Уфа: УГАТУ, 2005 Вестник УГАТУ Т. 6, № 1 (12). С. 212–226

УДК 004.4'242

А. К. СКУРАТОВ

АЛГОРИТМЫ АНАЛИЗА И МОНИТОРИНГА ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИЧЕСКИХ МЕТОДОВ

Рассматриваются проблемы, возникающие в телекоммуникационных системах и компьютерных сетях, и предлагаются пути их решения на основе анализа и мониторинга с использованием статистических методов. *Телекоммуникации; гетерогенные сети; мониторинг; статистические методы*

Развитие телекоммуникационных систем и компьютерных сетей обуславливает необходимость создания и надежного функционирования большого набора инфокоммуникационных сервисов, обеспечивающих эффективную работу пользователя с разнородной информацией в гетерогенной телекоммуникационной сети. Практика использования и эксплуатации гетерогенных телекоммуникационных систем и компьютерных сетей, связанная с недостаточной их прозрачностью, сложностью, организационными ограничениями и спецификой, определяет необходимость более широкого и научно обоснованного внедрения статистических методов их анализа и мониторинга на основе открытой потоковой информации, которую можно получить, используя предлагаемые методы и средства [9, 10, 13].

1. ПОКАЗАТЕЛИ ФУНКЦИОНИРОВАНИЯ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

Главным требованием, предъявляемым к сетям, является обеспечение пользователям возможности доступа к разделяемым ресурсам всех компьютеров, объединенных в сеть [2, 6, 7]. Из исследования и анализа состояния, перспектив и тенденций развития телекоммуникационных сетей вытекают следующие показатели функционирования сети — **производительность, надежность и безопасность, совместимость, управляемость, расширяемость и масштабируемость**. Рассмотрим их подробнее для формирования нормального профиля функционирования глобальной телекоммуникационной сети [11].

1.1. Производительность

Как показано в [2, 6, 7, 15], высокая производительность — это одно из основных свойств распределительных систем, к которым относится глобальная телекоммуникационная сеть. Оно обеспечивается возможностью распараллеливания работ между несколькими компьютерами сети с использованием прокси-серверов. Рассмотрим основные характеристики производительности сети.

Время реакции сети является интегральной характеристикой производительности сети с точки зрения пользователя. В общем случае время реакции определяется как интервал времени между возникновением запроса пользователя к какому-либо инфокоммуникационному сервису и получения ответа на этот запрос. Складывается из нескольких составляющих: времени подготовки запросов на клиентском компьютере, времени передачи запросов между клиентом и сервером через сегменты сети и промежуточное коммуникационное оборудование, времени обработки запросов на сервере, времени передачи ответов от сервера клиенту и времени обработки получаемых от сервера ответов на клиентском компьютере.

Пропускная способность отражает объем данных, переданных сетью или ее частью в единицу времени. Эта характеристика говорит о скорости выполнения внутренних операций сети — передачи пакетов данных между узлами сети через различные коммутационные устройства.

Задержка передачи определяется как задержка между моментом поступления пакета

Выполнено при поддержке Российского фонда фундаментальных исследований, проект № 05-07-90360.

на вход какого-либо сетевого устройства или части сети и моментом появления его на выходе этого устройства. Этот параметр по смыслу близок ко времени реакции сети, но отличается тем, что всегда характеризует только сетевые этапы обработки данных без задержек обработки компьютерами сети.

1.2. Надежность и безопасность

Для оценки надежности сложных систем применяется следующий набор характеристик [2, 6, 7, 15].

Готовность или коэффициент готовности означает период времени, в течение которого система может быть использована. Готовность может быть улучшена путем введения избыточности в структуру системы: ключевые элементы системы должны существовать в нескольких экземплярах, чтобы при отказе одного из них функционирование системы обеспечивали другие. Чтобы систему можно было отнести к высоконадежным, она должна как минимум обладать высокой готовностью, но только этого недостаточно. Необходимо обеспечивать **сохранность данных** и защиту их от искажения. Кроме того, должна поддерживаться согласованность (непротиворечивость) данных.

Вероятность доставки пакета узлу назначения без искажения, поскольку сеть работает на основе механизма передачи пакетов между конечными узлами.

Безопасность — способность системы защитить данные от несанкционированного доступа, поскольку всегда существует потенциальная угроза взлома защиты сети.

Отказоустойчивость — способность системы скрыть от пользователя отказ отдельных ее элементов. В отказоустойчивой системе отказ одного из ее элементов приводит к некоторому снижению качества ее работы, а не к полной остановке функционирования.

1.3. Управляемость

Управляемость подразумевает возможность централизованно контролировать и изменять состояния основных элементов сети, выявлять и разрешать проблемы, возникающие при работе сети, выполнять анализ производительности и планировать развитие сети, осуществлять текущий мониторинг [2, 7, 6, 15, 10, 13].

Правильно организованная система управления осуществляет интеллектуальное наблюдение за сетью и, обнаружив проблему, активизирует определенное действие, исправляет ситуацию и уведомляет администратора о том, что произошло и какие шаги предприняты. Одновременно с этим система управления должна накапливать данные, на основании которых можно планировать развитие сети. Примером такой интеллектуальной системы управления является комплекс программных средств фирмы IBM — Tivoli [13].

Полезность системы управления особенно ярко проявляется в больших сетях [2, 7, 15, 13].

1.4. Совместимость и интегрируемость

Совместимость в случае телекоммуникационной сети означает, что сеть способна включать в себя самое разнообразное программное и аппаратное обеспечение, т. е. в ней могут сосуществовать различные операционные системы, поддерживаться стеки коммуникационных протоколов и работать аппаратные средства и приложения от разных производителей. Сеть, состоящая из разнотипных элементов, называется неоднородной или гетерогенной [7, 15].

1.5. Качество обслуживания

Хотя все приведенные выше параметры функционирования телекоммуникационной сети весьма важны, часто понятие «качество обслуживания» (QoS) телекоммуникационной сети трактуется более узко — в него включаются только две самые важные характеристики сети — **производительность и надежность** [7, 15, 4].

Параметр QoS определяет, какая сетевая полоса пропускания должна быть назначена трафику каждого конкретного приложения и как следует управлять ею. Кроме того, он обеспечивает предсказуемый уровень полосы пропускания на базе IP в зависимости от важности процессов, связанных с трафиком.

1.6. Расширяемость и масштабируемость

Данные термины иногда используются как синонимы, но это неверно — каждый из них имеет четко определенное значение [7, 15].

Расширяемость означает возможность сравнительно легкого добавления отдельных элементов сети для наращивания (расширения) сегментов сети и замены существующей аппаратуры на более производительную или обладающую расширенным набором функций.

Масштабируемость означает, что сеть позволяет наращивать количество узлов и протяженность связей в очень широких пределах, при этом производительность сети не ухудшается.

1.7. Прозрачность

Прозрачность сети достигается в том случае, когда сеть представляется пользователям не как множество отдельных компьютеров, связанных между собой сложной системой кабелей, а как единая традиционная вычислительная машина с системой разделения времени [7, 15].

Основные практические результаты (методы, алгоритмы и программы) данной работы, связанные со статистической обработкой информации, циркулирующей в телекоммуникационных сетях, должны использоваться системными администраторами для управления работой сети. Рассмотрим эти задачи системы управления.

2. ЗАДАЧИ УПРАВЛЕНИЯ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТЬЮ

Выделяют пять функциональных групп задач системы управления, определенных стандартами ISO/ITU-T [7].

1-я группа: управление конфигурацией сети и именованием — эти задачи заключаются в конфигурировании параметров как отдельных элементов сети, так и телекоммуникационной сети в целом. Для элементов сети с помощью этой группы задач определяются сетевые адреса, идентификаторы (имена), географическое положение. Для сети в целом управление конфигурацией обычно начинается с построения карты сети, т. е. отображения реальных связей между элементами сети и изменения связей между элементами сети. ***Статистические результаты работы телекоммуникационной сети могут служить основой при принятии решения в рамках данной группы задач управления.***

2-я группа: обработка ошибок — эта группа задач включает выявление, определение и устранение последствий сбоев и отказов сети, ***в том числе и на основе статистических результатов работы сети.***

3-я группа: анализ производительности и надежности — задачи этой группы связаны с оценкой ***на основе накопительной статистической информации*** таких параметров, как время реакции системы, пропускная способность реального или виртуального канала связи, интенсивность трафика в отдельных сегментах и каналах сети, вероятность искажения данных при их передаче через сеть, а также коэффициент готовности сети. Функции анализа производительности и надежности сети нужны как для оперативного управления сетью, так и для планирования развития сети.

4-я группа: управление безопасностью — задачи этой группы включают в себя контроль доступа к данным при их хранении и передаче через сеть. Базовыми элементами управления безопасностью являются процедуры аутентификации пользователей, назначение и проверка прав доступа к ресурсам сети, управления полномочиями и т. д. ***При решении задач управления данной группы следует учитывать результаты статистической обработки информации об атаках на сеть и попытках несанкционированного доступа к ее ресурсам.***

5-я группа: учет работы сети — задачи этой группы — заниматься регистрацией времени использования различных ресурсов сети — устройств, каналов и транспортных служб, ***в том числе с учетом статистических параметров работы телекоммуникационной сети.***

Таким образом, практические задачи настоящей статьи непосредственным образом связаны с задачами управления телекоммуникационными сетями.

3. МОНИТОРИНГ И АНАЛИЗ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ

Процесс контроля работы сети обычно делят на два этапа — мониторинг и анализ [7, 10, 12–14].

На этапе **мониторинга** выполняется процедура сбора первичных данных о работе сети: статистика о количестве циркулирующих в сети пакетов различных протоколов, о состоянии портов концентраторов, коммутаторов и маршрутизаторов и т. п. По состоянию первичных параметров телекоммуникационной сети определяют вышеуказанные параметры функционирования сети.

Назовем ***текущим профилем*** телекоммуникационной сети совокупность текущих параметров телекоммуникационной сети, измеренных в заданный временной интервал. Таким образом, под мониторингом телекоммуникационной сети будем понимать сбор и фиксацию текущего профиля телекоммуникационной сети в заданный временной интервал.

Далее выполняется этап **анализа**, под которым понимается более сложный и интеллектуальный процесс осмысления, с использованием специальных инструментальных средств, собранной на этапе мониторинга информации, сопоставления ее с данными, полученными ранее, и выработки предположений о возможных причинах замедленной или ненадежной работы сети.

Назовем **нормальным профилем** телекоммуникационной сети совокупность параметров телекоммуникационной сети, которые установлены регламентом ее функционирования в заданный временной интервал. Таким образом, под анализом телекоммуникационной сети будем понимать процесс сравнения текущего и нормального профилей телекоммуникационной сети в заданный временной интервал. Под результатом анализа будем понимать совокупность зафиксированных значений расхождения между текущим и нормальным профилем телекоммуникационной сети по каждому параметру.

Задача анализа требует более активного участия человека и использования таких сложных средств, как экспертные системы, аккумулирующие практический опыт многих сетевых специалистов. Полученную информацию о работе телекоммуникационной сети можно анализировать с различной степенью глубины (или детализации). Выделяют три основных уровня глубины исследования: анализ статистической информации, декодирование протоколов, экспертный анализ [7, 13].

Статистическая информация представляет собой первый уровень детализации. Статистика может быть текущей (с интервалом усреднения информации от одной до нескольких секунд) и долговременной (с интервалом усреднения информации от одной минуты до нескольких часов).

Второй уровень детализации в анализе функционирования телекоммуникационной сети может быть осуществлен с помощью функций **захвата пакетов и декодирования содержащихся в пакетах протоколов**. Декодирование протоколов позволяет выявить такие дефекты, которые невозможно понять на основе анализа статистической информации. Чаще всего декодирование протоколов используется тогда, когда необходимо определить причину отсутствия связи между узлами сети.

Третий уровень детализации в анализе трафика проводится на основе **экспертного анализа** проходящих по сети пакетов.

Следует подчеркнуть, что системы управления представляют собой сложные и дорогие программно-аппаратные комплексы, поэтому существует граница целесообразности применения таких систем управления — она зависит от сложности, целевого назначения сети и степени ее территориальной распределенности. В небольшой сети можно применять отдельные программы управления наиболее сложными устройствами, поставляемыми производителем. Однако при росте сети могут возникнуть проблемы объединения разрозненных программ управления устройствами в единую систему управления, и для решения этой проблемы придется, возможно, отказаться от этих программ и заменить их интегрированной системой управления [7, 13].

4. ПРАКТИЧЕСКИЕ ЗАДАЧИ, ДЛЯ РЕШЕНИЯ КОТОРЫХ ПРОВОДИТСЯ МОНИТОРИНГ И АНАЛИЗ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

А) *Предсказание изменения параметров сетевого трафика на основе обработки статистической информации о работе элементов сети.* Эта информация носит, как правило, статистический характер и представляет собой временные последовательности. В этом случае речь идет о статистическом анализе сетевого трафика как анализе временных рядов, а анализируемая статистика может быть как текущей (с интервалом усреднения информации от одной до десятков секунд), так и долговременной (с интервалом усреднения информации от одной минуты до нескольких часов или суток).

Б) *Интеллектуальное управление компьютерными сетями для перераспределения сетевых ресурсов, в частности, пропускной способности виртуальных каналов.* Это достигается за счет статистического мультиплексирования с временным разделением пропускной способности между различными информационными приложениями. Методы управления перераспределением пропускной способности позволяют оптимально распределить информационные потоки по виртуальным каналам. При этом учитываются ограничения на доступную пропускную способность и уровень показателей качества. Указанные методы представляют собой симбиоз алгоритмов резервирования пропускной способности виртуальных каналов и статистического мультиплексирования ресурсов. Последние освобождаются вследствие случай-

ного характера распределения нагрузки по видам сервиса.

В) *Исследование временных задержек вдоль маршрута прохождения пакета, снижение которых повышает качество работы сети.* Временные задержки являются важным фактором, влияющим на пропускную способность сети. Чем больше эти задержки, тем меньше пропускная способность сети. В соответствии с протоколом TCP/IP пропускная способность со стороны источника пакетов определяется текущим окном перегрузки, равным числу разрешенных к передаче пакетов до прихода пакета подтверждения. При более или менее регулярном поступлении пакетов подтверждения величина окна увеличивается в два раза. В итоге достигается максимально возможная для принятого протокола пропускная способность, уменьшается окно перегрузки и, соответственно, пропускная способность соединения.

Г) Следующая *задача вытекает из предыдущей и заключается в формировании прогноза времени появления перегрузки и ее величины.* Задержка и потеря пакетов в пути могут происходить из-за очередей в промежуточных узлах — маршрутизаторах, в компьютерах — источниках и приемниках пакетов, а также из-за переполнения буферов в этих узлах. В таком случае пакеты подтверждения не отсылаются, и протоколом TCP на стороне источника формируется окно перегрузки уменьшенного размера. Интервал между моментами отсылки пакета из источника и получения пакета подтверждения называется RTT-задержкой (англ. round-trip time — задержка). Указанная задержка является важной характеристикой, обеспечивающей нормальное функционирование TCP-соединения в фазе медленного старта и поэтому требующей тщательной настройки и контроля. Для избежания простоев из-за ожидания потерянных и задержавшихся пакетов вводится пороговое значение RTT-задержки. Пакеты считаются потерянными, если RTT превышает заданный порог. По величине спрогнозированной RTT-задержки можно судить об уровне перегрузки и перенастроить величину окна, т. е. определить закон изменения окна перегрузки.

Д) *Контроль и прогнозирование переполнения буферов.* На пропускную способность участка сети между i-м и j-м узлами, очевидно, влияет очередь в узле j. Эта очередь может возникнуть из-за ограниченного объема памяти данного буфера, низкой интенсивности разгрузки этого буфера, чрезмерно больших объемов информации, поступивших на него. В связи с этим интенсивность потока информации от узла i к узлу j понижается, а в случае переполнения буфера в узле j передача информации прекращается и часть пакетов теряется. Для предотвращения потери пропускной способности узла необходимо регулировать уровень загрузки буфера на основе прогноза его переполнения.

Е) *Сравнение текущего профиля телекоммуникационной сети с нормальным профилем и выявление сетевых аномалий.*

Постановка и решение каждой задачи связаны с получением, обработкой и интерпретацией информации о работе телекоммуникационной сети, которой располагает администратор (пользователь) сети для соответствующего исследования. Рассмотрим источники информации о функционировании телекоммуникационной сети и программные средства, с помощью которых эта информация может регистрироваться.

5. РАЗРАБОТКА АЛГОРИТМОВ СТАТИСТИЧЕСКОЙ СИСТЕМЫ АНАЛИЗА ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

5.1. Анализ данных с пропущенными значениями

Оценка одномерных статистических характеристик. Если в данных имеются пропуски, то для оценки одномерных статистических характеристик некоторой переменной x используются все измеренные значения этой переменной [3].

Оценки матрицы ковариаций и вектора средних. Многие методы многомерного статистического анализа, включая множественную линейную регрессию, анализ главных компонент, дискриминантный анализ и канонический корреляционный анализ, основаны на преобразовании матрицы данных в выборочные средние и матрицу ковариаций переменных. В этом разделе рассматриваются EM-*оценки* среднего и ковариационной матрицы по выборке многомерных данных с пропусками из нормального распределения, в предположении, что пропуски *возникают случайно* (СП). Хотя предположение многомерной нормальности выглядит ограничительным, рассматриваемая здесь реализация EM алгоритма обычно обеспечивает приемлемые оценки и при более слабых предположениях об исследуемых распределениях.

Оценивание по EM-алгоритму носит итеративный характер. Каждая итерация состоит

из двух шагов E и M (откуда и произошло название алгоритма).

Шаг M. На этом шаге оценивание параметров проводится по методу максимального правдоподобия (МП) так, как будто пропусков нет, т. е. как будто они заполнены. Таким образом, на шаге M EM-алгоритма используются те же вычислительные методы, что и при получении МП-оценок. Для нормального распределения это стандартные оценки вектора средних и ковариационной матрицы.

Шаг E. На шаге E находят условное ожидание «пропущенных» данных при фиксированных наблюденных данных и текущих оценках параметров. Затем пропущенные данные заменяются найденными ожидаемыми значениями. Однако на практике в EM-алгоритме не обязательно происходит действительное заполнение пропусков.

Предположим теперь, что рассматривается p-мерная нормально распределенная переменная $(x_1, x_2, ..., x_p)$ с вектором средних значений $\mu = (\mu_1, ..., \mu_p)$ и ковариационной матрицей $\Sigma = (\sigma_{ik})$. Пусть $\mathbf{X}$ представляет выборку объема n из p-мерных векторов $(X_1, ...X_n)$. Ее можно представить в виде $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, где $\mathbf{X}_{\text{obs}}$ — множество наблюдаемых значений, а $\mathbf{X}_{\text{mis}}$ — множество пропущенных значений.

Пусть на t-й итерации $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$. Шаг E алгоритма состоит в вычислении

$$E\left(\sum_{i=1}^{n} x_{ij} | \mathbf{X}_{\text{obs}}, \theta^{(t)}\right) = \sum_{i=1}^{n} x_{ij}^{(t)},$$
$$j = 1, ..., p;$$
$$E\left(\sum_{i=1}^{n} x_{ij}x_{ik} | \mathbf{X}_{\text{obs}}, \theta^{(t)}\right) = \sum_{i=1}^{n} \left(x_{ij}^{(t)} x_{ik}^{(t)} + c_{jki}^{(t)}\right),$$
$$j, k = 1, ..., p,$$

где $x_{ij}^{(t)} = x_{ij}$, если x_{ij} присутствует и $E(x_{ij} | \mathbf{X}_{\text{obs}}, \theta^{(t)})$, если x_{ij} пропущено, $c_{jki}^{(t)} = 0$, если x_{ij} или x_{ik} присутствует и $\text{cov}\,(x_{ij}, x_{ik} | \mathbf{X}_{\text{obs},i}, \theta^{(t)})$, если x_{ij} и x_{ik} пропущены.

Таким образом, отсутствующие значения x_{ij} заменяются средними x_{ij}, условными по присутствующим значениям $x_{\text{obs},i}$ в этом наблюдении.

Оценка $\theta^{(t+1)}$ вычисляется по формулам:

$$\mu_j^{(t+1)} = n^{-1} \sum_{i=1}^{n} x_{ij}^{(t)} \quad j = 1, ..., p;$$

$$\sigma_{jk}^{(t+1)} =$$
$$= n^{-1} E\left(\sum_{i=1}^{n} x_{ij}x_{ik} | \mathbf{X}_{\text{obs}}\right) - \mu_j^{(t+1)} \mu_k^{(t+1)} =$$
$$= n^{-1} \sum_{i=1}^{n} \left[(x_{ij}^{(t)} - \mu_j^{(t+1)})(x_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)}\right],$$
$$j, k = 1, ..., p.$$

Начальные значения параметров получаются по методу всех возможных измеренных значений.

5.2. Анализ резко выделяющихся (аномальных) наблюдений

Устойчивые оценки. Данная процедура предназначена для визуального анализа резко выделяющихся наблюдений и получения устойчивых оценок средних, дисперсий и корреляций [3].

Подход на основе целенаправленного проецирования. Рассмотрим метод с целью его реализации в **статистической системе анализа телекоммуникационной сети** [11, 13] для выделения наблюдений (объектов), которые сильно отклоняются от центра распределения. Иногда такие большие отклонения возникают в результате случайного сдвига десятичной запятой при записи наблюдения или подготовке данных, неправильного чтения показаний измерительного прибора и т. д. В **статистической системе анализа телекоммуникационной сети** [11, 13] реализуем визуальный метод анализа резко выделяющихся наблюдений (заметим, что иногда с этой целью можно попытаться использовать метод главных компонент и проекции на исходные переменные). Этот метод основан на целенаправленном проецировании (ЦП) с выбором проекций, подходящих для анализа аномальных наблюдений.

Пусть $U = (u_1, ..., u_p)'$ есть некоторое направление проецирования, так что от p-мерных объектов X_i, (заданных в виде строк матрицы данных $\mathbf{X}i = \overline{1, n}$), переходим к их одномерным проекциям $y_i = (U'X_i)$. В качестве критерия выбора направления проецирования, приспособленного для выявления аномальных наблюдений, возьмем величину

$$l(U) = s^2(y)/s_{\text{rob}}^2(y), \qquad (1)$$

где $s^2(y)$ — обычная оценка дисперсии для y; $s_{\text{rob}}^2(y)$ — некоторая устойчивая оценка параметра масштаба для y.

Таким образом, выбирается направление проецирования, на котором стандартная оценка дисперсии в наибольшей степени «испорчена» присутствием аномальных наблюдений. В качестве устойчивой оценки в **статистической системе анализа телекоммуникационной сети** [11, 13] используем экспоненциально-взвешенную оценку. Требуемое направление U_1 получается как направление, для которого величина $l(U_1)$ максимальна.

Хорошее приближение для определения направления U_1, максимизирующего (1), получается из решения обобщенной задачи на собственные значения:

$$(\mathbf{S} - l\mathbf{S}_{\text{rob}})U = \mathbf{0}, \qquad (2)$$

где $\mathbf{S}$ – обычная оценка ковариационной матрицы для X; $\mathbf{S}_{\text{rob}}$ – устойчивая оценка матрицы рассеивания.

В качестве устойчивой оценки используется экспоненциально-взвешенная оценка, получаемая в результате решения (с помощью итеративной процедуры) системы уравнений

$$M = \sum_{i=1}^{n} w_1(t_i)X_i \Big/ \sum_{i=1}^{n} w_1(t_i),$$

$$\mathbf{S} = \beta(w_1, w_2)\sum_{i=1}^{n} w_2(t_i)\frac{(X_i - M)(X_i - M)'}{\sum\limits_{i=1}^{n} w_2(t_i)}, \qquad (3)$$

где $w_1(t) = w_2(t) = \exp(-\lambda t/2)$; λ – малый параметр, $\lambda = 1/(p+1)$;

$$t_i = (X_i - M)'\mathbf{S}^{-1}(X_i - M);$$

X_i – i-е наблюдение (объект), $i = \overline{1,n}$; $\beta = 1 + \lambda$ – скалярный множитель, обеспечивающий несмещенность оценки в случае выборки из нормального распределения.

Направлению U_1 соответствует собственный вектор с максимальным собственным значением. Решение задачи (2) является приближенным решением задачи (3), поскольку равенство $s^2_{\text{rob}} = U'\mathbf{S}_{\text{rob}}U$, вообще говоря, выполняется лишь приближенно в отличие от точного равенства $s^2(y) = U'\mathbf{S}U$.

Второе направление проецирования U_2 получается из решения оптимизационной задачи

$$U_2 = \arg\max_{U} l(U), \quad (U_1'U_2) = 0.$$

В качестве приближенного решения для второго направления проецирования используется собственный вектор задачи (2), соответствующий второму по величине собственному числу. Всего же получается p векторов, упорядоченных по убыванию собственных чисел.

Для визуального анализа аномальных наблюдений будем использовать интерактивную графическую галерею **статистической системы анализа телекоммуникационной сети** [11, 13].

5.3. Алгоритм k-средних при заданном числе классов

Идея метода. Реализованные в **статистической системе анализа телекоммуникационной сети** [11, 13] алгоритмы, краткое описание которых следует ниже, представляют собой частный случай общего метода динамических сгущений. Идея метода состоит в таком разбиении множества объектов на заранее известное число классов k, чтобы минимизировался функционал, заданный равенством:

$$W = \sum_{j=1}^{k} W_j, W_j = \sum_{X_i \in D_j} d^2(X_i, C_j),$$

где C_j – центр j-го класса.

Если центры C_j являются центрами тяжести классов, то этот метод сводится к методу центра тяжести или k средних. Если, кроме того, расстояние d определяется как евклидово, то функционал W заменяют на

$$W = \sum_{j=1}^{k} W_j, \quad W_j = \sum_{X_i \in D_j} d^2(X_i, C_j).$$

Вычислительная процедура носит итерационный характер и может быть описана следующим образом.

Шаг 1. Из n исходных объектов случайным образом выбираются k объектов (например, первые k объектов), которые объявляются центрами классов. Затем остальные $n - k$ объектов разбиваются относительно этих k центров следующим образом: любой объект из $n - k$ объектов считается принадлежащим к j-му классу, если этот объект ближе к j-му центру, чем к остальным (в смысле введенной меры расстояния). Полученные k классов обозначим через $D_1^{(0)}, ..., D_k^{(0)}$.

Шаг 2. Вычисляются центры $C_1^{(1)}, ..., C_k^{(1)}$ классов $D_1^{(0)}, ..., D_k^{(0)}$. Затем n объектов разбиваются на k классов относительно новых центров классов по правилу: объект относится к j-му классу, если центр j-го класса является наиболее близким к этому объекту из имеющихся центров классов. В результате получаются классы $D_1^{(1)}, ..., D_k^{(1)}$.

Шаг 3. Вычисляются центры классов $C_1^{(1)}, ..., C_k^{(1)}$. Исходные n объектов разбиваются снова на k классов согласно правилу, сформулированному на шаге 2. В результате получаются классы $D_1^{(2)}, ..., D_k^{(2)}$.

Процедура завершается либо когда стабилизируются центры классов, т. е. будут выполняться неравенства:

$$C_i^{(t+1)} = C_i^{(t)} (i = \overline{1,k}),$$

либо когда число итераций t превысит $t_{\max}$.

Сходимость алгоритма за конечное число шагов следует из общей теории [1]. Вообще говоря, описанный выше алгоритм не определяет глобального максимума критерия W на множестве разбиений, но зато обладает довольно высоким быстродействием.

В настоящей версии кластер-анализа в качестве центра класса на некотором шаге работы алгоритма выбирается тот объект, принадлежащий классу, сумма расстояний до которого от остальных объектов этого класса минимальна.

Если используется евклидова метрика, то предложенный способ выбора центра класса тесно связан с вычислением центра тяжести объектов из класса. А именно, пусть E есть объект, выбранный в качестве центра класса при первом подходе, а X — центр тяжести этого класса. Тогда можно показать, что E есть объект, ближайший к X среди объектов, входящих в класс.

Результат работы алгоритма зависит, особенно если количество объектов n невелико, от выбора начальных центров на шаге 1.

5.4. Иерархическая восходящая классификация

В статистической системе анализа телекоммуникационной сети [11, 13] реализуем ряд алгоритмов восходящей, или агломеративной, иерархической классификации (ВИК). Суть алгоритмов ВИК достаточно проста [5] и может быть описана следующим образом.

Пусть имеется множество из n объектов, между которыми определена некоторая функция расстояния $d(i, j)$ (в частности, расстояния могут быть заданы в качестве исходной матрицы расстояний). Должно быть определено и правило вычисления расстояния между двумя множествами a и b (группами, классами) объектов. Обозначим его $V(a, b)$. Алгоритмы ВИК носят пошаговый (последовательный) характер, число шагов работы любого из этих алгоритмов равно $n - 1$.

На первом шаге каждый из n объектов рассматривается как отдельная группа (класс) и два ближайших (в смысле введенной меры расстояния $d(i, j)$) объединяются в одну группу. Следовательно, при переходе ко второму шагу число групп будет $n - 1$.

Рассмотрим теперь k-й шаг алгоритма. При входе в него имеется $(n - k + 1)$ классов. Два ближайших класса (по введенной мере расстояний между классами) объединяются в один класс. Так что при выходе на $(k + 1)$-й шаг будет уже $(n - k)$ групп. При выходе же на последний $(n - 1)$-й шаг имеется всего две группы объектов, которые и объединяются. Получается один класс, объединяющий всю совокупность объектов. Таким образом, на первом шаге и по завершении работы алгоритмов ВИК мы имеем дело с тривиальными группировками: на первом шаге — «каждый объект — группа», на последнем — «все объекты — одна группа».

Основные причины, делающие алгоритмы ВИК полезными для анализа данных, следующие:

- получается иерархическая последовательность вариантов разбиения (кластеризации) исходного множества объектов, содержащая разбиения с любым допустимым числом классов от 1 до n. Эти разбиения называются частичными иерархиями и для них используется обозначение C_{k-1}, где k — номер шага; следовательно, иерархии C_{k-1} соответствует число классов $n - k + 1$;
- результаты работы алгоритмов ВИК допускают весьма удобное для анализа графическое представление: бинарное иерархическое дерево (дерево классификаций, или дендрограмму).

Бинарное иерархическое дерево. Бинарное дерево состоит из узлов и ребер. Каждому узлу соответствует класс (группа), полученный на некотором шаге алгоритма.

Последовательность узлов дерева может быть упорядочена по уровням снизу вверх, так что классифицируемым объектам соот-

ветствуют узлы самого нижнего уровня, а единственному классу, являющемуся объединением всех объектов и получаемому по завершении работы алгоритма, — узел самого высокого уровня. Узлы нижнего уровня часто называют терминальными. Из каждого узла более высокого уровня выходят два ребра, идущие к двум разным узлам более низкого уровня (поэтому дерево называется бинарным). Узлу, полученному объединением двух объектов на первом шаге, присваивается номер $(n+1)$. Узлу, получаемому объединением двух классов на втором шаге, присваивается номер $(n+2)$ и т. д.

В случае, когда некоторые объекты не участвуют в классификации, номер класса, получаемого на первом шаге, будет не $n+1$, а $k+1$, где k — максимальный номер объекта из всех объектов, участвующих в классификации.

Пример. Имеем $N = 10$ объектов, их номера $1, 2, 3, ..., 10$. Из них в классификации участвуют только $n = 6$ объектов с номерами 1, 2, 3, 5, 6, 7. Номера получаемых классов будут соответственно 8, 9, 10, 11, 12.

Определение вертикальной координаты узла. При графическом построении бинарного дерева для определения вертикальных координат узлов-преемников используются следующие шкалы (индексации):

- по мощности класса (под мощностью подразумевается число объектов в классе), для терминальных классов $V = 1$ (power-индексация);
- по значению меры расстояния между объединяющимися классами, для терминальных классов $V = 0$ (index-индексация);
- по порядку образования классов, для терминальных классов $V = 0$, для первого образовавшегося класса $V = 1$, для второго — $V = 2$ и т. д. (order-индексация).

Для некоторых типов расстояний между классами необходимо сделать следующие замечания:

10) Не гарантировано, что уровень $V(r)$ превышает уровни объединяемых классов, т. е. возможны так называемые инверсии. В таких случаях можно в качестве $V(r)$ использовать максимальное значение уровня объединяемых классов.

2) Если минимальное значение V достигается сразу на двух парах классов, то произвольно объединяется только одна из них. Несмотря на то, что с математической точки зрения иерархии различны, это сказывается на интерпретации не слишком сильно.

Некоторые свойства кластеров (классов), порождаемых ВИК. По определению в результате работы этих алгоритмов получаются неперекрывающиеся кластеры. Однако они являются вложенными в том смысле, что каждый кластер может рассматриваться как элемент кластеров, расположенных на более высоком, чем он, уровне иерархического дерева.

Иерархия кластеров, порождаемая ВИК, зависит от нормы и расстояния между объектами, меры расстояния между классами.

Основная операция с деревом — разрезание по некоторому уровню вертикальной шкалы. В этом случае получается классификация, соответствующая узлам, расположенным непосредственно ниже уровня разреза. Относительно этих классов доступна информация о мерах разброса, значениях переменных и т. д. Существует также возможность разрезания дерева на заданное число классов.

5.5. Быстрый метод иерархической восходящей классификации (метод сводимых окрестностей)

В алгоритме иерархической восходящей классификации необходимо вычислить все расстояния между классами текущего разбиения. Затем берется минимальное значение указанных расстояний и классы, на которых оно реализуется, объединяются. Для экономии памяти и уменьшения необходимого числа сравнений при поиске минимального расстояния следует исключить из рассмотрения те расстояния, которые не влияют на выполнение вычислений. Эта идея [5] реализована в быстродействующем алгоритме иерархической восходящей классификации. Она основана на специальном свойстве расстояний между подмножествами — свойстве «сводимости».

Свойство сводимости: иерархия обладает свойством сводимости, если окрестность класса, образованного объединением двух классов, включена в объединение окрестностей этих двух классов.

Свойство сводимости позволяет предложить следующую модификацию традиционных процедур иерархической классификации. Устанавливается некоторое значение порогового расстояния t между классами. Затем составляется список-граф, куда входят дуги (и их вершины — классы), длина которых не превышает t. Два ближайших класса

из списка объединяются. Расстояния между классами пересчитываются только по элементам списка, а не по всей исходной совокупности данных. В то же время из списка удаляются дуги, ведущие в вершины, соответствующие двум объединенным классам, и добавляются дуги (длина которых не превышает t), ведущие в вершину, соответствующую образованному на этом шаге классу. После того как список будет исчерпан, порог t увеличивается и составляется новый список. Таким образом, по завершении работы алгоритма сводимых окрестностей мы также имеем группировку «все объекты — одна группа».

5.6. Алгоритмы двухфакторного разложения для анализа сезонно трендовых моделей (ST-моделей)

Алгоритмы, объединенные в эту группу, имеют общую входную форму и некоторые части также совпадают, поэтому программно они могут быть реализованы в одной процедуре.

Для применения этих алгоритмов временной ряд разворачивается в двухвходовую таблицу. Это возможно в том случае, когда наблюдения временного ряда содержат внутренний цикл (сезонную составляющую). Тогда из реализаций временного ряда формируется таблица размера $N \times M$ (N строк, M столбцов), где $N \times M$ — общее число наблюдений; M — количество внутренних циклов (сезонов); N — число наблюдений во внутреннем цикле.

5.6.1. Алгоритм, основанный на удалении средних значений

Цель применения алгоритма. Алгоритм позволяет разложить временной ряд на 3 аддитивные компоненты: *тренд, сезонную составляющую и случайную компоненту.*

Описание алгоритма. Непараметрический алгоритм обработки временного ряда, «свернутого» в двухвходовую таблицу, состоит из следующих шагов.

Шаг 1. Исходный ряд разворачивается в двухвходовую $N \times M$ таблицу.

Шаг 2. Вычисляется среднее по всем значениям в таблице и затем вычитается из всех наблюдений.

Шаг 3. Вычисляется среднее значение по каждой строке, и эти значения вычитаются из каждого наблюдения в соответствующей строке.

Шаг 4. Вычисляется среднее по каждому столбцу таблицы и затем эти значения вычитаются от каждого наблюдения в соответствующем столбце.

Алгоритм не нуждается в итерировании, для получения аддитивной модели, наилучшей в смысле наименьших квадратов, достаточно выполнить вышеописанную процедуру один раз.

В случае, когда данные не содержат «выбросов», использование этого алгоритма предпочтительнее алгоритма медианного сглаживания.

Достоинства алгоритма:

- не нуждается в итерировании;
- минимизирует сумму квадратов остатков;
- результаты легко интерпретируются.

5.6.2. Итерационный алгоритм медианного сглаживания

Цель применения алгоритма. Как и предыдущий, этот алгоритм позволяет разложить временной ряд на 3 аддитивные компоненты: ***тренд, сезонную составляющую и случайную компоненту.*** При этом обеспечивается высокая устойчивость к выбросам в данных.

Описание алгоритма. Непараметрический алгоритм обработки временного ряда, «свернутого» в двухвходовую таблицу, состоит из следующих шагов.

Шаг 0. Исходный ряд разворачивается в двухвходовую N*M таблицу.

Дальнейшие шаги выполняются итеративно, в зависимости от требуемой точности разложения. Заметим, что на первой итерации все значения параметров, которые в дальнейшем берутся с предыдущей итерации $n-1$, полагаем равными 0.

Шаг 1. Вычисляются медианы по строкам.

Шаг 2. Вычисляется медиана вектора медиан строк.

Шаг 3. Медианы по строкам вычитаются из соответствующих элементов строк.

Шаг 4. Вычисляются медианы по столбцам.

Шаг 5. Вычисляются медианы вектора медиан столбцов.

Шаг 6. Медианы столбцов вычитаются из соответствующих элементов столбцов.

Шаг 7. Формирование главного значения таблицы: сумма главного значения с предыдущей итерации (для первой итерации = 0) и ме-

дианы вектора медиан строк и медианы вектора медиан столбцов.

Шаг 8. Вычисляется поправка вектора медиан строк.

Шаг 9. Вычисляется поправка вектора медиан столбцов.

Далее сравниваются матрицы остатков в k и $k-1$ итерациях; если разности суммы квадратов элементов меньше указанной величины, итерационная процедура прекращается (параметр eps1).

Несмотря на внешне более громоздкий алгоритм, число вычислительных операций для одной итерации примерно такое же, что требуется для метода, основанного на удалении средних.

Данный алгоритм сходится к аддитивному представлению, которое является наилучшим в смысле минимизации суммы *абсолютных значений* остатков.

5.7. Алгоритм разложения по целевым факторам

Цель применения алгоритма. Разложение временного ряда остатков, «свернутого» в двухходовую таблицу (см. метод, основанный на средних, и метод медианного сглаживания), непараметрическим путем на таблицы, равные по размерности исходной. С каждой таблицей связывается пара векторов. Аддитивная сумма таблиц образует исходную таблицу остатков.

Пары векторов, связанные с таблицами, являются поправками к тренду и сезонному поведению, полученных ранее с помощью алгоритма, основанного на удалении средних, или алгоритма медианного сглаживания.

Описание алгоритма:

Исходный временной ряд предполагается рядом остатков после выполнения аддитивного разложения методом, основанным на средних, или с помощью медианного сглаживания.

Алгоритм заключается в последовательном применении двухфакторного разложения к вновь образующейся матрице остатков. Итерационная процедура разложения по целевым факторам, в случае если исследуемая таблица соответствует гипотезе об аддитивной природе временного ряда, довольно быстро приводит к «вычищению» матрицы остатков. Под «вычищением» понимается то, что в результате каждой итерации ряд остатков разбивается на три аддитивные компоненты: поправку к главному циклу, поправку к внутреннему циклу и новую таблицу остатков; и с каждой итерацией абсолютные значения элементов таблицы остатков убывают. Содержательный смысл имеют поправки первого порядка: поправка первого порядка к тренду имеет смысл поправки в i-й момент времени $1 \leqslant i \leqslant N$ к амплитуде внутреннего цикла (сезонной компоненте), поправка первого порядка к сезонной компоненте описывает смещение тренда по среднесезонному поведению за j-й внутренний временной момент.

Определение 1: Пусть $v_1 \in R^n$, $w_1 \in R^m$ – фиксированные векторы и C – некоторая $n \times m$-таблица. Двухфакторным разложением таблицы C называется представление ее в виде

$$C = v_1 w_2' + v_2 w_1' + C_1, \qquad (4)$$

где $v_k \in R^n$, $w_k \in R^m$, $k = 1,2$ – векторы-столбцы, $'$ – символ транспонирования, $v_1 w_2'$ и $v_2 w_1'$ – $(n \times m)$-таблицы, представляющие собой произведения векторов-столбцов на векторы-строки, C_1 – $(n \times m)$-таблица остатков.

Разложение C_1 строится на основе функционала F, выражающего критерий малости таблицы остатков. Мы ограничимся критерием метода наименьших квадратов, т. е.

$$F(C_1) = \sum_{i=1}^{n}\sum_{j=1}^{m} C_1(i,j)^2 = S_p(C_1', C_1), \qquad (5)$$

где $C_1(i,j)$ – (i,j) матричный элемент, а $S_p(\cdot)$ – след матрицы – сумма ее диагональных элементов.

Итак, приходим к задаче:

Найти ($v_2 \in R^n$ и $w_2 \in R^m$), такие, что $F(C_1)$ принимает минимальное значение для данных $v_1 \in R^n$ и $w_1 \in R^m$.

Решение

$$F(C_1) = S_p(C_1', C_1) =$$
$$= S_p[(C' - w_2 v_1' - w_1 v_2')(C - v_1 w_2' - v_2 w_1')] =$$
$$= S_p C' C - 2(v_1' C w_2 + v_2' C w_1) +$$
$$+(\|v_1\|^2 \|w_2\|^2 + 2(w_1, w_2)(v_1, v_2) + \|v_2\|^2 \|w_1\|^2),$$

где $\|\cdot\|$ – норма вектора, а $(\cdot,\cdot)$ – скалярное произведение.

Следовательно:

$$\mathrm{grad}_{w_2} F(C_1) = -2C'v_1 + 2\|v_1\|^2 w_2 + 2(v_1, v_2) w_1;$$
$$\mathrm{grad}_{v_2} F(C_1) = -2Cw_1 + 2(w_1, w_2) v_1 + 2\|w_1\|^2 v_2.$$

Из условия $\mathrm{grad}_{w_2, v_2} = 0$ при $\|v_1\|^2 \neq 0$ и $\|w_1\|^2 \neq 0$ получаем:

$$w_2 = \frac{C'v_2}{\|v_1\|^2} - \frac{(v_1, v_2)}{\|v_1\|^2} w_1; \quad (6)$$

$$v_2 = \frac{Cw_1}{\|w_1\|^2} - \frac{(w_1, w_2)}{\|w_1\|^2} v_1. \quad (7)$$

Подставляя (6) и (7) в (4), получаем:

$$C = \frac{v_1 v_1' C}{\|v_1\|^2} - \frac{(v_1, v_2)}{\|v_1\|^2} v_1 w_1' + \frac{Cw_1 w_1'}{\|w_1\|^2} v_1 w_1' + C_1 = \frac{v_1 v_1' C}{\|v_1\|^2} + \frac{Cw_1 w_1'}{\|w_1\|^2} - \left(\frac{(v_1, v_2)}{\|v_1\|^2} + \frac{(w_1, w_2)}{\|w_1\|^2} \right) v_1 w_1' + C_1.$$

Из (6) и (7) получаем:

$$\frac{(v_1, v_2)}{\|v_1\|^2} + \frac{(w_1, w_2)}{\|w_1\|^2} = \frac{v_1' C w_1}{\|v_1\|^2 \|w_1\|^2}.$$

Окончательно получаем решение задачи в виде

$$C = \frac{v_1' C w_1}{\|v_1\|^2 \|w_1\|^2} + v_1 \tilde{w}'_2 + \tilde{v}_2 w_1' + C_1, \quad (8)$$

где

$$\tilde{w}_2 = \frac{C'v_1}{\|v_1\|^2} - \frac{v_1' C w_1}{\|v_1\|^2 \|w_1\|^2} w_1; \quad (9)$$

$$\tilde{v}_2 = \frac{Cw_1}{\|w_1\|^2} - \frac{v_1' C w_1}{\|v_2\|^2 \|w_1\|^2} v_1.$$

Следствие 1. Двухфакторное разложение таблицы в случае $v_1' = (v_1(1), ...v_1(n)), w'' = (w_1(1), ..., w_1(m))$, где $v_1(i) = 1, 1 \leqslant i \leqslant n$ и $w_1(j) = 1, 1 \leqslant j \leqslant m$, дает разложение таблицы C по методу средних.

Доказательство. Для $v_1' = (1, ..., 1)$ и $w_1' = (1, ..., 1)$ имеем: $\|v_1\|^2 = n, \|w_1\|^2 = m$; $\frac{v_1' C w_1}{\|v_1\|^2}$ — среднее матричных элементов; $\frac{v_1' C}{\|v_1\|^2}$ — вектор-строка, составленная из средних значений векторов-столбцов таблицы C; $\frac{Cw_1}{\|w_1\|^2}$ — вектор-столбец, составленный из средних значений таблицы C.

Заметим теперь, что если $\|\tilde{w}_2\| \neq 0$ и $\|\tilde{v}_2\| \neq 0$, то метод двухфакторного разложения таблиц можно применить к таблице остатков C_1 и получить ее разложение для векторов $\tilde{v}_2$ и $\tilde{w}_2$:

$$C_1 = \tilde{v}_2 \tilde{w}_3' + \tilde{v}_3 \tilde{w}'_2 + C_2,$$

где $\tilde{w}_3$ и $\tilde{v}_3$ рассчитываются по формулам (5) и (6), где $C = C_1, v_1 = \tilde{v}_2$ и $w_1 = \tilde{w}_2$.

Следствие 2. Для $v_1' = (1, .., 1)$ и $w_1' = (1, ..., 1)$ по формулам (5) и (6) получим векторы $\tilde{w}_2$, $\tilde{v}_2$, $\tilde{w}_3 = (w_3(j))$, $\tilde{v}_3 = (v_3(i))$. Тогда $v_3(i)$ — поправка в i-й год к амплитуде годового колебания, описываемого вектором $\tilde{w}_2$, $1 \leqslant i \leqslant n$, $\tilde{w}_3(j)$ — смещение тренда, описываемого вектором v_2, при расчете тренда только по среднемесячным концентрациям за j-й месяц года.

Доказательство непосредственно вытекает из формулы

$$C = \frac{v_1' C v_1}{\|v_1\|^2 \|w_1\|^2} + (w_1 + \tilde{v}_3)\tilde{w}'_2 + \tilde{v}_2(w_1' + \tilde{w}_3') + C_2.$$

5.8. Алгоритмы вычисления оценок спектральной плотности

Спектральная плотность мощности стационарного процесса определяется как преобразования Фурье автокорреляционной функции [8] и описывается как мощность случайного процесса распределения по частоте и является для действительных реализаций четной неотрицательной функцией частоты. Аналогично взаимная спектральная плотность мощности двух стационарных процессов определяется как преобразование Фурье взаимнокорреляционной функции и в общем случае является комплекснозначной функцией.

Оценки СПМ (ВСПМ) могут быть получены с использованием периодограммного метода Уэлша, если длина реализации X состоит из N отсчетов и используется K последовательных сегментов длиной $L = [N/(K-1)]$ с 50%-м перекрытием. Взвешенный p-й сегмент будет состоять из

$$x_i^p = w_i x_{i+pL}, \quad p = 0, \quad K-1, \quad (10)$$

где w_i — выбранное временное окно. Выборочный спектр

$$S_{xx}^p(f) = \frac{1}{L\Delta t \cdot U} |X^p(f)|^2, \quad (11)$$

где L — длина сегмента, Δt — шаг дискретизации, $U = \Delta t \sum_{i=0}^{L-1} w_i^2$ — энергия временного окна, $X^{(p)}(f) = \Delta t \sum_{M=0}^{L-1} x_n^{(p)} \exp(-i2\pi f n \Delta t)$ —

дискретное преобразование Фурье, выполненное на M' частотных отсчетах с использованием быстрого преобразования Фурье. Среднее значение выборочных спектров дает оценку периодограммы Уэлша

$$\tilde{S}(f) = \frac{1}{K}\sum_{p=0}^{K-1} S_x^p(f). \qquad (12)$$

Для ВСПМ процедура производится аналогично, только

$$S_{xy}^p(f) = \frac{1}{L\Delta t U} X^p(f)\cdot\overline{Y^p(f)},$$

где $\overline{Y^p(f)}$ — сопряженное значение дискретного преобразования Фурье для p-го взвешенного сегмента реализации y.

Рекомендации по выбору окна. Предлагается использование окон трех типов:

Прямоугольное окно

$$w_i = 1.$$

Окно Хэмминга

$$w_i = 0{,}538 + 0{,}462\cos(2\pi t[i]).$$

Окно Натолла

$$w_i = 0{,}3635819 + 0{,}4891775\cos(2\pi t[i]) + \\ +0{,}1365995\cos(4\pi t[i]) + 0{,}0106411\cos(6\pi t[i]),$$

где $t[i] = (i-(N-1)/2)/(N-1)$.

Обработка с использованием временных окон используется для управления эффектами, обусловленными наличием боковых лепестков в спектральных оценках. В силу конечной длительности «обрабатываемой реализации появляется эффект просачивания». Поэтому в спектре одиночной гармонической компоненты присутствует один главный лепесток, характеризующий частоту и амплитуду этой компоненты, и несколько боковых лепестков. Просачивание приводит не только к появлению амплитудных ошибок в спектрах, но также маскирует присутствие слабых сигналов и, следовательно, препятствует их обнаружению.

Окна Хэмминга и Натолла существенно снижают уровень боковых лепестков, что снижает смещение спектральных оценок. Однако это дается ценой расширения главного лепестка спектра окна, что, естественно, приводит к ухудшению разрешения. Поэтому при выборе окна должен выбираться компромисс между шириной главного лепестка и уровнем подавления боковых лепестков. При выборе окна необходимо учитывать несколько количественных показателей их качества.

Замечание 1. Ширина полосы на уровне половинной мощности, т. е. на уровне, который на 3 дБ ниже максимума главного лепестка. Измеряется в бинах — безразмерных величинах, определяющих величину элемента разрешения дискретного преобразования Фурье; 1 бин$= 1/N$, где N — длина реализации.

Пусть q — ширина полосы по уровню половинной мощности окна, используемого при расчете спектра реализации из N отсчетов с шагом дискретизации Δt. Тогда в тексте могут быть разрешены гармонические компоненты одинаковой амплитуды, частоты которых отличаются более чем на величину $q/(N\Delta t)$. Также может быть использован показатель ширины полосы главного лепестка по уровню 6дБ.

Замечание 2. Максимальный уровень боковых лепестков, который позволяет судить, насколько хорошо подавляется просачивание. Если c — уровень боковых лепестков по мощности (не в дБ) и a — мощность максимальной компоненты, то устойчиво могут быть обнаружены компоненты, мощность которых превышает $c\cdot a$.

Также полезной характеристикой окна является скорость спада боковых лепестков на каждые 8 бин. Эти характеристики отражены в табл.

Вывод. Если основная задача — разрешение компонент равной амплитуды, предпочтительным является использование прямоугольного окна. Если необходимо обнаружение слабых компонент, целесообразно использование окна Натолла. При отсутствии априорной информации предпочтительнее использовать окно Хэмминга, обладающее средним разрешением и достаточным уровнем боковых лепестков.

Рекомендации по выбору числа сегментов. При выборе K — числа сегментов, на которых последовательно рассчитывается и усредняется периодограммная оценка Уэлша, как и при выборе окна, приходится искать компромисс между степенью гладкости спектральной оценки и требуемым спектральным разрешением. Действительно, при большом K будет получаться много сегментов, по спектрам которых будет проводиться усреднение, а следовательно, будут получаться оценки с меньшей дисперсией, но также и с меньшим разрешением. Уменьшение K повышает спектральное разрешение, но, естественно, за счет

Таблица

Окно	Ширина 3 дБ	Максимальный уровень БЛ	Скорость спада	Эквивалентная ширина окна
Прямоугольное	0,89	–13 дБ	–6 дБ/онг	1,00
Хэмминга	1,30	–43 дБ	–6 дБ/онг	1,36
Натолла.....	1,70	–98 дБ	–6 дБ/онг	1,80

увеличения дисперсии из-за меньшего числа усредняемых сегментов.

Если отсутствует априорная информация о сигнале, позволяющая определить необходимое спектральное разрешение и дисперсию оценок СПМ, то рекомендуется использовать процедуру «закрытие окна», которая заключается в нескольких повторных проходах по данным при различных длинах и числах сегментов данных. Это необходимо для исследования данных при различных параметрах частотного разрешения и устойчивости оценки, что позволит получить больше полезной информации об изучаемой реализации. Процедура «закрытия окна» начинается с использованием низкого разрешения и высокой устойчивости оценок, с последующим переходом к оценкам с большим разрешением и более низкой устойчивостью. Визуальный анализ спектров позволит остановиться на той спектральной оценке, которая наиболее соответствует представлению об исследуемой реализации.

Аналитические оценки качества спектральных (справедливые лишь для гауссовских процессов):

• математическое ожидание периодограммы Уэлша

$$M(\tilde{S}(f)) = \frac{1}{K}\sum_{p=0}^{K-1} S_x^{(p)}(f) = \\ = S(f) * |W(f)|^2/U, \quad (13)$$

где $*$ — операция свертки, $W(f)$ — дискретное преобразование Фурье окна данных w_i, $S(f)$ — истинный спектр процесса;

• дисперсия периодограммы Уэлша примерно равна

$$D(\tilde{S}(f)) \approx \frac{S^2(f)}{K}.$$

Математическое ожидание и дисперсия периодограмм Уэлша для случая расчета ВСПМ обладают аналогичными свойствами.

Если заранее известно необходимое спектральное разрешение Δf, то число необходимых сегментов может быть определено следующим образом:

$$K = [N * \Delta t \Delta f / q] + 1,$$

где N — число отсчетов, Δt — временная длина одного отсчета, q — ширина главного лепестка используемого окна. Дисперсия спектральной оценки может быть рассчитана по (С8) с заменой $S(f)$ на $\tilde{S}(f)$.

ВЫВОДЫ

1) Представлен алгоритм анализа данных с пропущенными значениями для применения в случае многомерных временных рядов.

2) Модифицирован подход на основе целенаправленного проецирования с целью реализации в статистической системе анализа телекоммуникационной сети визуального метода анализа резко выделяющихся наблюдений.

3) Адаптирован алгоритм k-средних при заданном числе классов для реализации в статистической системе анализа телекоммуникационной сети.

4) Представлен ряд алгоритмов восходящей или агломеративной, иерархической классификации с целью реализации в статистической системе анализа телекоммуникационной сети.

5) Адаптированы следующие алгоритмы обработки рядов, содержащих компоненты тренда и сезонности, основанные на развертке ряда в двухвходовую таблицу: алгоритм удаления средних значений, итерационный алгоритм медианного сглаживания, алгоритм разложения по целевым факторам.

6) Отобраны типы окон в области данных для вычисления оценок спектральной плотности с целью их реализации в статистической системе анализа телекоммуникационной сети.

СПИСОК ЛИТЕРАТУРЫ

1. **Айвазян С. А., Бухштабер В. М.** Прикладная статистика. Классификация и снижение размерности: Справ. изд. М.: Финансы и статистика, 1989.
2. **Бройдо В. Л.** Вычислительные системы, сети и телекоммуникации. СПб.: Питер, 2002. 688 с.
3. **Бугай А. И., Гугель Ю. В., Скуратов А. К. и др.** Применение агрегирующих и разностных операторов для анализа потоков информации в сетях // Вестник РГРТА. Рязань, 2003. № 13. С. 41–46.
4. **Вегешна Ш.** Качество обслуживания в сетях IP. М.: Вильямс, 2003. 368 с.
5. **Голубев В. В., Никитин В. М., Никитина Д. А.** Статистика. Определение общей тенденции развития рядов динамики. М., 2002. 105 с.
6. **Пятибратов А. П., Гудыно Л. П.** Вычислительные системы, сети и телекоммуникации. М., 2001. 512 с.
7. **Сажин Ю. В., Катынь А. В. и др.** Статистические методы прогнозирования на основе временных рядов. Саранск: Изд-во Морд. ун-та, 2000. 113 с.
8. **Скуратов А. К.** Обеспечение доступа межшкольных методических центров проекта «Информатизация системы образования» к Интернет в рамках планируемого кредита МБРР // Телекоммуникации и информатизация образования. 2004. № 3, С. 14–37.
9. **Скуратов А. К.** Мониторинг функционирования телекоммуникационных сетей на основе статистической системы исследования и анализа информационных потоков // Вестник Моск. гос. ун-та леса. Лесной вестник. 2004. № 2 (33). С. 133–146.
10. **Скуратов А. К., Безрукавный Д. С.** Администрирование телекоммуникационной сети на основе статистического анализа трафика // Вестник Тамбовск. гос. техн. ун-та. 2004. № 4 а. С. 919–923.
11. **Скуратов А. К., Ретинская И. В. и др.** Анализ трафика научно-образовательных сетей // Автоматизация, телемеханизация и связь в нефтяной промышленности: Науч.-техн. журн. 2003. № 2. С. 4–7.
12. **Статистическое** проектирование информационно-управляющих систем. анализ и мониторинг научно-образовательных интернет-сетей / И. С. Енюков, И. В. Ретинская, А. К. Скуратов; Под. ред. А. Н. Тихонова. М.: Финансы и статистика, 2004. 320 с.
13. **Столлингс В.** Компьютерные системы передачи данных. М.: Вильямс, 2002. 928 с.
14. **Компьютерные** сети. Модернизация и поиск неисправностей. БХВ-Петербург: 2001. 1008 с.
15. **Толковый** словарь сетевых терминов и аббревиатур: Офиц. изд. Cisco Systems. М.: Вильямс, 2000. 368 с.

ОБ АВТОРЕ

Скуратов Алексей Константинович, доцент, зам. дир. ГНИИ ИТТ «Информика». Дипл. инж.-системотехник (МИЭМ, 1983). Канд. техн. наук по САПР (М., 1989). Лауреат прем. Правит. РФ. Иссл. в обл. инф. технол. и телекоммуникации.