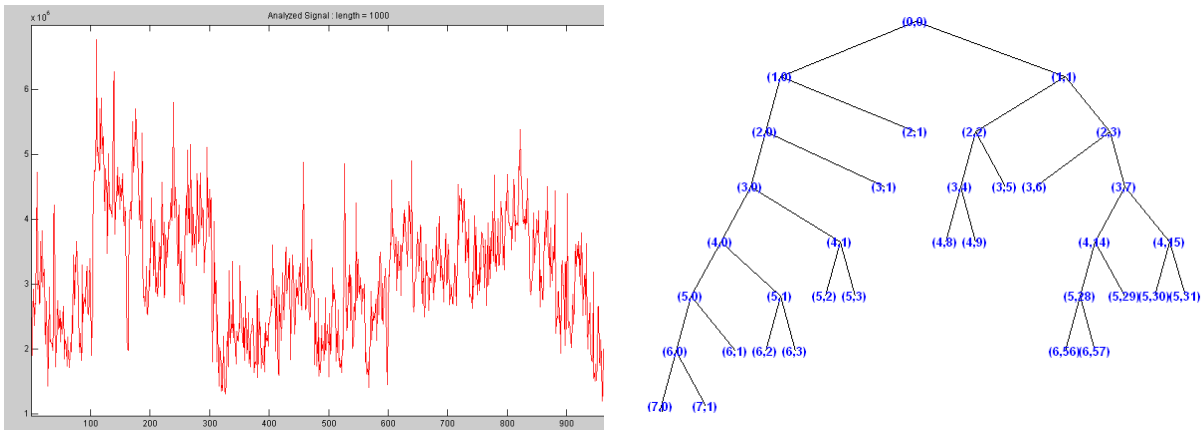


Рисунок 8.5 – Оптимальное вейвлет-пакетное разложение

Пусть значения номеров уровней и номеров узлов определены по результатам выбора оптимального базиса вейвлет-пакетов, и среди них номера уровней m_{o1}.. m_{ol} и номера узлов p_{o1} ..p_{oz} характеризуют аномалии, а номера уровней $m_{\phi 1}$.. $m_{\phi l}$ и номера узлов $p_{\phi 1}$ ..$p_{\phi z}$, характеризуют шум. Тогда формула, описывающая аномалии сетевого трафика со случайной шумовой составляющей трафика, принимает вид [13,14]:

$$f_d(t_i) = \varepsilon_a(t_i) + \varepsilon_\phi(t_i) = \sum_{m=m_{o1}}^{m_{ol}} \sum_{p=p_{o1}}^{p_{oz}} \sum_{k=0}^{N/2^m} d_{m,p,k} \Psi_{m,p,k}(t_i) + \sum_{m=m_{\phi 1}}^{m_{\phi l}} \sum_{p=p_{\phi 1}}^{p_{\phi z}} \sum_{k=0}^{N/2^m} d_{m,p,k} \Psi_{m,p,k}(t_i) \quad (8.17)$$

Тогда разность между эталонным $f_{d^э}(t_i)$ уровнем аномального поведения сетевого трафика с учетом случайной шумовой помехи, определяемым в нормальном режиме, и регистрируемым уровнем $f_{d^р}(t_i)$ с предположением равенства флюктуаций $\varepsilon_\phi^э(t)$ и $\varepsilon_\phi^р(t)$, определяет текущий уровень отклонения от нормального поведения трафика ККС:

$$\tilde{\varepsilon}_a(t_i) = f_{d^э}(t_i) - f_{d^р}(t_i) = \sum_{m=m_{o1}}^{m_{ol}} \sum_{p=p_{o1}}^{p_{oz}} \sum_{k=0}^{N/2^m} d^э_{m,p,k} \Psi_{m,p,k}(t_i) - \sum_{m=m_{o1}}^{m_{ol}} \sum_{p=p_{o1}}^{p_{oz}} \sum_{k=0}^{N/2^m} d^р_{m,p,k} \Psi_{m,p,k}(t_i) . \quad (8.18)$$

Предложенный метод вейвлет – преобразований сетевого трафика (в виде массива цифровых данных) позволяет адекватно описать его модель с учетом нестационарности и неоднородности процессов, в частности, стохастическую его

часть, характеризующую аномалии трафика ККС $\varepsilon_a(t_i)$ и шумовой помехи $\varepsilon_\phi(t_i)$.

8.3 Прогнозирование текущего состояния трафика ККС

Предложенная модель (8.18) не обеспечивает обнаружение аномалий в режиме реального времени в силу необходимости постоянного пересчета флуктуационной составляющей $\varepsilon_\phi^э(t)$ нормального состояния трафика сети [13].

Для решения этой проблемы, предлагается прогнозировать величину $f_{d^э}(t)$, используя модель «черного ящика», на выходе которого генерируются стохастические процессы в зависимости от управляющего сигнала $D^э(t)$, зашумленного некоторым неконтролируемым сигналом $\varepsilon_\phi(t)$ (рисунок 8.6).

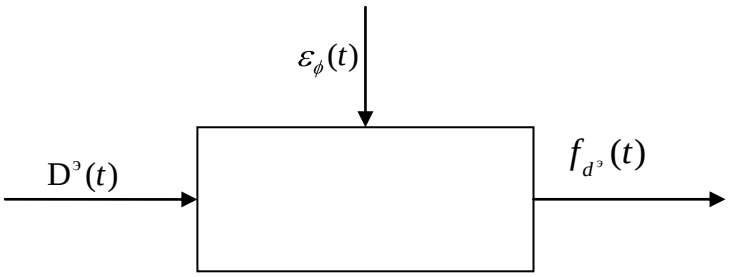


Рисунок 8.6 – Модель «черного ящика»

Из множества структур моделей для рассматриваемой совокупности наблюдений наиболее предпочтительной является класс ARIMAX-моделей линейной разностной динамической системы, компоненты которого могут самостоятельно использоваться как авторегрессионные, стационарные, нестационарные, нелинейные и робастные к различным проявлениям среды модели [7].

Из всех представителей класса ARIMAX-моделей широко применяется при обработке различного рода сигналов ARX-модель.

Одна из форм записи ARX-модели имеет вид [7]

$$y(t) = \sum_{i=1}^{p} a_i y(t-i) + \sum_{i=1}^{q} b_i x(t-i) + e(t), \qquad (8.19)$$

где a_i и b_i настраиваемые параметры модели,

$y(t\text{-}i)$ – предыдущие значения выхода (образцы);

$x(t\text{-}i)$ – входные значения (регрессор);

$e(t)$ – белый шум.

Тогда предсказатель для ARX-модели случайной составляющей сетевого трафика с регрессором в форме вейвлет-коэффициентов примет вид [13]

$$\tilde{f}_{d^э}(t) = \sum_{i=1}^{r} a_i f_{d^э}(t-i) + \sum_{i=1}^{q} b_i D^э(t-i), \qquad (8.20)$$

где a_i и b_i – параметры модели;

$f_{d^э}(t)$ – значения случайной составляющей, определяющие предыдущие значения выхода (образцы);

$D^э(t) = \{ d_{m,p,k}, \} m = m_{o1},..m_{ol};\ p = p_{j1},...p_{jz};\ k = 0..(N+1)/2^m$ – вектор коэффициентов, определяющий входные значения (регрессор);

r, q – показатели глубины истории.

Ошибка прогноза определяется зависимостью вида [13]:

$$\xi(t) = \tilde{f}_{d^э}(t) - f_{d^э}(t), \qquad (8.21)$$

где $f_{d^р}(t)$ – эталонные значения случайной составляющей.

Значения параметров $\theta=\{a_i,b_i\}$ определяются из условия минимума ошибки прогноза (8.21) по методу наименьших квадратов.

Тогда прогнозируемый уровень отклонения от нормального состояния трафика определяется как разность прогнозируемого и текущего значения случайной составляющей:

$$\tilde{\varepsilon}_a(t) = \tilde{f}_{d^э}(t) - f_{d^р}(t), \qquad (8.22)$$

где $f_{d^р}(t)$ – текущие значения случайной составляющей.

Для линейных регрессий формула (8.20) может быть записана в виде

$$\tilde{f}_{d^э}(t) = \varphi^T(t)\theta, \qquad (8.23)$$

где $\varphi(t) = (-f_{d^э}(t-1),..., -f_{d^э}(t-r), d^э(t-1),...,d^э(t-q)^T$ – вектор регрессий.

Тогда согласно методу наименьших квадратов вектор θ оптимальных параметров можно вычислить путём решения линейной системы с симметричной матрицей

$$R(N)\theta = f(N), \qquad (8.24)$$

где

$$R(N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\cdot\varphi^{T}(t),\ \ f(N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\cdot y(t).$$

Вектор регрессий $\varphi(t)$ составляется из значений трафика и вейвлет-коэффициентов.

Значения для r и q определены экспериментально и равные соответственно 7 и 8. На рисунке 8.7 изображены графики высокочастотной составляющей текущего трафика с аномалией (точечная кривая) и её прогнозируемые значения с помощью построенной модели. Аномалия в данном случае есть результат сканирования сети. Рисунок 8.9 показывает, что отклонения прогнозируемых значений от текущих в месте аномалии сильно отличаются от остальных.

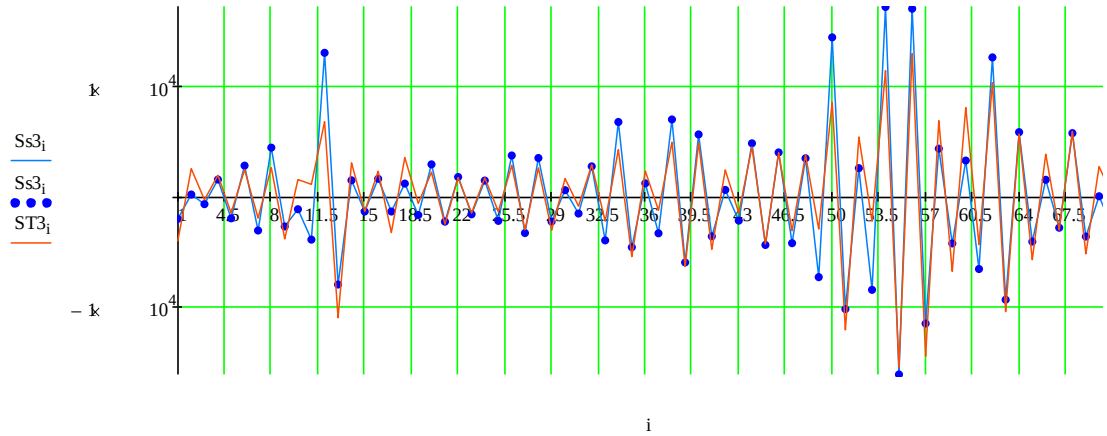


Рисунок 8.7 –Прогнозирование случайной составляющей трафика

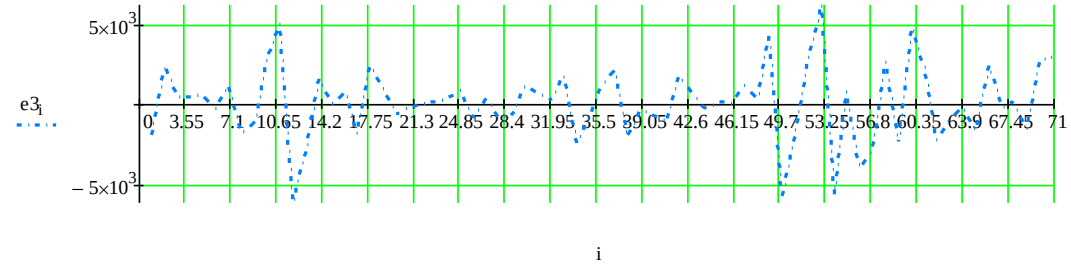


Рисунок 8.8 – Текущая аномальная активность

Таким образом, прогнозирующая ARX-модель (8.20) устанавливает зависимость состояния трафика в момент времени t от предыдущих состояний в моменты времени t-1, t-2, …, t-r и прогнозирует текущие значения случайной составляющей $f_{d'}(t)$ трафика сети на величину глубины прогноза, тем самым повышая оперативность принятия решения при выявлении сетевых аномалий, не требуя постоянного пересчета текущих значений [13].

Вопросы для самоконтроля

1 Что затрудняет адекватное описание математических моделей информационных процессов с помощью методов гармонического анализа?

2 Почему для математического моделирования сетевого трафика и выявления угрозы безопасности информационных ресурсов сети целесообразно использовать теорию мультиразрешающего анализа?

3 Сформулируйте понятие вейвлетов.

4 Чем отличаются вейвлеты от гармонических функций?

5 В чем преимущество вейвлетов по сравнению с разложением сигналов на ряды Фурье?

6 Какова цель построения математической модели сетевого трафика?

7 Для чего нужно прогнозирование состояния трафика сети в задаче обнаружения сетевых аномалий?

8 Что понимается под мониторингом информационных процессов компьютерных корпоративных сетей?

9 Какие характеристики трафика используются при построении математической модели сетевого трафика?

10 Какие составляющие входят в обобщенную модель сетевого трафика?

11 Что дает решение масштабирующего уравнения?

12 В чем состоит физический смысл в мультиразрешающего анализа?

13 Какие коэффициенты составляют модель сетевого трафика по методологии мультиразрешающего анализа и что они характеризуют?

14 Сущность операций алгоритма Малла.

15 Какой критерий следует использовать для выбора самого эффективного или лучшего базиса вейвлет-разложения для трафика?

16 Что нужно для обнаружения аномалий в режиме реального времени?

17 Для чего предназначен класс ARIMAX-моделей?

18 Исходя из какого условия вычисляются параметры для ARX-модели?

19 Как можно вычислить вектор оптимальных параметров для ARX-модели?

20 Как определяется прогнозируемый уровень отклонения от нормального состояния трафика?

21 Что позволяет прогнозирующая ARX-модель в задаче обнаружения сетевых аномалий?

Глава 9 Идентификация уровня аномальности трафика корпоративной сети

В основу идентификации уровня аномальности трафика ККС положена математическая модель сетевого трафика (8.7) и прогнозирующая авторегрессионная модель (8.20), используемая в двух режимах: обучение и анализ [13].

9.1 Методика идентификации уровня аномальности трафика корпоративной сети

В режиме обучения системы поддержки принятия решений (СППР) идентификации аномалий проводится моделирование нормального трафика сети, которое состоит из этапов, представленных на рисунке 9.1 [13].

Рисунок 9.1 – Моделирование нормального трафика

СППР собирает информацию о трафике, воспринимая его как нормальный режим работы сети. Длительность обучения и временные интервалы регистрации трафика составляют отдельную научную задачу и обосновываются ниже.

При этом СППР на основе пакетов, прошедших через средство разграничения доступа (маршрутизатор), фиксирует отсчеты сигналов, и запоминает их в базе данных в текстовом формате цифровых данных в виде конечного набора отсчетов.

В ходе вейвлет-разложения полученные цифровые массивы отсчётов, представляющие трафик сети, преобразуются в наборы коэффициентов с помощью одного из быстрых алгоритмов кратноразрешающего анализа – вейвлет-пакетов (ВП).

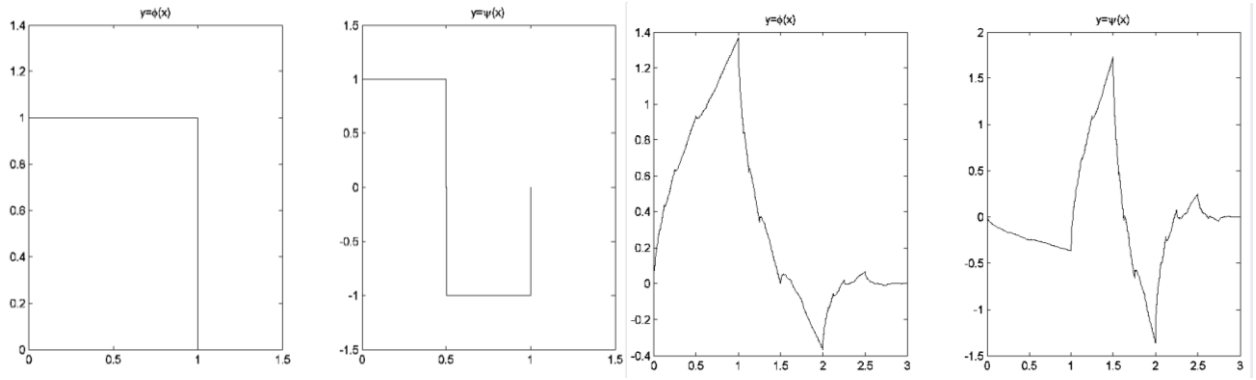
Далее на основе полученных коэффициентов с помощью модели (8.20) прогнозируется нормальный уровень трафика $\tilde{f}_{d^p}(t)$. Входной вектор $D^э(t)$ модели формируется из высокочастотных коэффициентов $d^э_{m,p,k}$. Вектор предыдущих значений выхода $f_{d^э}(t)$ составляется из значений характеристик трафика, также восстановленных из высокочастотных компонент. В режиме анализа по очередным отсчётам трафика, рассчитывается $f_{d^p}(t)$ и текущий уровень отклонения от нормального поведения сетевого трафика по формуле (8.21).

В настоящее время в МРА разработано большое количество вейвлетов. Выбор анализирующего вейвлета во многом определяется тем, какую информацию необходимо извлечь из массива цифровых данных. С учетом характерных особенностей различных вейвлетов во временном и в частотном пространстве, можно выявлять в анализируемых массивах цифровых данных те или иные свойства и особенности, которые незаметны на гистограммах, особенно в присутствии шумов $\varepsilon_ф(t_i)$. Если целью СППР является выявление аномалий трафика ККС, необходимо извлечение детализирующей информации из массива цифровых данных.

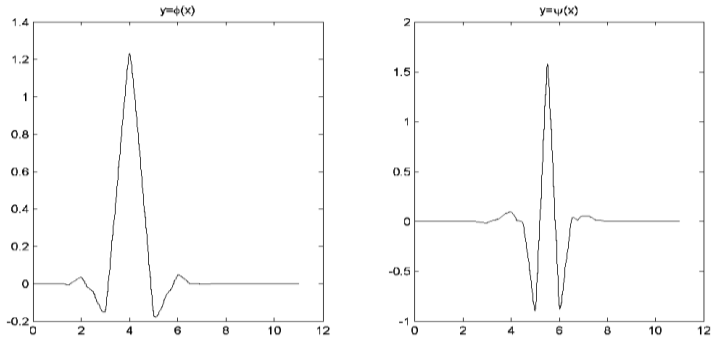
Для выбора системы базисных вейвлетов проанализированы вейвлеты с компактным носителем: Хаара, Добеши – 2, вейвлеты Койфмана – койфлеты –

2 (рисунок 9.2), которые выделяют локальные особенности сигналов.

Для мониторинга сетевого трафика в качестве базиса целесообразно использовать систему вейвлетов – койфлеты-2, которые имеют близкую к симметричной форму, обеспечивают большее количество малозначимых, близких к нулевым, коэффициентов разложения и имеют более высокую крутизну среза полосы пропускания, а, соответственно, обеспечивают лучшее качество разложения сигналов и их реконструкции.



а) б)



в)

Рисунок 9.2 – Вейвлеты: а) Хаара; б) Добеши – 2, в) койфлеты – 2

Вид массива данных сетевого трафика во временном пространстве с наложением графика койфлетов – 2 представлен на рисунке 9.3 .

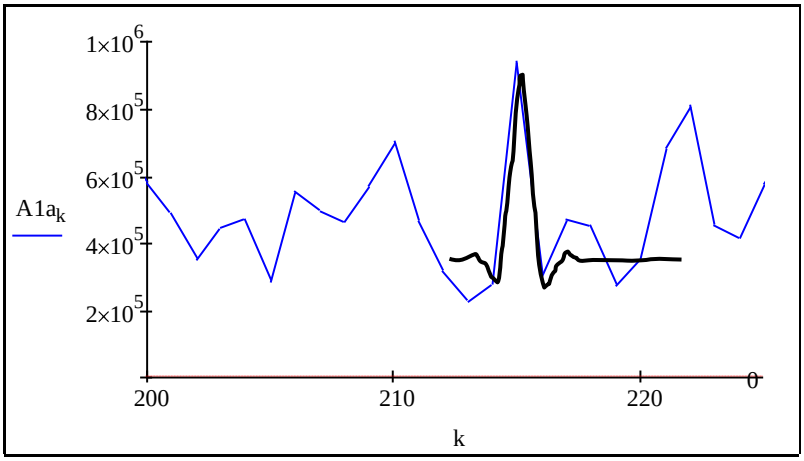


Рисунок 9.3 – Трафик сети с наложением графика койфлета

Таким образом, предложенную методику на основе интеграции вейвлет-пакетной модели сетевого трафика с базисными функциями койфлетами – 2 и прогнозирующей авторегрессионной модели можно использовать для выявления аномальности трафика ККС.

9.2 Разработка алгоритмов мониторинга информационных процессов

На этапе алгоритмизации модель вида (8.19, 8.23) воплощается в конкретную машинную модель. В соответствии с рассмотренным ранее математическим аппаратом выявления аномалий сетевого трафика разработан соответствующий алгоритм функционирования СОВ. На рисунке 9.4 представлена укрупненная схема алгоритма мониторинга информационных процессов ККС – мониторинга сетевого трафика [13].

На последующих рисунках 9.5 – 9.6 представлены схемы алгоритмов функций, запускаемых при мониторинге сетевого трафика [13].

Мониторинг трафика начинается непосредственно со считывания текущих значений трафика. Данное действие осуществляется функцией StFW. Считанные значения трафика сохраняются в массив TrafRow.

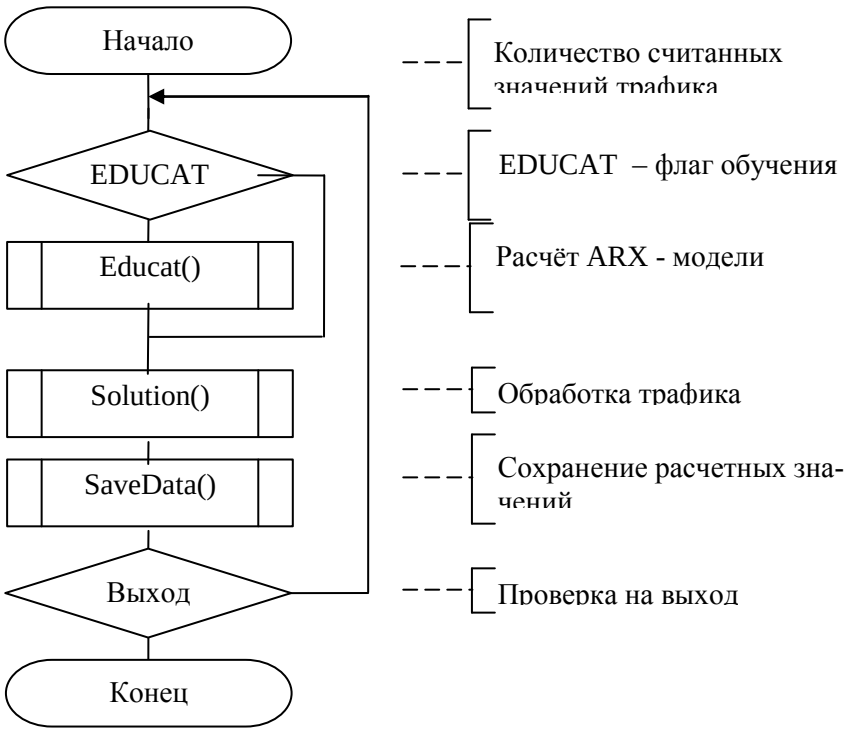


Рисунок 9.4 – Укрупненная схема алгоритма мониторинга информационных процессов ККС

Рисунок 9.5 – Алгоритм определения уровня угрозы безопасности Solution()

Далее полученные данные (массив TrafRow) подвергаются вейвлет-анализу (функция Solution), по результатам которого можно судить о наличии угрозы ИБ.

Затем полученные результаты сохраняются в базе данных (функция SaveData), после чего цикл повторяется снова либо происходит выход из мониторинга.

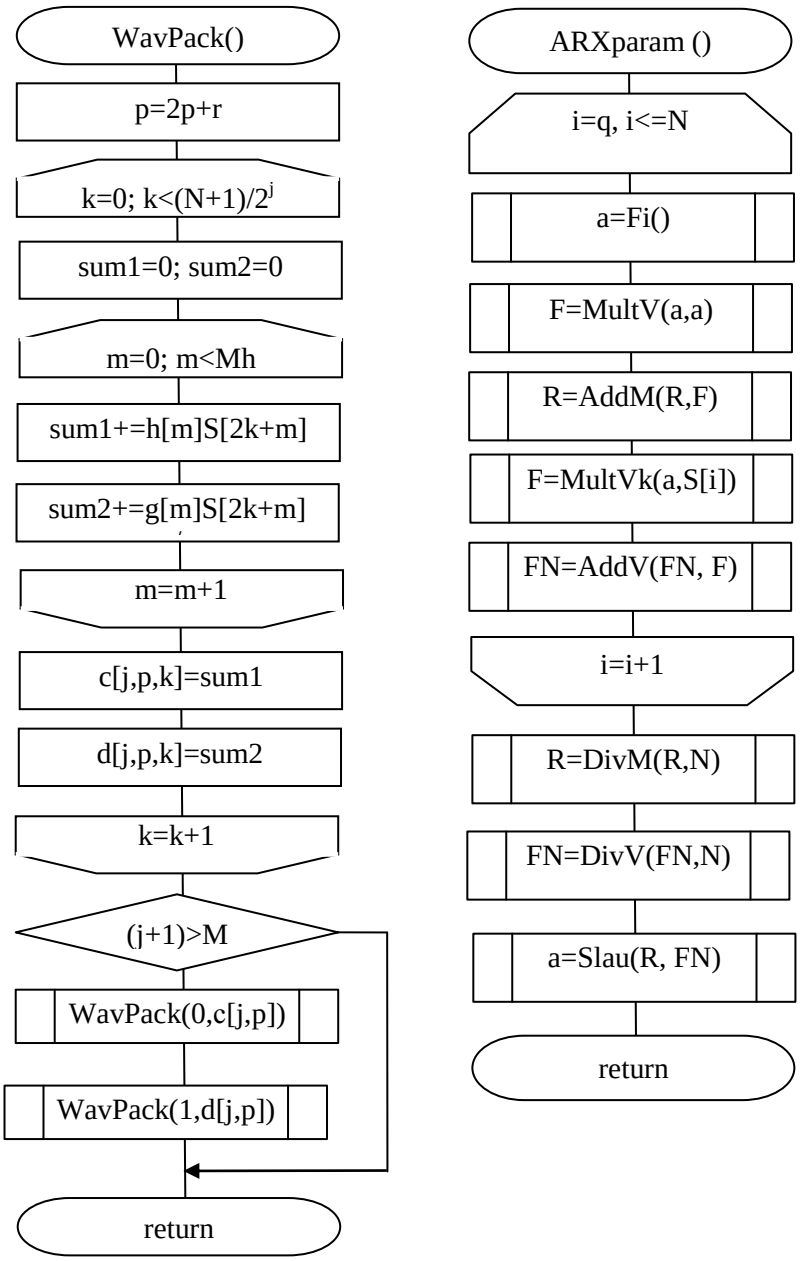


Рисунок 9.6 – Схема алгоритма функции нахождения коэффициентов вейвлет разложения сигнала на определенном уровне WavPack и расчёта параметров ARX-модели

В функции Solution происходит расчет коэффициентов разложения вейвлет-анализа, получение восстановленных значений трафика и определение уровня угрозы безопасности (методика описана в главе 10).

В функции WavPack происходит расчет всех аппроксимирующих и детализирующих коэффициентов по алгоритму вейвлет-пакетов.

Затем прогнозируется высокочастотная составляющая трафика, в результате чего определяются значения аномальной составляющих трафика, а также полный восстановленный массив.

По значению аномальной составляющей трафика происходит определение уровня угрозы безопасности с помощью функции OpredAlarm.

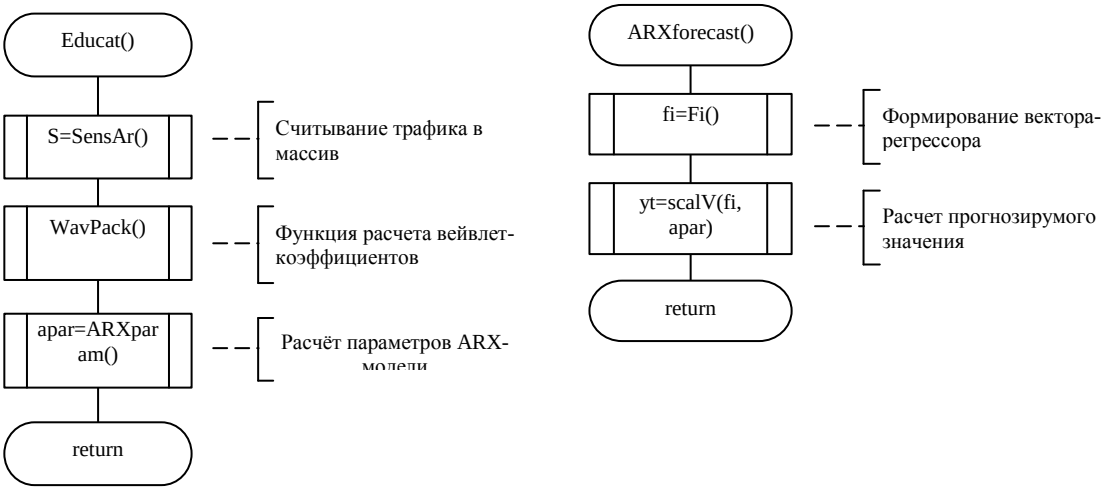


Рисунок 9.7 – Алгоритмы расчёта параметров ARX – модели и прогнозирования

Таким образом, предложенные алгоритмы позволяют в реальном масштабе времени проводить мониторинг нестационарных во времени или неоднородных в пространстве информационных процессов с целью выявления аномалий трафика ККС.

Вопросы для самоконтроля

1. Из каких этапов состоит процесс моделирования нормального трафика

сети?

2. Поясните этапы процесса моделирования нормального трафика сети.

3. Какие существуют базисные вейвлеты?

4. Какие базисные вейвлеты целесообразно использовать для мониторинга сетевого трафика?

5. В чем состоит преимущество базисной системы вейвлетов – койфлеты-2?

6. Поясните алгоритм мониторинга информационных процессов ККС.

7. Поясните алгоритм расчета коэффициентов вейвлет разложения сигнала.

8. Поясните алгоритм расчета аппроксимирующих и детализирующих коэффициентов по алгоритму вейвлет-пакетов.

9. Поясните алгоритм расчёта параметров ARX-модели

Глава 10 Методика обоснования порога аномального состояния трафика корпоративной сети

Результатом мониторинга информационных процессов ККС являются данные сетевого трафика. Основная цель обработки данных сетевой трафика – получение результата измерения, оценка его погрешности и расчет порогового значения, за которым информационный процесс признаётся аномальным.

В литературе достаточно полно изложены методы обработки экспериментальных данных [4,5,6,10,12,16]. Выбор метода обработки данных сетевой активности зависит от вида распределения погрешностей измерений**,** вида измерений и требований к быстроте вычислений или трудоемкости. При выборе метода для программной обработки данных важное значение имеют быстрота получения результата измерения и вычислительная трудоемкость метода. Поэтому в качестве результата измерений используют математическое ожидание, а в качестве характеристики его погрешности – среднее квадратическое отклонение.

Для обоснования порога аномального состояния ККС необходимо проанализировать экспериментальные данные мониторинга информационных процессов экспериментального участка корпоративной сети, разработать методику и алгоритмы расчета порога аномального состояния ККС.

10.1 Проверка гипотезы о виде распределения результатов мониторинга информационных процессов сети

Проверить гипотезу о том, что распределение данных не противоречит теоретическому распределению, можно по ряду критериев: Колмогорова, *w*-критерий и χ^2 – критерий – критерий Пирсона [11,13].

Число экспериментальных данных мониторинга информационных процессов ККС значительно больше 50, поэтому для проверки критерия согласия теоретического распределения с экспериментальным целесообразно использовать критерий Пирсона.

Результаты преобразования данных сенсора сводятся в файл. При этом данные группируются, вычисляются середины интервалов χ_i^2, соответствующие им эмпирические частоты, определяются математическое ожидание (МО) m и среднеквадратическое отклонение σ (выборочные значения).

При обработке экспериментальных данных мониторинга возникают особенности, связанные со значительным объемом информации о состоянии информационных процессов ККС. Поэтому необходимо так организовать фиксацию и обработку результатов мониторинга, чтобы оценки для искомых характеристик формировались в реальном масштабе времени, т.е. без специального запоминания всей информации о состоянии процесса. Так при обработке экспериментальных данных можно подойти к оценке возможных значений случайной величины, т.е. закона распределения. Область возможных значений случайной величины разбивается на ***r*** интервалов. Затем накапливается количество попаданий случайной величины в эти интервалы n_i, $i = 1,\ldots,r$.

Число данных n_i, которое должно быть в i-ом интервале, если бы их рас-

пределение соответствовало предполагаемому (теоретическая частота), определяется зависимостью [11]

$$n_i = n \cdot \frac{\mathrm{h}}{\sigma_i} \varphi(x^{н}{}_i), \qquad (10.1)$$

где n – объем выборки (сумма всех частот);

h – шаг (разность между двумя соседними вариантами);

$x_i^{н}$ – нормированное значение отсчетов, $x^{н}{}_i = \frac{x_i - \overline{x_в}}{\sigma_в}$;

σ_i – выборочное среднекрадратичекое отклонение,

$$\varphi(x^{н}{}_i) = \frac{1}{\sqrt{2\pi}} e^{-x^{н}{}_i{}^2/2}.$$

В качестве оценки дисперсии случайной величины $\boldsymbol{x}$ **,** как правило, используется известная зависимость

$$\sigma_x^2 = \sum_{i=1}^{r} (x_i - m_x)^2 / r. \qquad (10.2)$$

Непосредственное вычисление по (10.2) нерационально, т.к. среднее значение $\bar{x}$ изменяется в процессе накопления значений $\boldsymbol{x_i}$. Поэтому более рационально организовать фиксацию результатов мониторинга для оценки дисперсии по разработанному для программирования рекуррентному алгоритму вида

$$\sigma_x^2 = \left[\sum_{i=1}^{r} x_i^2 + \left(\sum_{i=1}^{r} x_i \right)^2 / r \right] / (r-1). \qquad (10.3)$$

Тогда для вычисления дисперсии достаточно накапливать только две суммы, что приведет к экономии вычислительные ресурсы.

Аналогично при оценке МО целесообразно применять для расчета при-

ближенную асимптотическую зависимость вида.

$$m_x = (\Delta t / T) \sum_{i=1}^{T/\Delta t} x(t_i), \qquad (10.4)$$

где T – интервал мониторинга;

Δt – шаг квантования.

Для каждого интервала вычисляется

$$\chi_i^2 = \frac{(\tilde{n}_i - n_i)^2}{n_i}. \qquad (10.5)$$

Просуммировав (10.5) по всем *r* интервалам, определяется наблюдаемое значение критерия с заданным числом степеней свободы *k*

$$\chi_{набл}^2 = \sum_{r=1}^{r} \frac{(\tilde{n}_i - n_i)^2}{n_i}. \qquad (10.6)$$

Для нормального распределения $k = r - 3$.

По таблицам распределения χ^2 с заданным уровнем значимости α и числом степеней свободы k находится критическая точка $\chi_{кр}^2$.

Гипотеза о соответствии теоретического распределения экспериментальному принимается, если $\chi_{кр}^2 < \chi_{набл}^2$, т.е. теоретическое и практическое распределения различаются незначимо (случайно). В противном случае гипотеза отвергается.

В процессе опытной эксплуатации программной системы «Анализатор Аномальности -2» на базе ККС «ТБинформ», г. Оренбург получены экспериментальные данные мониторинга информационных процессов (гистограмма частот поступления информации для переменной X_2 представлена на рисунке 10.1), подтвердившие гипотезу о нормальности закона распределения при уровне значимости $\alpha = 0{,}05$.

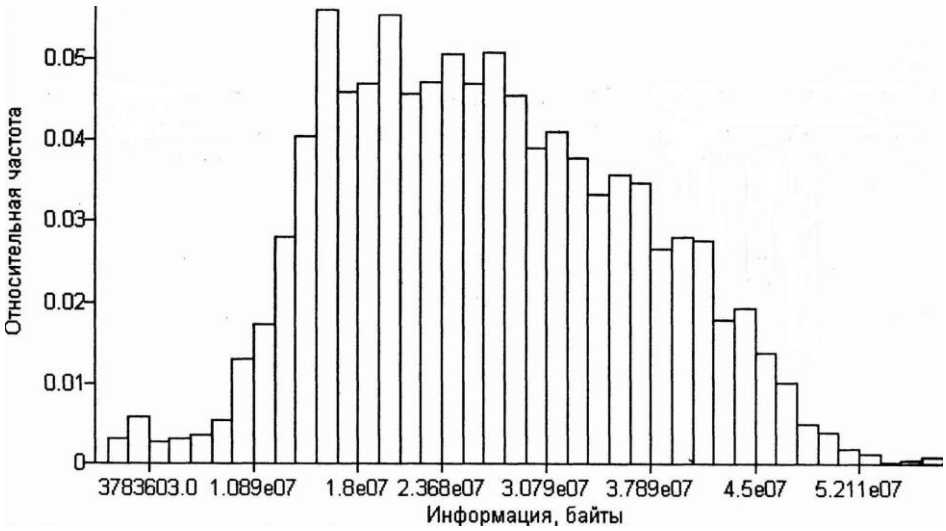


Рисунок 10.1 – Гистограмма частот поступления информации

Таким образом, проверка гипотезы о виде распределения результатов мониторинга информационных процессов ККС свидетельствует о возможности дальнейших исследований по обоснованию порогового уровня аномального поведения трафика сети с учетом нормальности законов распределения случайных составляющих информационного процесса экспериментальной ККС.

10.2 Методика обоснования порогового уровня аномального состояния трафика сети

В основу обоснования положен метод статистических решений для задачи проверки двухальтернативной гипотезы: H_0 и H_1 выражают предположения об отсутствии или наличии аномалии на текущем уровне сетевого трафика $f_{dp}(t)$ ККС [11].

Для того, чтобы задача обнаружения аномалий обрела математическую содержательность введены показатели – вероятности ложной тревоги $p_{лт}$ и пропуска аномалии $p_{па}$, понимая под ложной тревогой факт решения $\hat{H}_1$ об обнаружении аномалии при условии, что в наблюдаемом $f_{dp}(t)$ аномалия отсутствует, а под пропуском аномалии – принятие решения $\hat{H}_0$ о том, что аномалии в $f_{dp}(t)$ нет при условии, что в действительности она имеет место.

Исходя из того, что реальный сетевой трафик $f(t)$ является суперпозицией

большого числа некоторых элементарных случайных процессов, на основании центральной предельной теоремы теории вероятностей (утверждение о нормализации суммы случайных слагаемых с произвольными плотностями вероятности (ПВ) по мере увеличения их числа) и используя результаты мониторинга информационных процессов доказано, что ПВ сетевого трафика удается аппроксимировать нормальным законом.

Отсюда вероятности ошибок p_{na}, $p_{лт}$ определяются исходя из графического представления двух нормальных случайных процессов, представленных на рисунке 10.2, где площади заштрихованных областей равны $p_{лт}$ (косая штриховка) и p_{na} (прямая штриховка).

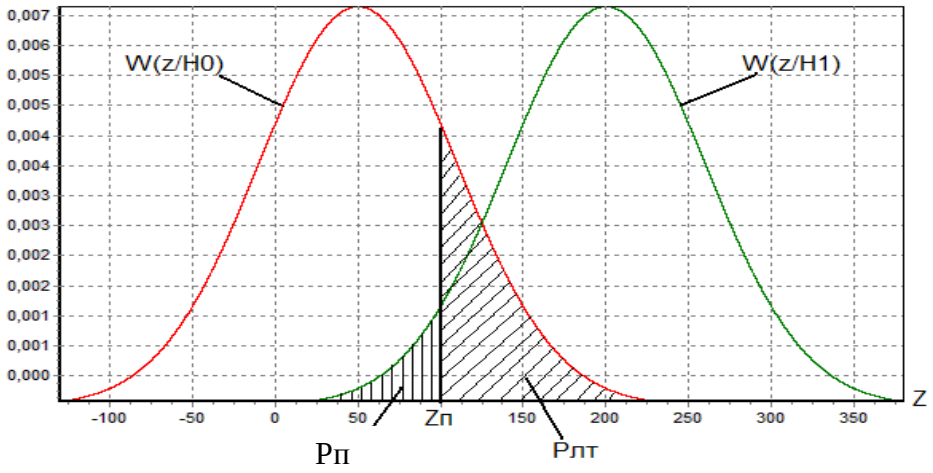


Рисунок 10.2 – Методика определения вероятностей $p_{лт}$ и p_{na}

Тогда p_{na}, $p_{лт}$ будут определяться зависимостями вида

$$p_{лт} = P(\hat{H}_1 \mid H_0) = P(z \geq z_{П} \mid H_0) = \int_{z_n}^{\infty} W(z \mid H_0) dz, \qquad (10.7)$$

$$p_{na} = P(\hat{H}_0 \mid H_1) = P(z < z_{П} \mid H_1) = \int_{-\infty}^{z_n} W(z \mid H_1) dz, \qquad (10.8)$$

где $z = \xi(t_i)$ – погрешности прогноза, определяющие степень сходства наблюдаемой реализации $f_{d^p}(t)$ с прогнозируемым состоянием сетевого трафика $\tilde{f}_{d^p}(t_i)$;

z_n – пороговый уровень аномальности сетевого трафика;

$W(z/H_1) = P(Z < Zn/H_i)$ - плотность вероятности корреляции z при ги-

потезе H_i, i = 0,1;

$\boldsymbol{T}$ – интервал наблюдения.

$W(z/Hi)$ – одномерные нормальные плотности вероятностей. Определив их параметры: математическое ожидание $\bar{z}$ и дисперсию $D\{z\}$, расчетные зависимости искомых вероятностей примут вид

$$p_{лт} = \frac{1}{\sqrt{2\pi}} \int_{z_n}^{\infty} \frac{1}{\sqrt{D(z)}} \exp\left(-\frac{z^2}{2D(z)}\right) dz = 1 - \Phi(h), \qquad (10.9)$$

$$p_{па} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_n} \frac{1}{\sqrt{D(z)}} \exp\left(-\frac{(z-\bar{z})^2}{2D(z)}\right) dz = \Phi(h-q) \ldots\ldots\ldots\ldots (10.10)$$

где $\Phi(x) = (1/\sqrt{2\pi} \int_{-\infty}^{x} \exp(-k^2/2dk$ – интеграл вероятности при $k = z(\sqrt{D(z)})^{-1}$;

$h = z_n(\sqrt{D(z)})^{-1}$ – нормированный пороговый уровень;

$q = z(\sqrt{2\bar{z}/N_0})^{-1}$ – параметр обнаружения, равный соотношению атака/шум.

С помощью соотношений (10.9), (10.10) рассчитывается z_n в соответствии с принятым критерием оптимальности (критерий Неймана – Пирсона) – необходимо минимизировать $p_{па}$ при фиксированном значении $p_{лт}$: для целей настоящего исследования $p_{лт} \leq 0{,}05$. Полученные оценки $p_{па}$ и $p_{лт}$ при этих условиях обеспечивают оптимальную величину среднего риска.

Переформулируем критерий оптимальности Неймана – Пирсона в форме условной минимизации целевой функции $p_{па} + \mu\,(p_{лт} - 0{,}05)$, где μ – неопределенный множитель Лагранжа, решение которой возможно на основе методов

нелинейного программирования путем отыскания условных (в заданных областях аргументов) экстремумов функций многих переменных.

В основу решения задачи условной оптимизации положена теорема Куна-Такера. Функция Лагранжа запишется в виде [11,13]:

$$L(h,\mu) = p_{па} + \mu(0{,}05 - p_{лт}) = \Phi(h-q) + \mu(0{,}05 - (1-\Phi(h)) = \\ = \Phi(h-q) + \mu(\Phi(h) - 0{,}95) \quad , \quad (10.11)$$

где ограничение критерия Неймана – Пирсона приведены к виду $0{,}05 - p_{лт} \geq 0$.

Условия Куна-Такера примут вид следующей системы уравнений:

$$\begin{cases} \begin{cases} \dfrac{\partial L}{\partial h} \geq 0 \\ h \cdot \dfrac{\partial L}{\partial h} = 0 \end{cases} \\ \begin{cases} \dfrac{\partial L}{\partial \mu} \geq 0 \\ \mu \cdot \dfrac{\partial L}{\partial \mu} = 0 \end{cases} \end{cases} \Rightarrow$$ Решения частных производных функции (10.11) приве-

дут к системе вида:

$$\begin{cases} \begin{cases} \Phi(h-q)'_h + \mu(\Phi(h) - 0{,}95)'_h \geq 0 \\ h \cdot (\Phi(h-q)'_h + \mu(\Phi(h) - 0{,}95)'_h) = 0 \end{cases} \\ \begin{cases} \Phi(h) - 0{,}95 \geq 0 \\ \mu \cdot (\Phi(h) - 0{,}95) = 0 \end{cases} \end{cases} . \quad (10.12)$$

Дальнейшие преобразования (10.12) позволяют получить систему с двумя условиями а) и б):

$$\begin{cases} \begin{cases} e^{-\frac{(h-q)^2}{2}} + \mu e^{-\frac{h^2}{2}} \geq 0 \\ h(e^{-\frac{(h-q)^2}{2}} + \mu e^{-\frac{h^2}{2}}) = 0 \qquad a) \end{cases} \\ \begin{cases} \Phi(h) \geq 0{,}95 \\ \mu(\Phi(h) - 0{,}95) = 0 \qquad b) \end{cases} \end{cases} \quad \ldots\ldots\ldots \quad \ldots(10.13)$$

Исследование системы уравнений (10.13) осуществляется путем выполнение проверки условий а) и б) при всех возможных значений h и μ.

1 Пусть $h = 0;\ \mu = 0$, тогда система (9.13) принимает вид:

$$\begin{cases} \begin{cases} e^{-\frac{q^2}{2}} + 0 \geq 0 \\ 0(e^{-\frac{q^2}{2}} + 0) = 0 \qquad a) \end{cases} \\ \begin{cases} \Phi(0) \geq 0{,}95 \\ 0(\Phi(0) - 0{,}95) = 0 \qquad b) \end{cases} \end{cases}.$$

При этом не выполняются условия *b*) системы (10.13), а именно ложно условие $\Phi(0) \geq 0{,}95$.

2 Пусть $h \neq 0;\ \mu = 0$, тогда система (10.13) принимает вид:

$$\begin{cases} \begin{cases} e^{-\frac{(h-q)^2}{2}} \geq 0 \\ h(e^{-\frac{(h-q)^2}{2}}) = 0 \qquad a) \end{cases} \\ \begin{cases} \Phi(h) \geq 0{,}95 \\ 0(\Phi(h) - 0{,}95) = 0 \qquad b) \end{cases} \end{cases}.$$

При этом не выполняются условия *a)* системы (10.13), а именно ложно условие $h(e^{-\frac{(h-q)^2}{2}}) = 0,\ h \neq 0$.

3 Пусть $h = 0;\ \mu \neq 0$, тогда система (10.13) принимает вид

$$\begin{cases} \begin{cases} e^{-\frac{q^2}{2}} + \mu \geq 0 \\ 0(e^{-\frac{q^2}{2}} + \mu) = 0 \qquad a) \end{cases} \\ \begin{cases} \Phi(0) \geq 0{,}95 \\ \mu(\Phi(0) - 0{,}95) = 0 \qquad b) \end{cases} \end{cases},$$

При этом не выполняются условия *b*) системы (10.13), а именно ложно условие $\Phi(0) \geq 0{,}95$ и $\mu(\Phi(0) - 0{,}95) = 0,\ \mu \neq 0$.

4 Пусть $h \neq 0;\ \mu \neq 0$, тогда система (10.13) принимает вид

$$\begin{cases} \begin{cases} e^{-\frac{(h-q)^2}{2}} + \mu e^{-\frac{h^2}{2}} \geq 0 \\ e^{-\frac{(h-q)^2}{2}} + \mu e^{-\frac{h^2}{2}} = 0 \qquad a) \end{cases} \\ \begin{cases} \Phi(h) \geq 0{,}95 \\ \Phi(h) - 0{,}95 = 0 \qquad b) \end{cases} \end{cases}.$$

При этом все условия теоремы Куна-Такера выполнимы одновременно и из системы *a*) будет получено равенство вида

$$e^{-\frac{(h-q)^2}{2}} + \mu e^{-\frac{h^2}{2}} = 0. \qquad (10.14)$$

После преобразования (10.14) получено выражение для расчета неопределенного множителя Лагранжа вида [11,13]

$$\mu = -e^{\frac{2hq-q^2}{2}}. \qquad (10.15)$$

Из условия системы *b*) получено равенство $\Phi(h) - 0{,}95 = 0$, тогда нормированный пороговый уровень активности субъектов сети примет значение $h = 1{,}64485$ [11].

Таким образом, значение минимума функция $p_{па}$ при условии $p_{лт} \leq 0{,}05$

принимается в точке $h = 1.64485$, причем значение минимума меняется в зависимости от параметра обнаружения q.

Истинное пороговое значение z_n рассчитывается из принятого ранее условия определения нормированного порогового уровня h, т.е.

$$z_n = h(\sqrt{D(z)}) \quad . \qquad (10.16)$$

Тестовая проверка разработанной методики обоснования порогового уровня аномальной активности субъектов сети выполнена с использованием пакета MATCAD. Анализ результатов тестирования позволяет оценить влияние параметра обнаружения q (соотношение атака/шум) на оценки вероятностей $p_{па}$ и $p_{лт}$ и свидетельствует:

– при росте соотношения атака /шум вероятность пропуска $p_{па}$ атаки снижается;

– при превышении критического значения нормированного порогового уровня h критерий Неймана-Пирсона выполняется при меньших вероятностях ложной тревоги $p_{лт}$

Таким образом, разработанная методика обоснования порога аномальной активности субъектов ККС является развитием методов распознавания теории статистических решений в задаче обнаружения угроз информационной безопасности ККС.

10.3 Разработка алгоритмов расчета порогового уровня аномальности трафика корпоративной сети

Для создания приложений выбрана среда разработки приложений NetBeans, предоставляющее множество компонентов для работы с базами данных и сетевыми ресурсами, что отвечает требованиям решаемой задачи.

На рисунке 10.3 представлена схема алгоритма определения уровня угрозы безопасности OpredAlarm() [13].

Рисунок 10.3 – Схема алгоритма определения уровня угрозы ИБ OpredAlarm()

Алгоритм определения уровня угрозы ИБ OpredAlarm() является составной частью укрупненной схемы алгоритма функции расчета коэффициентов вейвлет-преобразования сетевого трафика Solution() в ПС «Анализатор Аномальности - 2» [13].

В функции OpredAlarm происходит сравнение текущего уровня аномалии сетевого трафика с рассчитанным на текущий промежуток времени пороговым значением аномалии. В случае если зафиксировано превышение значения, выдается сообщение об угрозе и предпринимаются определенные в настройках программы меры по устранению угрозы безопасности. В качестве данных мер могут выступать: блокирование канала передачи на определенный промежуток времени, закрытие всех текущих сеансов с атакующим узлом, передача преду-

преждающего сообщения на атакуемый узел, либо обеспечение возможности игнорирования атаки. В случае если значение текущего уровня аномалии не превышает порогового, то фиксируется нормальный режим работы сети.

В схеме укрупненного алгоритма мониторинга информационных процессов ККС включен модуль *SrF*, обеспечивающий расчет порогового значения аномального состояния трафика сети. Схема алгоритма *SrF* представлена на рисунке 10.4.

Результаты реализации алгоритма *SrF* при различных значениях параметр обнаружения q в ПС «Анализатор Аномальности - 2» [13] представлен на рисунке 10.5.

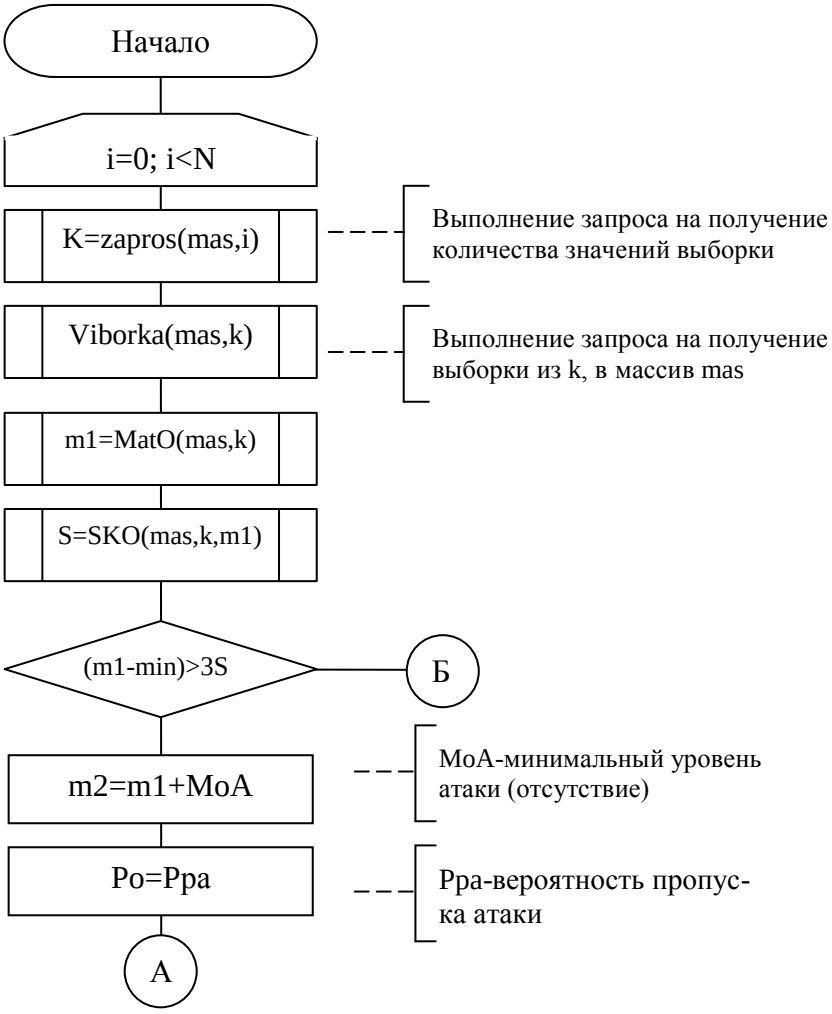


Рисунок 10.4 – Схема алгоритма расчета порогового значения SrF (лист1)

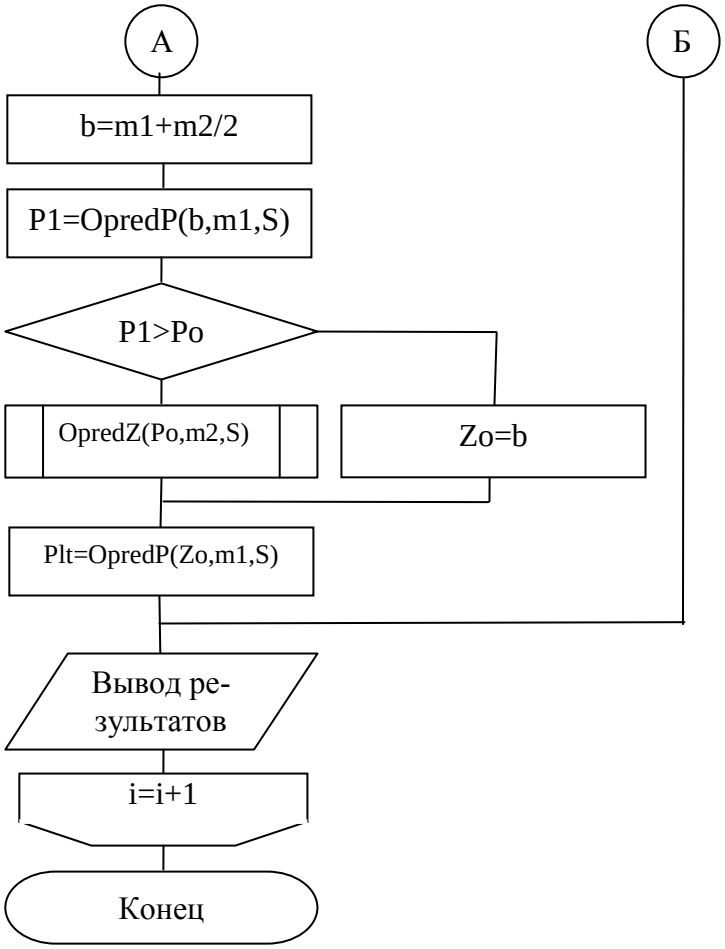


Рисунок – 10.4 (лист 2)

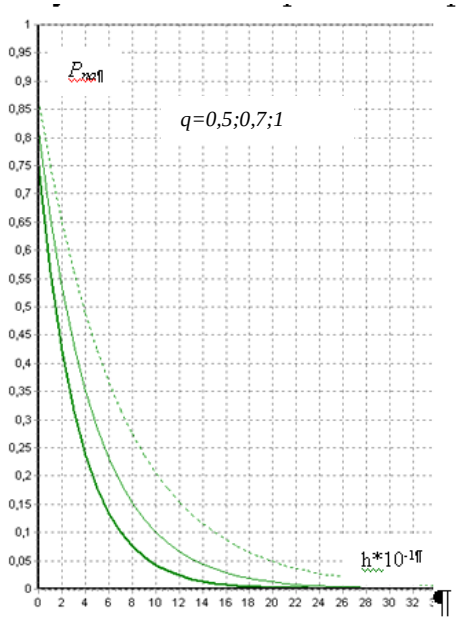


Рисунок 10.5 – Результаты расчета порогового уровня сетевого трафика

Результаты расчета порогового уровня сетевого трафика свидетельствуют, что с применением предложенной технологии принятия решений следует

ожидать обеспечение вероятности обнаружения вторжений на уровне 0,78 – 0,88 при вероятности ложной тревоги не более 0,05, причем положительный эффект новой технологии принятия решений усиливается по мере увеличения единого информационного поля ККС [13].

Таким образом, предложен алгоритм статистического обоснования порогового уровня аномального состояния трафика сети.

Вопросы для самоконтроля

1. Какие задачи нужно решить для обоснования порога аномального состояния ККС?
2. Как можно проверить гипотезу о том, что распределение данных не противоречит теоретическому распределению?
3. Какой критерий согласия нужно применить для проверки гипотезы о соответствии теоретического распределения экспериментальному для данных мониторинга информационных процессов ККС?
4. Для чего нужна проверка гипотезы о виде распределения результатов мониторинга информационных процессов ККС?
5. Что означают показатели вероятности ложной тревоги и пропуска аномалии в задаче обнаружения?
6. Почему плотность вероятностей сетевого трафика удается аппроксимировать нормальным законом?
7. Что определяет степень сходства наблюдаемой реализации с прогнозируемым состоянием сетевого трафика?
8. Как можно решить задачу условной минимизации?
9. В чем заключается суть теоремы Куна-Таккера?
10. Опишите алгоритм определения уровня угрозы безопасности
11. Какие можно принять меры реагирования на сетевые аномалии?
12. Принцип принятия решения о нормальном состоянии трафика сети?

Раздел 3 Исследование операций в задачах цифровой обработки электронных сообщений

Глава 11 Моделирование текстового контента электронных сообщений

В основе модели описания текстового контента электронных почтовых сообщений положен контент-анализ [15]. Под контент-ана́лизом (от англ.: contents - *содержание, содержимое*) понимается методика исследования, предметом которого является содержание текстовых массивов и продуктов коммуникативной корреспонденции. Согласно определению выделяют количественный и качественный контент-анализ текстов для последующей их обработки, анализа и определения числовых закономерностей.

Количественный контент-анализ (по содержанию) предназначен для анализа отдельных слов, словосочетаний, предложений сообщения.

Качественный контент-анализ (структурный) позволяет определять не ***что*** говориться в тексте, т.е. его содержание, а ***как*** отражается этот объект в тексте, не уделяя внимания анализу самого содержания.

Известно несколько способов описания текста [1,3,8,15,17], основным из которых являются векторная и графовая модель представления текста.

11.1 Векторная модель текстового контента

Текст электронного почтового сообщения можно представить в виде термов, что позволяет описать каждый документ в виде вектора в пространстве признаков.

Графическое отображение документа L состоящее из трех термов t_1, t_2, t_3 представлено на рисунке 11.1.

Трехмерное отображение документа, состоящее из трех термов может быть распространенно и на N-мерные документы, где N – количество термов в документе.

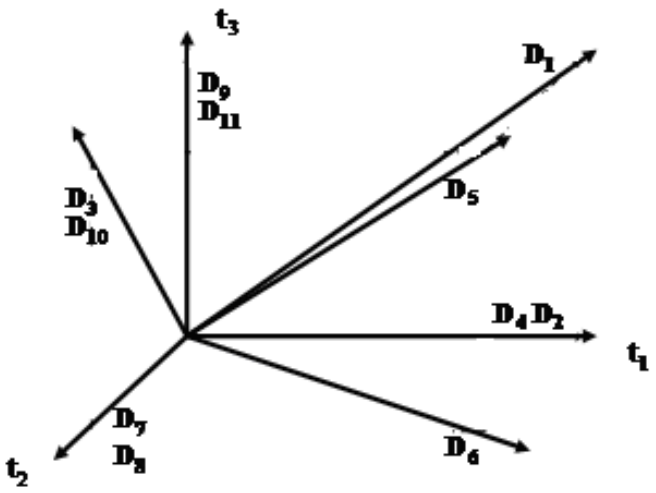


Рисунок 11.1– Пример графического отображения документа

В рамках векторной модели электронного сообщения (ЭС) описывается в некотором пространстве признаков, в котором каждому используемому в сообщении терму ставится в соответствие его вес:

$$L_i = (w_{1j}, \ldots, w_{|T|j}), \qquad (11.1)$$

Векторная модель ЭС представляется в виде матрицы столбца L_i, элементами которой являются веса соответствующих термов в сообщении.

$$L_i = \begin{bmatrix} w_{1j} \\ \vdots \\ w_{ij} \\ \vdots \\ w_{Mj} \end{bmatrix} \qquad (11.2)$$

где w_{ij} – вес терма j в сообщении i

M – число термов(слов) в сообщении

Если одно ЭС можно представить в виде 11.2, то вся коллекция сообщений примет вид матрицы столбцами которой будут являться письма, а строками термы содержащиеся в данных письмах.

$$Let = \begin{bmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1N} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{M1} & \cdots & w_{Mj} & \cdots & w_{MN} \end{bmatrix}, \quad (11.3)$$

где M – количество писем в выборке,

N – число термов(слов) в коллекции после удаления стоп-слов,

ω_{ij} – вес терма j в сообщении i;

$i = 1,...,M$, $j = 1,...,N$.

Основным достоинством векторной модели является возможность использования алгоритмов классификации, основанных на анализе статистических характеристик, а так же возможность сравнивать вектора в векторном пространстве признаков. Задача преобразования текста в вектор в пространстве признаков требует определения координаты признаков. Самым распространенным методом определения значимости терма является логическое взвешивание, заключающееся в том, что терму присваивается значение «1», если терм встречается в сообщении и «0» - в случае, если терм в сообщении не встречается.

Основным недостатком данной меры является то, что такой подход не учитывает частоту встречаемости отдельного терма во всей коллекции документов. Таким образом, при оценке близости векторов с использованием данной мерой значимости термов результаты могут быть не всегда удовлетворительными [15,17].

11.2 Модель представления текста на основе графа

Графовая модель представления текста заключается в выделении объектов, которые являются вершинами графа, и отношений между данными объектами (связи), т.е. ребрами графа. В качестве объектов могут выступать как от-

дельные слова, так и понятия. Графическое отображение модели представления текста показано на рисунке 11.2

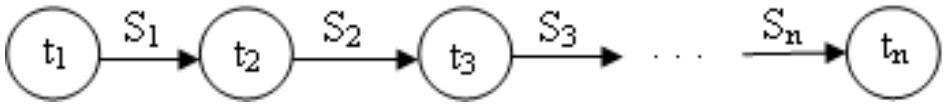


Рисунок 11.2 – Представление текста на основе графа (t_1,t_n – слова в тексте (вершины графа), S_1,S_n –мера между словами (дуги графа)

Использование понятий в качестве вершин графа требует ведения понятийных словарей. Вид графа в таком случае принимает вид показанный на рисунке 11.3

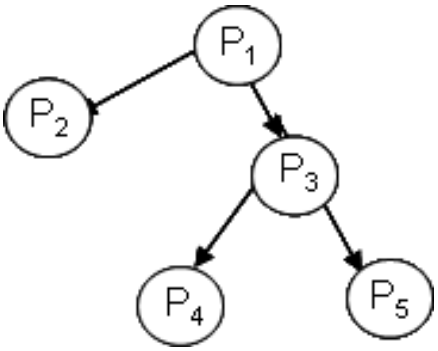


Рисунок 11.3 – Представление текста на основе графа – дерева зависимостей (P_n – понятия в тексте, O_n – отношение между понятиями)

Графовые модели текста часто представляются семантическими сетями, имеющими вид ориентированного графа, вершинами которого являются объекты предметной области, а ребра определяют отношение между ними [3,7,15]. Семантическая сеть отражает семантику предметной области в виде понятий и отношений. Семантический анализ позволяет строить семантическую структуру как одного предложения, так и всего текста.

В качестве вершин в семантических сетях используются понятия базы знаний, отношение между этими понятиями отображаются ребрами. В более сложном представлении текста применяют грамматические связи между словами или понятиям. Результатом обработки текста является граф грамматических связей.

Выбор конкретной формы графа определяется методом, применяемым для последующей обработки, основной целью и требуемым временем на обработку графа [1,3,8,17].

Сравнительный анализ векторного представления текста и описания на основе графа. Для проведения исследований взято 100 сообщений (50 спам сообщений и 50 легитимных сообщений в обучающей выборке системы фильтрации для одного пользователя). Анализировались сообщения, количество термов которых составляет от 0 до 20, от 20 до 50, от 50 до 70, от 70 до 100, и свыше 100 слов в сообщении. Результатом анализа является затраченное на обработку время – для векторной и графовой модели в качестве веса термов принята мера логического взвешивания. Логическое взвешивание заключается в присвоении весу значения 1, если слово встречается в документе и 0 в противном случае, т.е. $w_{ji} = \begin{cases} 1, & f_{ji} > 0 \\ 0 & f_{ji} = 0 \end{cases}$,

Так как графовая модель представления текста дополнительно требует определение связей между термами, то для этого использовалась мера Дайса. Результаты сравнения представлены на рисунке 11.3.

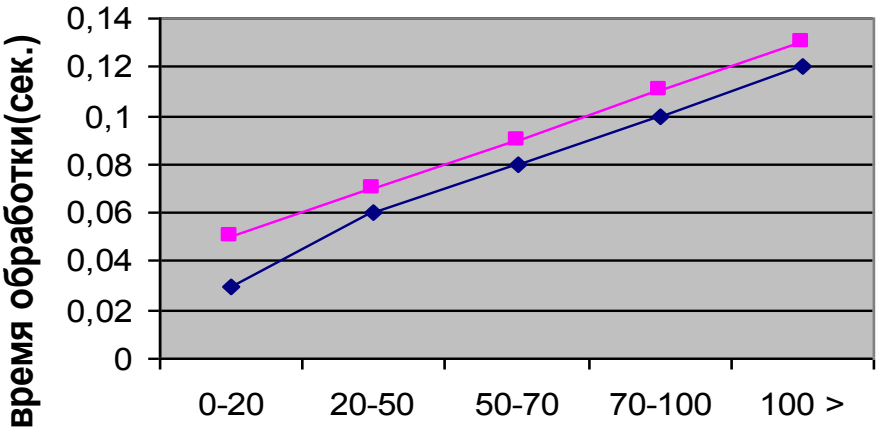


Рисунок 11.3 – Результаты сравнения векторной и графовой модели описания текста

Векторная модель позволяет определить значимость коллекции сообщений в диапазоне от 0,03 до 0,1секунды. Графовая модель на определение веса терма и нахождение связей между ними затрачивает от 0,05 до 0,13 секунды.

Второй параметр, по которому происходило сравнение – размерность полученной матрицы. При таком сравнении векторная модель и модель на основе графа показали практически идентичные результаты. При векторном представлении в сообщениях, в которых количество термов порядка двадцати, размерность составляет 3250 термов на 100 сообщений, и при увеличении числа термов в сообщениях она достигает 12000 термов для исследуемых сообщений. Результат сравнения представлен на рисунке 11.4.

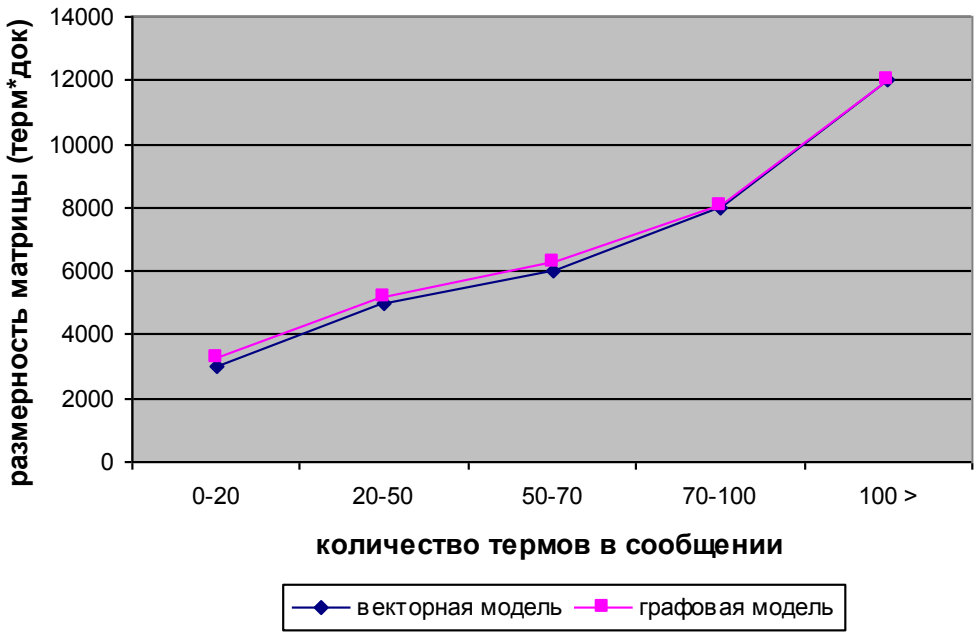


Рисунок 11.4 – Ограничения рассматриваемых моделей по размерности матрицы (терм*документ)

Анализ размерности базы термов занимаемой на жестком диске показал (рисунок 1.51), что векторная модель занимает 124 Кб при количестве слов в сообщении до 20, и возрастает до 385Кб при увеличении числа термов до 100. Графовая модель при количестве слов в сообщении до 20 занимает 155 Кб, и возрастает до 475Кб при увеличении числа термов до 100. При использовании графовой модели в служебной переписке, размерность баз возрастает многократно по сравнению с векторной моделью.

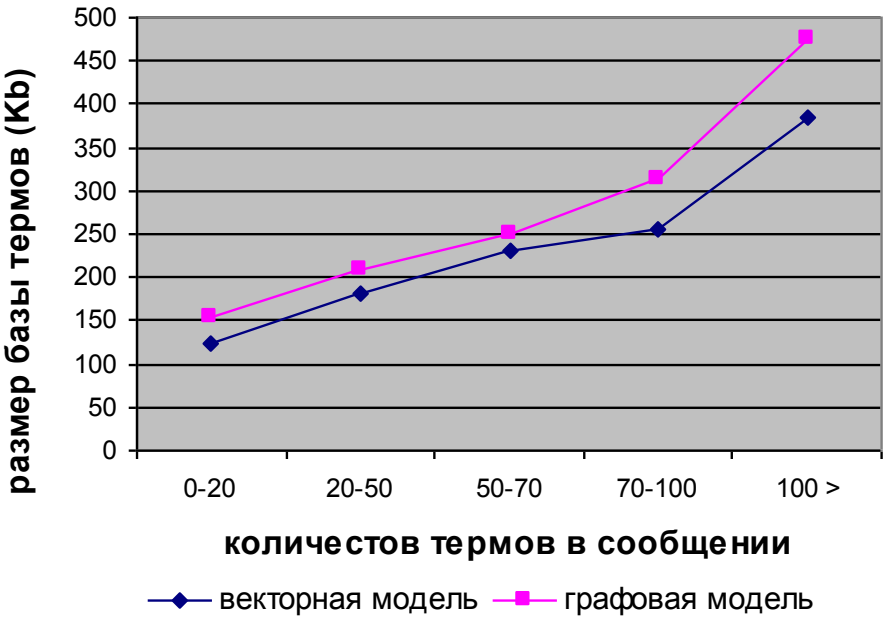


Рисунок 11.5 – Ограничения рассматриваемых моделей по размерности матрицы (терм*документ)

В результате проведенного анализа ограничений рассмотренных моделей определенно, что модель на основе графового описания текста требует значительных ресурсных и временных затрат на установление связей между термами. Кроме того, в случае более глубокого анализа текста при использовании графовой модели (в том случае если в качестве вершин используются понятия слов, или при необходимости выделения частей речи, падежей) временные и ресурсные затраты возрастают.

Таким образом, для дальнейших исследований используется векторная модель текстового описания ЭС.

11.3 Модель электронного сообщения в задаче классификации

При использовании векторной модели содержание ЭПС описывается с помощью термов t, множество которых будет образовывать тезаурус $T\{t_1,....t_q\}$ определенного класса k. В качестве термов используются слова, составляющие содержание сообщения. Анализ пространства признаков $S(p_i)$ представленных на рисунке 11.6, позволил выбрать в качестве признака вес терма.

Тогда модель ЭПС можно представить в виде

$$S(p_i)=<t_j,w(t_j)> \qquad (11.5)$$

где t – j-ый терм в сообщении;

p_i – пространство признаков, определяющих сообщение;

$w(t_j)$ – вес терма в сообщении после удаления стоп-слов.

Вся коллекция сообщений определенного класса L_k примет вид

$$L_k=<T_k ,w(t_j)> \qquad (11.6)$$

где T_k – k-ый тезаурус сообщения класса k;

$w(t_j)$ – вес терма в сообщении.

Рисунок 11.6 – Пространство признаков ЭС

Базовый подход к построению векторного представления текста при использовании логической меры значимости термов обладает следующим недостатком: при оценке близости векторов с использованием данной меры значимости термов результаты могут быть не всегда удовлетворительными. Кроме того на результат классификации также может влиять изменение количества термов в сообщении [1,3,8,15,17]. Преодоление данного недостатка возможно при изменении способа определения значимости терма. Следовательно, одной из основных задач при работе с текстовым содержимом

ЭПС становится вычисление весового коэффициента w_{iq}, определяющего значимость соответствующего терма t_j в i-ом документе.

Вопросы для самоконтроля

1 Поясните понятие контент – анализа?

2 В чем заключается количественный и качественный контент-анализ?

3 Опишите векторную модель представления текстового содержимого?

4 В чем заключается графовая модель представления текста. Основные достоинства и недостатки.

5 Перечислите ограничения графовой модели представления текста.

6 В чем заключается основной недостаток логической меры взвешивания

Глава 12 Методика формирования признаков классификации текста

12.1 Меры взвешивания термов в электронном сообщении

Существуют несколько различных мер определения значимости термов [3,8,15,17] (рисунок 12.1).

Рисунок 12.1 – Существующие меры взвешивания термов

Примем за f_{ij} - частоту терма t_j в сообщении S_i, N – число сообщений в выборке определенного класса, M – число термов в сообщениях класса k после удаления стоп-слов, n_j - общее количество сообщений, содержащих терм t_j.

Логическое взвешивание. Данная мера взвешивания основана на том, что присвоить терму значение 1 в том случае, если он встречается в документе и 0 в противном случае.

$$w_{ji} = \begin{cases} 1, & f_{ji} > 0 \\ 0 & f_{ji} = 0 \end{cases}.$$

Основным достоинством данного метода является простота реализации и использования. Однако, в качестве недостатка можно отметить, что при таком подходе никаким образом не учитывается важная информация о частоте встречаемости термина и отсутствует выделение информативных терминов.

TF – взвешивание (term freguency). Другим простым подходом является использование частоты слова в документе

$$w_{ij} = f_{ij},$$

TF – взвешивание выделяет в качестве информативных часто встречающиеся термины. Использование частоты слова дает примерно 25% увеличение эффективности классификации по сравнению с логическим взвешиванием.

TF - IDF взвешивание. Предыдущие два метода не используют частоту встречаемости термина во всех документах коллекции. TF-IDF- взвешивание присваивает вес слову *j* в документе *i* пропорционально числу вхождений слова в документ и обратно пропорционально числу документов в коллекции, в которые слово входит однажды. Логарифм в формуле используется для уменьшения веса часто встречающихся терминов и увеличения веса терминов которые встречаются редко. Таким образом, в TE-IDF мере взвешивания происходит выделение средне и низкочастотных терминов

$$w_{ij} = f_{ij} \log\left(\frac{N}{n_j}\right)$$

TFC-взвешивание. В TF-IDF методе взвешивания не учитывается, что документы могут быть различной длины, что существенным образом влияет на каче-

ство классификации TFC - взвешивание подобно TF-IDF - взвешиванию за исключением того, что используется нормализация длины документа.

$$w_{ij} = \frac{f_{ji} \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{l=1}^{M}\left[f_{li} \log\left(\frac{N}{n_l}\right)\right]^2}},$$

где n_l - общее количество документов исходной выборки, содержащих слово l.

TFC мера взвешивания, как и TF – IDF мера, выделяет средние и низкочастотные термы.

LTC – взвешивание. Данный подход заключается в использовании логарифма частоты слова вместо просто частоты слова и сокращает эффект больших различий в частотах

$$w_{ij} = \frac{\log(f_{ji}+1) \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{l=1}^{M}\left[\log(f_{li}+1) \log\left(\frac{N}{n_l}\right)\right]^2}}.$$

Таким образом, использование LTC меры позволяет сократить эффект больших различий в частотах, что делает использование данной меры взвешивания наиболее приемлемой.

Отсюда, сообщения, формирующие обучающую выборку, можно представить в виде матрицы, столбцами которой будут письма, а строками термы, содержащиеся в письмах:

$$L_k = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{j1} \\ w_{12} & w_{22} & \cdots & w_{j2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1\mathrm{i}} & w_{2\mathrm{i}} & \cdots & w_{ji} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1\mathrm{N}} & w_{2\mathrm{N}} & \cdots & w_{MN} \end{bmatrix}$$

где $w_{it_j} = Ltc_{it_j}$, $j = 1,\ldots,M$, $i = 1,\ldots,N$.

Получаемая матрица признаков ЭС имеет высокую размерность, обработка которой требует больших вычислительных ресурсов и времени.

Согласно законам Ципфа, слова, встречающиеся в тексте обучающей выборки чаще других, являются малоинформативными, не имеющими решающего смыслового значения, что становится основой снижения размерности матрицы признаков за счет избавления от малоинформативных термов без потери смыслового содержания ЭС.

Проведенный анализ частотности термов в сообщениях класса легитим и класса спам (рисунок 12.2) показал:

- всегда существуют термы, встречающиеся в одном классе, но не встречающиеся в другом классе. Термы, которые чаще всего встречаются в определенном классе, говорит о возможности данных термов характеризовать тот или иной класс;

- существуют термы, по которым трудно определить принадлежность к тому или иному классу, так как частота встречаемости данных термов в обоих классах различается незначительно.

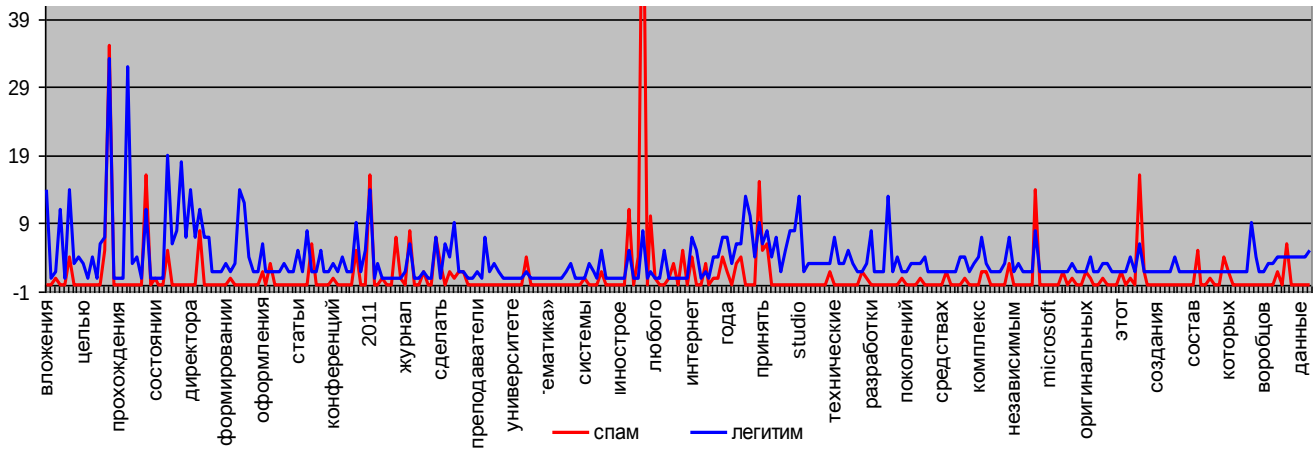


Рисунок 12.2 – Число термов по классам *(spam/legitim)* типового почтового электронного сообщения

Полученные выводы свидетельствуют о необходимости сокращения признакового пространства с целью избавления от малоинформативных термов и выделения термов способных характеризовать тот или иной класс.

12.2 Размерность пространства признаков текстовых документов

Для увеличения быстродействия алгоритмов классификации необходимо сокращение признакового пространства, известны следующие подходы [1,3,15,17]. Для их реализации известны методы многомерного статистического анализа, ориентированные на работу с текстовыми данными, такие как подсчет взаимной значимости термов, кластеризация термов относительно введенной метрики, выделение только тех термов, вес которых является максимальным.

Метод взаимной значимости термов позволят вычислить значимость термов в коллекции документов.

Данная мера рассчитывается по зависимости вида:

$$MI(t_j,k) = \sum P(t_j,k) * \log \frac{P(t_{j,}k)}{P(t_j) * P(k)} ,$$

где t_j – терм сообщения;

k – класс сообщения.

Данная мера основана на определении вероятности появления терма в сообщении. С точки зрения теории вероятности данная мера позволяет определить степень независимости термов. Сокращение числа термов происходит по установленному порогу. В качестве недостатка данной методики можно выделить зависимость значения меры *MI* от количества сообщений в классе, а также свойство данной меры завышать значимость термов, частота которых ниже по сравнению с другими термами сообщениях класса.

Метод Хи-квадрат. Статистический метод Хи-квадрат широко применяется как метод выбора признаков для работы с текстом. Критерий Хи-квадрат позволяет ранжировать признаки только по степени их полезности(важности) и не позволяет сделать вывод о статистической зависимости или независимости признаков. Критерий Хи-квадрат ориентированный для работы с текстом рассчитывается по зависимости вида:

$$\chi^2(t,k) = \frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)},$$

где A – количество сообщений, в которых *t* и *k* появились совместно;

B – количество сообщений, при которых слово *t* встречается с другим классом;

C – количество сообщений относящихся к классу *k*, но в которых не встречается *t*;

D – количество сообщений не относящихся к классу *k*, в которых нет слова *t*.

Метод главных компонент предназначен для сокращения размерности данных, при этом позволяет осуществить отбор наиболее информативных термов. Формально задача снижения размерности признакового пространства для некоторой выборки объектов $X = \{X_n\}$ размерности $n=1,\infty$ состоит в получении представления этой выборки в пространстве меньшей размерности, т.е. в приведении X к виду $X' = x_m$, $m<n$.

Однако метод не всегда эффективно снижает размерность при заданных ограничениях на точность. Прямые и плоскости не всегда обеспечивают хорошую аппроксимацию. Например, данные могут с хорошей точностью следовать какой-нибудь кривой, а эта кривая может быть сложно расположена в пространстве данных. В этом случае метод главных компонент для приемлемой точности потребует нескольких компонент (вместо одной), или вообще не даст снижения размерности при приемлемой точности.

Глобальный вес $RF_{t_q}^k$. *Для с*окращения признакового пространства предложен комбинированный подход, основанный на том, что для каждого терма в сообщениях определенного класса вычисляется величина $RF_{t_j}^k$, характеризующая значимость терма для определенного класса *k*:

$$RF_{t_j}^k = \log_2(2+\frac{a_i}{\max(1,b_i)}),$$

где a_i – количество ЭС, содержащих t_j-ый терм и относящихся к классу k;

b_i – количество ЭС, содержащих t_j-ый терм и не относящихся к классу k.

Термы, значимость которых $RF_{t_q}^k \le 1{,}5$, исключаются в сообщении класса k.

Сравнительный анализ методов сокращения признакового пространства показал: для 100 сообщений (50 сообщений спам тематики и 50 легитимных сообщений) и векторной модели ЭС с использованием значимости термов LTC, общее количество термов в сообщениях составляет 8723.

Компонентный анализ позволил сократить число термов на 1923, в результате чего количество термов стало 6800. Метод Хи-квадрат позволил сократить число термов всего на 1611, таким образом, количество термов стало составлять 7112. Метод на основе определения глобального веса $RF_{t_q}^k$ сократил число термов на 2566, и количество термов стало составлять соответственно 6157.

Другим признаком, по которому происходило сравнение рассматриваемых методов, является время, затраченное на сокращение размерности матрицы. Метод $RF_{t_q}^k$ позволяет сократить размерность термов за 17 секунд, компонентный анализ за 23 секунды, метод Хи-квадрат за 15 секунд и метод на основе определения информационной значимости за 16 секунд.

Результаты анализа сведены в таблицу 12.1 и представлены в виде диаграммы на рисунке 12.3.

Таблица 12.1 – Результаты работы методов

Вид проведенной работы	Название методов			
	RF	IG	Хи-квадрат	Компонентый анализ
размерность матрицы термов	6157	6200	7112	6800
время обработки (сек)	17	16	15	23

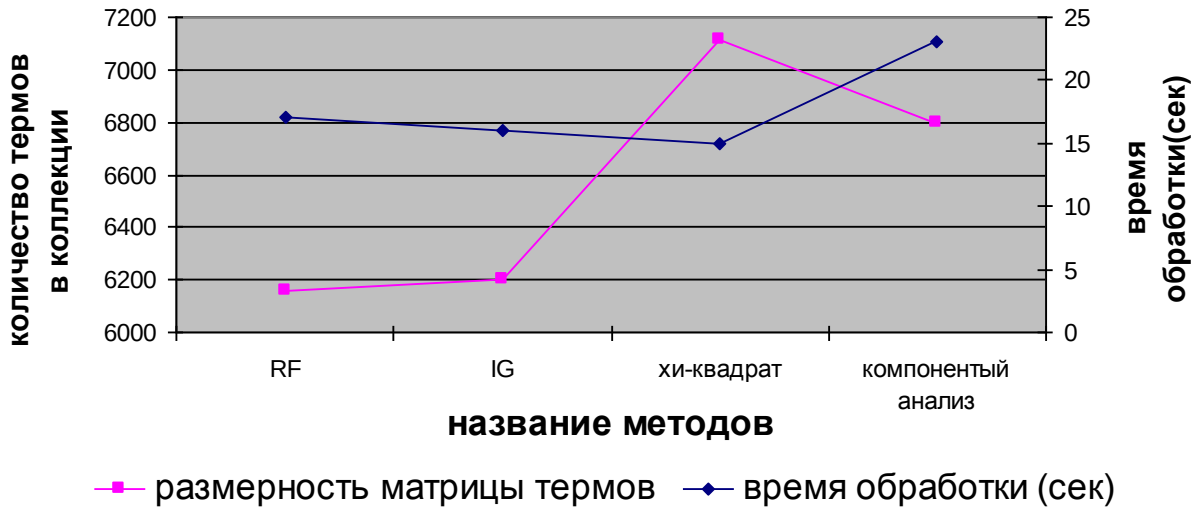


Рисунок 12.3 – Сравнительная характеристика методов

Таким образом, анализ показал, что больше всего на сокращение числа термов «затратил» компонентный анализ, при этом сократив количество термов с 8723 до 6800. Метод RF и IG показали близкие результаты, однако использование метода на основе определения глобального веса оправданно его результативностью.

12.3 Методика выявления устойчивых словосочетаний

В качестве примера выявления устойчивых словосочетаний рассмотрено легитимное электронное почтовое сообщение кафедры программного обеспечения вычислительной техники и автоматизированных систем (рисунок 11.4), на базе которой выполнялся эксперимент.

В соответствии с законом о персональных данных [15] адрес отправителя и получателя, а так же номера телефонов будут скрыты, а все личные данные (Ф.И.О.) заменены на Иванов Иван Иванович.

От: xxxxx@mail.osu.ru
Отправлено: 14 февраля 2012 г. 17:04
Кому: xxxx@unpk.osu.ru
Вложения: Информационные ресурсы научной библиотеки.doc

Уважаемые заведующие кафедрами!
С целью информированности об информационных ресурсах научной библиотеки, а также для успешного прохождения аккредитации в 2012 г. образовательных программ, рассылаю Вам справку о состоянии фонда библиотеки вуза.
С уважением,
Иванов Иван Иванович,
зам. директора Научной библиотеки ОГУ
тел.: (xxxx) xx-xx-xx; внутр. xx-xx

Рисунок 12.4 – Пример легитимного ЭПС

В таблице 12.2 приведены термы и значение глобального веса RF.

Таблица 12.2 – Значения RF термов в сообщении

Термы в сообщении	RF
внутр	3
аккредитации	1
вложения	3,90689
2012	4,45943
информационные	1
образовательных	2
ресурсы	1,32193
программ	2,32193
научной	3,32193
рассылаю	1
библиотеки	3,32193
справку	1
уважаемые	2,39232
состоянии	1
заведующие	2
фонда	1
кафедрами	2,32193
вуза	1
целью	2
уважением	2,48543
информированности	1
Иванов	2,80735
информационных	2
Иван	3,16993
ресурсах	1
Иванович	4,24793
также	1,80735
директора	3,58496
успешного	1

Согласно принятому порогу термы $RF_{t_q}^{k} \le 1{,}5$, следовательно, такие термы, как *аккредитации, информационные, рассылаю, справку, состоянии, вуза, информированности, ресурсах, успешного, прохождения* исключаются из данного сообщения.

Анализ значимости термов в сообщении после сокращения признакового пространства (рисунок 12.5) показал, что существуют термы, обладающие большим классификационным свойством и, следовательно, могут в большей степени повлиять на результат фильтрации, а некоторые меньшим свойством, что говорит о необходимости выделения термов, наиболее важных для детектирования сообщений. Данное обстоятельство говорит о необходимости применения методики выделения ключевых термов из текстового содержимого сообщения, которые отражают специфику ЭПС и позволяют выделить термы имеющие наиболее выраженные классификационные свойства.

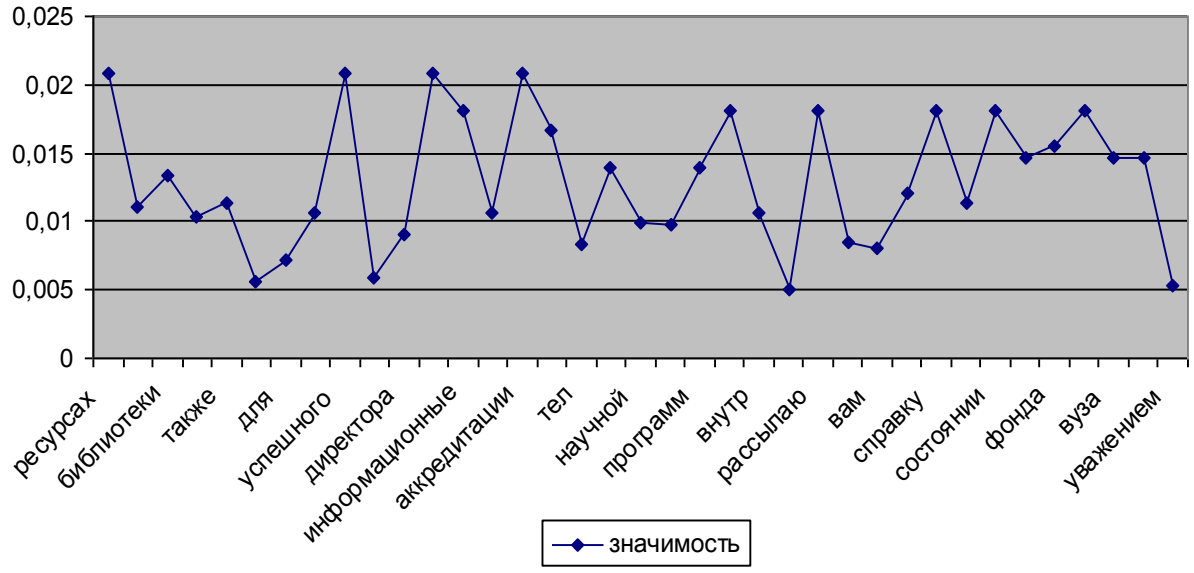


Рисунок 12.5 - Значимость термов сообщения после сокращения пространства признаков статистическими методами

В то же время, не все термы, содержащиеся в сообщении, отражают его тематику. Служебные слова не несут смысловой нагрузки, а используются для связи слов в предложениях (предлоги, союзы, частицы). Самостоятельные сло-

ва, частота употребления которых в тексте невелика, также не отражают тематику текста [15]. В связи с этим необходимо выделить термы способные отражать содержание сообщения. Извлечение таких термов может быть смоделировано через процедуры выделения ключевых слов текста.

Тогда решение задачи фильтрации входящей почтовой корреспонденции можно представить в виде последовательности, показанных на рисунке 12.6.

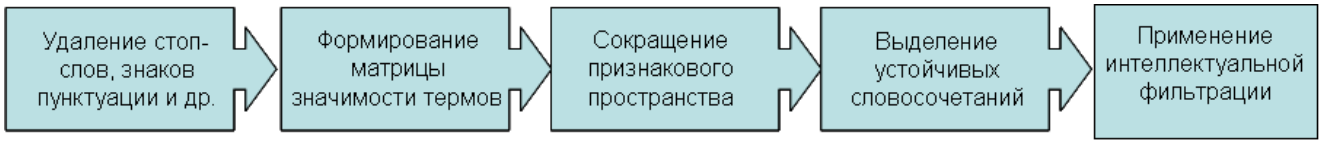


Рисунок 12.6 – Технология решения задачи фильтрации электронной корреспонденции

Отсюда, под устойчивым словосочетанием понимается комбинация двух и более термов, имеющих тенденцию к совместной встречаемости [8,15,17], т.е. речь идет об осмысленных последовательностях слов в тексте. Такое рассмотрение устойчивых словосочетаний позволяет применять к ним различные статистические меры, с помощью которых можно определить наличие связи между элементами словосочетаний и силу связи между ними.

В настоящее время в лингвистике существует несколько способов для вычисления степени связанности частей устойчивых словосочетаний [1,3,8,15,16,17]. В качестве таких статистических мер выбраны меры ассоциации, которые чаще всего используются при вычислении степени близости между компонентами словосочетаний.

Существуют два основных подхода к автоматическому выделению терминов [3,15,17].

Первый подход относится к области статической обработки естественного языка - вычисление различных мер ассоциативной связи, которые оценивают, является ли взаимное появление лексических единиц случайным, или оно статически значимо.

Второй подход опирается на семантическую близость термов., предполагающую определение семантической связности термов в сообщениях.

Пусть $D_i=\{d_{jq}\}, _{j=1,\ldots,N}$ характеристика связи между термами в i-ом сообщении, где d_{jq} – степень смысловой близости j-го и q-го термов.

В качестве меры близости между термами в сообщении предложено использовать расстояние Даейса. Данная статистическая мера позволит объединить термы в устойчивые (ключевые) словосочетания, характеризующие семантическое содержание сообщений.

Близость D и частота $f(t_1,t_2)$ совместной встречаемости термов становятся предпосылкой для нахождения устойчивых словосочетаний.

Последовательность формирования устойчивого словосочетания:

– выделение значимых термов для соответствующего класса k (spam/legitim);

– расчет близости термов и принятие решения о формирования устойчивого словосочетания;

– подтверждение смысловой значимости устойчивого словосочетания.

Мера Дайса D рассчитывается по зависимости вида:

$$D(t_1,t_2)=\log_2\left(\frac{2*(f(t_1,t_2))}{f(t_1)+f(t_2)}\right),$$

где $f(t_1)$ и $f(t_2)$ – частота встречаемости термов t_1 и t_2 в сообщении;

$f(t_1,t_2)$ – частота совместной встречаемости термов t_1 и t_2.

Для задачи фильтрации электронных почтовых сообщение предлагается формировать устойчивое словосочетание, если значение D_{jq} равное или выше, чем в соседних парах термов.

Для подтверждения смысловой значимости полученных устойчивых словосочетаний предлагается оценить тесноту взаимосвязи между термами в словосочетании, метрикой которой могут выступать меры ассоциации или контингенции.

Наиболее распространенными мерами ассоциации являются *MI-score*, *t-score* и *log-likelihood* [8,15,16], которые признаны показателями силы смысловой (синаптической) связи между качественными признаками (термами) словосочетаний. В качестве меры тесноты взаимосвязи двух качественных признаков словосочетаний предложено использовать коэффициенты ассоциации K_a и контингенции K_k, которые рассчитываются по следующим зависимостям:

$$K_a = \frac{ad - bc}{ad + bc},$$

$$K_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}},$$

где a – число сообщений, имеющих терм t_1, который встречается в классе k;

b – число сообщений, в которых терм t_1 встречается с другим классом;

c – число сообщений, имеющих терм t_2, который встречается в классе k;

d – число сообщений, в которых терм t_2 встречается с другим классом.

Экспериментально установлено [15], что связь между элементами словосочетания считается подтвержденной, если $K_a \geq 0{,}5$ или $K_к \geq 0{,}3$.

Тогда модель почтового ЭС можно представить в виде:

$$L(p_i) = <T^k, w^*(t_j)>,$$

где T^k – терм устойчивых словосочетаний в сообщении;

$w^*(t_j)$ – вес терма в сообщении после сокращения матрицы признаков.

Таким образом, модель ЭС в форме устойчивых словосочетаний позволяет без потери смыслового содержания обеспечить интеллектуальную классификацию почтовой электронной корреспонденции.

Вопросы для самоконтроля

1 Каким пространством признаков можно определить текстовое содержимое документа?

2 Перечислите основные этапы представления текстового содержимого электронного сообщения?

3 Какие существуют меры определения значимости термов в тексте?

4 В чем заключается закон Ципфа?

5 Для чего необходимо сокращать признаковое пространство?

6 Какие существуют подходы для сокращения признакового пространства задачи классификации?

7 Какие основные этапы включает в себя методика выявления устойчивых словосочетаний?

8 Какие основные шаги включает в себя технология решения задачи фильтрации электронной корреспонденции?

Заключение

Учебный материал настоящего учебного пособия ориентирован на подготовку студентов магистратуры по направлениям ФГОС ВО «Информатика и вычислительная техника и «Программная инженерия» в рамках дисциплин «Методы теории принятия решений», «Системы поддержки принятия решений», «Теория систем и системный анализ». Содержание и объем материала определяется рабочей программой читаемой дисциплины.

В первом разделе рассмотрены специальные методы исследования операций в задачах компьютерного зрения применительно к идентификации и распознаванию поверхностных дефектов тонколистового проката.

Во втором разделе раскрыты методы исследования операций в задачах информационной безопасности информационно-телекоммуникационных систем применительно к корпоративным компьютерным системам с территориально распределенной структурой.

В третьем разделе изложены методы исследования операций в задачах цифровой обработки текстовой информации применительно к распознаванию электронной почтовой корреспонденции.

Содержание разделов по главам соответствует последовательности лекций, причем практическая часть материала и алгоритмы могут быть использованы на практических занятиям и при выполнении лабораторных работ.

Список использованных источников

1 Аверьянов, Л.Я., Контент-анализ: уч. пособие / Л.Я. Аверьянов – М.: КноРус. - 2007 г., – 284 с.

2 Блаттер, К. Вейвлет – анализ. Основы теории / К. Блаттер. – М.: ТЕХ-СФЕРА, 2006. – 272 с.

3 Большакова, Е.И., Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие /Е.И.Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова — М.: МИЭМ, 2011. — 272 с.

4 Вентцель, Е.С. Исследование операций: задачи, принципы, методология: учеб. пособие для вузов /Е.С. Вентцель – 3-е изд., стер. – М.: Дрофа, 2004. – 208 с.: ил.

5 Грешилов, А.А. Математические методы принятия решений /А.А. Грешилов – М.: Изд. МГТУ им. Н.Э. Баумана, 2006 г. – 584 с.

6 Кузьмин М.И. Информационно-измерительная система распознавания поверхностных дефектов листового проката на основе метода окрестностей: диссертация кандидата технических наук; 05.11.16, Оренбург, 2016 г.

7 Льюнг, Л. Идентификация систем. Теория для пользователя /Л.Льюнг; Пер. с англ. А.С. Манделя, А.В. Назина; под ред. Я.З. Цыпкина. – М.: Наука, 1991. – 432 с.

8 Маннинг, К.Д. Введение в информационный поиск / К.Д. Маннинг, П. Рагхаван, Х. Шютце – М.:ООО «И.Д. Вильямс», 2011. – 528с.

9 Ногин, В.Д. Принятие решений в многокритериальной среде /В.Д. Ногин – М.: Физматлит, 2002. – 408 с.

10 Соловьев Н.А. Основы теории принятия решений для программистов: уч. пособие /Н.А. Соловьев, Е.Н. Чернопрудова, Д.А. Лесовой. – Оренбург: ООО ИПК «Университет», 2012. - 153 с.- ISBN 978-5-4417-0092-4

11 Соловьев, Н.А. Методы спектрального анализа в задаче обнаружения аномалий информационных процессов телекоммуникационных сетей [Электронный ресурс]: монография /Н.А. Соловьев [и др.] – Оренбург: ОГУ – 2013. –

171 с.- ISBN 978-5-4417-0330-7.

12 Соловьев, Н.А. Исследование операций в задачах программной инженерии [Электронный ресурс]: учебное пособие / [А.Ф. Валеев и др.] – Оренбург: ОГУ. - 2017. - 202 с.

13 Тишина, Н.А. Информационное и программное обеспечение идентификации аномалий корпоративных компьютерных сетей [Текст]: монография / Н.А. Тишина, И.А. Щудро. - Оренбург : Экспресс-печать, 2016. - 164 с.: ил.; 10,25 печ. л. - Библиогр.: с. 116-136. - Прил.: с. 127-160. - ISBN 978-5-9906460-1-8.

14 Тишина, Н.А. Цифровая обработка информации в задачах и примерах [Электронный ресурс]: учебное пособие / Н.А. Соловьев, Н.А. Тишина, Л.А. Юркевская - Оренбург : ОГУ. - 2016. - 122 с. - ISBN 978-5-7410-1614-5.

15 Чернопрудова, Е.Н. Программное обеспечение защиты почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений [Электронный ресурс]: монография /[Н.А. Соловьев и др.]; - Оренбург : ОГУ. - 2017. - 128 с. - ISBN 978-5-7410-1724-1.

16 Черноруцкий, И.Г. Методы принятия решений /И.Г. Черноруцкий. – СПб.: БХВ – Петербург, 2005. – 416 с.: ил.

17 Шевелев, О.Г. Методы автоматической классификации текстов на естественном языке: уч. пособие /О.Г. Шевелев. – Томск: ТМЛ-Пресс, 2007. – 144 с.

Приложение А

(*обязательное*)

Тематика магистерских исследований

№	Руководитель	Тематика ВКР
1	2	3
1.	Доктор технических наук, профессор Соловьев Николай Алексеевич	1 Системы обнаружения аномалий информационных процессов телекоммуникационных сетей корпоративных предприятий территориально-распределенной структуры. 2 Системы компьютерного зрения со средствами поддержки принятия решений. 3 Автоматизированные системы предприятий (организаций) со средствами поддержки принятия решений 4 Имитационное моделирование в задачах анализа технологических процессов
2.	Доктор технических наук, профессор Зубкова Татьяна Михайловна	1. Проектирование адаптивных графических пользовательских интерфейсов по заданной предметной области с учетом пользовательской аудитории. 2. Разработка адаптивных интерфейсов web-приложений. 3. Разработка интерфейсов web-приложений для мобильных устройств с построением (поиском) дерева приоритетов. 4. ПС расчета затрат при проектировании программного обеспечения. 5. Формирование технического задания на программное обеспечение с учетом пользовательской аудитории. 6. ПС интеллектуальной поддержки решений креативных проблем. 7. АИС формирования приоритетов развития персонала. 8. ПС ведения электронного классного журнала учителя 9. Обработка и анализ информации. 10. Математическое моделирование технологических объектов.
3.	Кандидат технических наук, доцент Горбачев Дмитрий Владимирович	1.Имитационное моделирование пространства состояний пациента на основе искусственной нейронной сети. 2. Разработка системы поддержки принятия решений для обеспечения доступности первичной медико-санитарной помощи сельскому населению.
4.	Кандидат технических наук, доцент Семенов Анатолий Михайлович	1 Разработка компонентов автоматизированных информационных систем с элементами искусственного интеллекта. 2 Разработка экспертной системы или (СППР) для выбранной предметной области.
5.	Кандидат технических наук, доцент Костин Владимир Николаевич	1 Разработка автоматизированных информационных систем физической защиты потенциально опасных объектов. 2. Средства принятия решений на основе методов оптимизации

6.	Кандидат технических наук, доцент Волкова Татьяна Викторовна	1 Автоматизация управления ресурсами корпоративной автоматизированной информационной системы (КАИС) (анализ наполнения данных - полнота, актуальность; анализ состояния компонентов технического обеспечения и их соответствия функционалу АС (по заданным критериям); управление правами доступа персонала системы; мониторинг, оптимизация сложных процессов обработки данных в интегрированной базе данных большой АС). 2 Разработка веб-приложений, сайтов предприятия на основе корпоративной автоматизированной информационной системы 3 Автоматизация хранения, модерирования и представления ресурсов пользователей (личные кабинеты пользователей на основе веб-приложений и баз данных). 4 Автоматизация управления компонентами сайта (мониторинг состава и состояния компонентов, управление доступом, анализ реализация инклюзивных требований, реализация мобильной версии и др.) 5 Автоматизация исследования структуры и контента веб-сайта на основе заданных параметров (мониторинг наличия/ отсутствия информации для целевого пользователя, актуальность обновления, соответствие требованиям законодательства (обработка персональных данных, проверка информационных ресурсов на присутствие в Федеральном списке экстремистских материалов и др.).
7.	Кандидат педагогических наук, доцент Тагирова Лилия Фаритовна	1 Разработка интеллектуальных адаптивных интерфейсов. 2 Использование средств информационно-коммуникационных технологий в сфере образования.
8.	Кандидат педагогических наук, доцент Наточая Елена Николаевна	1 Разработка автоматизированных информационных систем управления образовательной организацией. 2 Разработка электронных образовательных ресурсов.
9.	Кандидат технических наук, доцент Щудро Игорь Анатольевич	1 Экспертные системы поддержки принятия решений. 2 Использование технологии Blockchain для системы финансового учета.
10.	Кандидат технических наук, Чернопрудова Елена Николаевна	1. Разработка автоматизированной системы для анализа неструктурированных документов. 2. Исследование и разработка методов классификации в задачах многокритериального принятия решений. 3. Разработка системы поддержки принятия решений для информационных систем.
11.	Кандидат технических наук Тишина Наталья Александровна	1. Информационная безопасность облачных технологий. 2. Защита информации информационно-коммуникационных систем 3. Цифровая обработка изображений и звуков
12.	Кандидат технических наук, доцент Юркевская Любовь Аркадьевна	1. Автоматизированная информационная система предметной области со средствами поддержки принятия решений. 2. Безопасность информационно-коммуникационных систем