УДК 621.372.542

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING В ОБРАБОТКЕ СИГНАЛЬНОЙ ИНФОРМАЦИИ (В ГЕОФИЗИЧЕСКИХ ИССЛЕДОВАНИЯХ)

В.В. Геппенер[1], А.Б. Тристанов[2], П.П. Фирстов[2]

[1] Санкт-Петербургский государственный электротехнический университет, г. Санкт-Петербург
[2] Институт вулканологии и сейсмологи ДВО РАН, г. Петропавловск-Камчатский,
neringa@mail.kamchatka.ru

Введение

В связи бурным развитием вычислительной техники в последние два десятилетия значительно возросли возможности современных систем сбора, передачи, хранения и обработки данных, получаемых с целью поиска предвестниковых аномалий землетрясений в тех или иных геофизических полях. Значительно возрос и объем получаемой геофизической информации, что требует современного подхода к обработке геофизических данных. В работах А.А. Любушина (мл.) [4] разрабатываются методы поиска аномалий в низкочастотных фоновых процессах в земной коре и атмосфере (временные вариации геофизических полей: наклоны, деформация земной поверхности, уровень подземных вод в скважинах, интенсивность сейсмоакустической эмиссии, интенсивность выделения газов и т. д., длительностью от нескольких минут до нескольких месяцев) на основе многомерного анализа временных рядов. Для анализа геофизических данных также успешно внедряются новые математические методы [4, 11], основанные на применение нелинейных процедур (фрактальный анализ, нейронные сети, самоподобие и самоорганизация).

В развитии работ такого направления выделение в геофизических сигналах особенностей, повторяющихся процессов, скрытых аномалий и закономерностей в автоматическом режиме с применением современных методов математического аппарата является актуальной задачей геофизики. В данной работе сигнал рассматривается как последовательность следующих друг за другом участков, обладающих на некотором временном интервале постоянными свойствами (структурная модель сигнала). Исходя из такого допущения, задача обработки сигнала рассматривается как выделение отдельных участков (сегментов) с последующей их кластеризацией. Следует отметить, что однозначно эти участки выделены быть не могут, вследствие априорной неопределенности в выборе модели сигнала, и одному сигналу может быть поставлено в соответствие множество последовательностей сегментов, зависящих от выбора модели. Такая последовательность может рассматриваться как сжатая информация, описывающая сигнал с позиции ограниченного круга свойств, определяемых моделью.

Построение подобного описания может формулироваться как построение дескрипторной (описательной) модели сигнала, что является, с одной стороны, обобщением задачи выделения особенностей, а с другой стороны, частным случаем решения задачи регрессии. Обнаружение закономерностей в последовательности сегментов и скрытых закономерностей в сигнале является задачей сиквенциального анализа, который может быть выполнен в рамках современного направления, называемого интеллектуальным анализом данных или технологией Data Mining [1], которая изначально разрабатывалась как технология для бизнес-приложений. Наиболее широкое применение технология Data Mining нашла в маркетинге, банковском деле, страховании и т. д. и применение ее к задачам геофизики, несомненно, должно помочь в решении задачи выделения аномального поведения геофизического сигнала, обусловленного изменением напряженно-деформированного состояния геосреды.

Предлагаемый подход к обработке геофизических сигналов

Предлагаемый подход к обработке сигналов геофизических полей, состоящий из трех этапов, заключается в решение (явно или не явно) следующих формальных задач (рис. 1):

1 - сегментации;
2 - классификации;
3 - сиквенциального анализа.

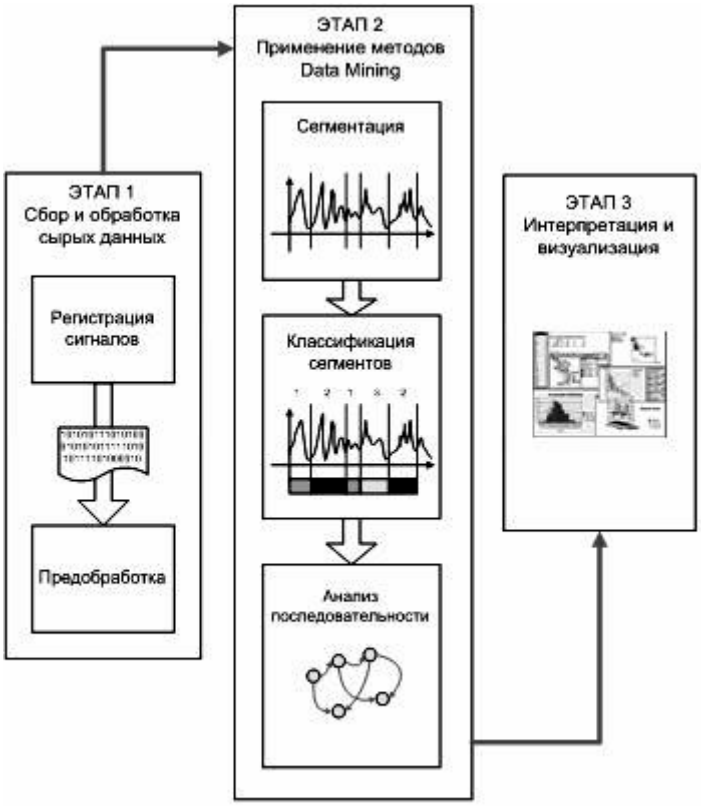


Рис. 1. Схема анализа геофизических сигналов на основе технологии Data Mining.

Этап 1. Сбор и обработка первичных данных. Задача получения экспериментальных данных решается на аппаратном уровне и включает в себя регистрацию сигналов и подготовку их цифровых записей. Далее происходит подготовка первичных данных к применению методов Data Mining.

Процесс предобработки включает в себя типовые операции цифровой обработки сигналов: 1 - преобразование форматов данных; 2 - удаление тренда; 3 - удаление аномальных значений (выбросов); 4 - фильтрация и очистка сигнала от шумов.

На этом этапе может быть выполнен предварительный анализ данных – анализ спектрального состава и основных статистических параметров.

Помимо описанных действий, на этапе интерпретации результатов может быть полезным наличие базы данных, содержащей параметры и условия регистрации данных.

Этап 2. Применение методов Data Mining. На данном этапе рассматриваются задачи, решение которых невозможно или представляет сложность при применении классических методов. К этим задачам относятся: 1 - выделение скрытых особенностей сигнала; 2 - выделение фоновых процессов; 3 - классификация особенностей; 4 - выявление скрытых закономерностей в последовательности условно стационарных участках сигнала.

Явное решение данных задач в геофизике затруднительно вследствие неопределенности в структуре наблюдаемых данных.

Решение перечисленных задач предлагается проводить по следующей схеме.

1) Использование методов выделения особенностей и формирование признакового описания сегментов. Этот пункт заключается в применении методов сегментации сигналов и является одним из вариантов задачи регрессии.

Сегментацию можно рассматривать как процесс преобразования сигнала к дискретной последовательности сегментов с известными свойствами. В связи с тем, что выделенные в первом пункте сегменты являются разнородными, как по длине, так и по структуре, следовательно, необходимо выбрать единые для всех сегментов признаки (представление сегментов).

2) Классификация сегментов. После получения подготовленного материала необходимо его систематизировать. В зависимости от выбранного алгоритма сегментации, выделенные

участки сигнала (сегменты) могут быть разделены по классам автоматически, либо это требует применения методов автоматической кластеризации.

В результате выделяется последовательность сегментов, каждому из которых поставлена в соответствие строка (код), определяющая класс сегмента. Данный код включает номер (название или другой идентификатор) класса и длительность сегмента. Данная последовательность подлежит дальнейшему анализу.

3) Анализ последовательностей. На данном этапе происходит решение задачи поиска ассоциативных правил (в частности сиквенциального анализа), т. е. поиск закономерностей в последовательности.

Предлагается вести анализ последовательности сегментов по двум направлениям: 1 - поиск повторяющихся групп; 2 - построение статистической модели.

Второе направление предполагает рассмотреть последовательность как реализацию случайного процесса, обладающего марковским свойством. Для описания системы требуется определить множество состояний и вероятности перехода из одного состояния в другое, а также параметры потока, приводящего систему в действие.

Этап 3. Интерпретация и визуализация. Данный этап связан с визуализацией результатов и предполагает использование методов когнитивной графики и прочих средств визуализации, упрощающих понимание полученных результатов пользователем.

Методы сегментации

Исследования по обнаружению изменений свойств случайных процессов начали сформировываться в отдельное направление теоретической статистики после публикации Е. Пейджа [10], в которой была поставлена задача обнаружения момента скачкообразного изменения среднего значения последовательности независимых случайных величин.

Задача сегментации появилась из задачи выделения локальных неоднородностей в сигналах со сложной структурой. Будем понимать под сегментацией процесс разделения сигнала на участки (сегменты), соответствующие некоторым структурным единицам. В свою очередь, сегмент - это участок сигнала, заданные свойства которого могут быть приняты постоянными. Для сегмента можно определить его границы, т. е. момент времени изменения свойств сигнала.

Ключевым объектом в алгоритме сегментация является критерий сегментации. Под критерием сегментации будем понимать функцию, определяющую поведение свойств сигнала.

Задача сегментации может быть сформулирована как задача поиска границ сегментов или задача поиска моментов изменения свойств сигнала [6, 7].

Рассмотрим временной ряд $\{x_t\}$. Требуется построить детектор, выбирающий одну из двух гипотез H_0 и H_1. Гипотеза H_0 предполагает, что исследуемый сигнал $\{x_t\}$ соответствует модели M_1. Гипотеза H_1 предполагает, что существует момент времени τ, в который исследуемый сигнал $\{x_t\}$ соответствует модели M_1 при $t < \tau$ и модели M_2 при $t \geq \tau$, где τ - граница сегмента.

Свойства, изменение которых могут быть обнаружены, определяются гипотезой о виде моделей M_1 и M_2. Эти модели, в свою очередь, определяют выбор критерия сегментации. Модели M_1 и M_2 представляют варианты параметрической модели M, зависящей от набора свойств (параметров) θ.

Параметры модели M_1 идентифицируются или задаются заранее и служат отправной точкой процедуры сегментации. Параметры модели M_2 в задаче обработки геофизических данных считаются априори неизвестными.

Следует отметить, что оценка параметров модели M_1 в задаче сегментации, хотя и может быть полезной, в общем случае, не является обязательной. Вместе с тем, некоторые алгоритмы, например, алгоритм адаптации АР-модели при обнаружении последовательности границ, могут неявно производить такую оценку.

С точки зрения функционального подхода, детектор определяет, когда реальный сигнал имеет значительное расхождение с моделью M_1. В этот момент считается, что свойства модели изменились на M_2.

Как описано выше, алгоритм сегментации основан на выборе модели, описывающей исследуемый сигнал. Поскольку адекватность модели определяется поставленной задачей, то и адекватность выделенных сегментов определяется из тех же соображений.

Рассмотрим некоторый сигнал *S*. Пусть имеется *M* - множество моделей, *I* – множество критериев сегментации и $\{A_i\}$- набор алгоритмов сегментации, определенных парой (M_k, I_l), где $M_k \in M$ – модель, $I_l \in I$ – критерий сегментации, выбранные для *i*-го алгоритма. Таким образом, множество алгоритмов определяется как отношение на декартовом произведении $M \times I$. Отметим, что один критерий сегментации может быть применен к разным моделям, и разные модели могут поддерживать различные критерии сегментации. Любая модель из *M* описывает исследуемый сигнал с некоторой степенью адекватности. Пусть в результате сегментации получен набор сегментов $\{S^{A_i}\}$. $S^{A_i} = \{t_r\}$, где t_r – граница сегмента. В общем случае $S^{A_i} \neq S^{A_j}$, $i \neq j$.

Из сформулированной задачи сегментации следует, что основой разрабатываемых алгоритмов сегментации является гипотеза о виде модели M, которая описывает исследуемый сигнал. Выбор модели определяется прикладной задачей (целью), видом сигнала и накопленным опытом в прикладной области исследования.

Авторами предлагается применить алгоритмы сегментации с использованием трех моделей: АР-модели, модели сигналов с постоянной гладкостью, модели сигнала с постоянной частотно-временной структурой.

Классической моделью, применяемой в исследовании геофизических сигналов, является модель авторегрессии – скользящего среднего (АРСС), как частного случая модели авторегрессии (АР-модель). Данная модель позволяет определить изменения в спектральной структуре сигнала [2].

Разработаны два алгоритма сегментации сигналов на основе метода вейвлет-преобразования [3]. Первый алгоритм на основе вейвлет-пакетов лучше приспособлен к описанию широкополосных сигналов с меняющейся частотно-временной структурой и позволяет исследовать нестационарные сигналы, например, сигналы сейсмического шума.

Также предлагается алгоритм сегментации сигналов на основе исследования показателя Гельдера [8]. Данный алгоритм в большей степени ориентирован на исследование низкочастотных сигналов или их сглаженных компонент. Критерием сегментации в данном случае является изменение гладкости сигнала, математической характеристикой которой служит показатель Гельдера. Одним их блоков алгоритма является вычисление непрерывного вейвлет-преобразования (численного варианта НВП). Если этот блок реализован на алгоритме быстрого непрерывного вейвлет-преобразования, то алгоритм может быть использован в реальном масштабе времени. Данный метод основан на предположении, что сигнал описывается виннеровским процессом с меняющимся показателем Херста [5]. Критерием сегментации в этом алгоритме является показатель Гельдера, оцениваемый методом вейвлет-анализа, значение которого совпадает со значением показателя Херста [5].

Предложенные алгоритмы основаны на функциональном подходе к решению задач обработки сигналов, что позволяет применять их без предварительного накопления статистических данных или обучения.

Экспериментальные исследования

Для реализации предлагаемого подхода авторами разработан программный комплекс на базе системы Matlab. Интерфейс программы разработан в Delphi.

Применение метода сегментации на основе модели авторегрессии рассмотрено в работе [3]. Покажем возможность сегментации геофизических сигналов методами на основе вейвлет-преобразования.

На рис. 2 приведен пример сегментации сейсмического сигнала, зарегистрированного вертикальным сейсмометром в скважине на глубине 16 м с частотой дискретизации 128 Гц [9]. Данный сигнал является широкополосным и для его сегментации применялся метод, основанный на вейвлет-пакетном разложении. На верхней части рисунка изображены классы сегментов в каждый момент времени. Параметры алгоритма следующие: вейвлет - db3 (Добеши), число уровней вейвлет-разложения - 3, длина блока (разрешение) - 10 с. Три уровня разложения обеспечивают разбиение частотной области на 8 поддиапазонов. Видно, что преобладающими классами является второй и третий. Это может свидетельствовать о том, что данные сегменты являются фоновыми. Проецируя сегменты на временную область, видно появление новых классов

сегментов в местах, где проявляется аномальное (нефоновое) поведение сигнала (отмечено стрелками на рис. 2).

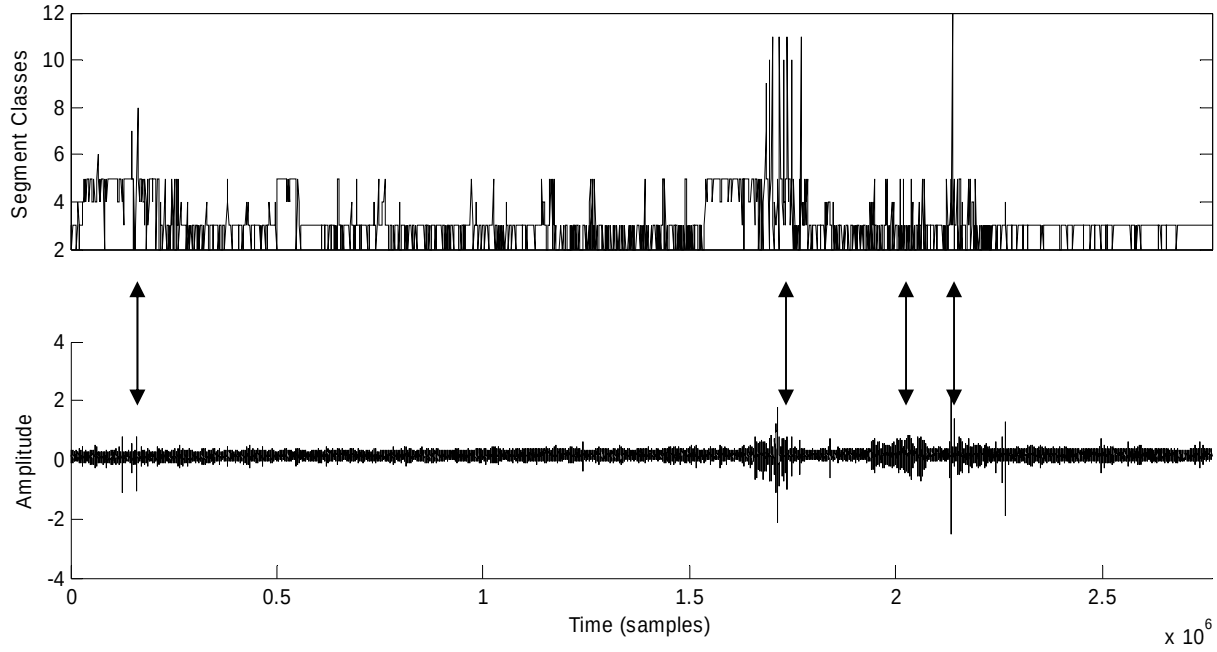
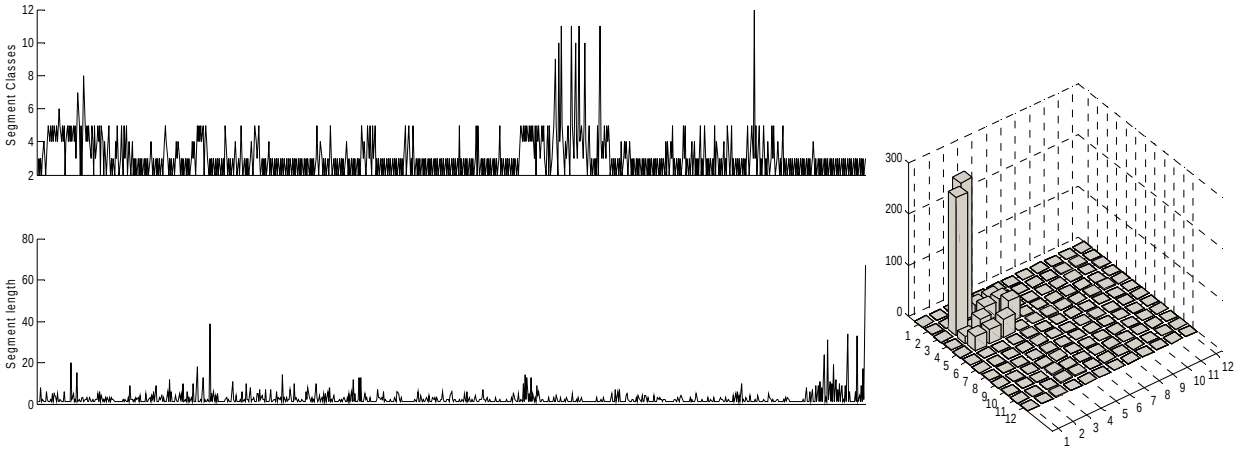


Рис. 2. Пример сегментации сейсмического сигнала.

Интерес представляет закономерности в последовательности появления сегментов сигнала. Для этого рассмотрим модель сигнала, представляющую собой последовательность упорядоченных пар (K, L), где K – класс сегмента, L – его длительность. На рис. 3 а представлены зависимости K (верхняя часть) и L (нижняя часть) от времени. Теперь можно рассмотреть частоты переходов *i*-го сегмента в *j*-й - p_{ij}. Матрицу P, элементами которой являются вероятности перехода p_{ij}, называют переходной матрицей, которая может служить основой для построения статистической модели сигнала на базе скрытых марковских процессов [6]. На рис. 3 б представлена визуализированная переходная матрица для рассматриваемого примера.

Для более формального анализа закономерностей переходов предлагается использовать сиквенциальный анализ [1]. Данный механизм позволяет построить ассоциативные правила вида «если A, то B», где A и B – подпоследовательности сегментов. Т. е. можно сказать, что после подпоследовательности A будет следователь подпоследовательность B, а также подсчитать статистические характеристики данного правила. Поддержка (supp) – показывает, какая часть подпоследовательностей поддерживает данное правило. Достоверность (conf) – показывает вероятность того, что из наличия в транзакции (A∪B) подпоследовательности A следует наличие в ней подпоследовательности B. Считается, что чем больше достоверность, тем лучше правило. В табл. 1 приведены правила, полученные для рассматриваемого сигнала. Все подпоследовательности, имеющие низкую поддержку (меньше 0.6), были отброшены. Правила, полученные по достаточно большой выборке, могут использоваться для прогноза поведения сигнала.

Для апробации алгоритма, основанного на исследовании динамики показателя Гельдера, авторами были использованы данные регистрации концентрации подпочвенного радона. Регистрация проводилась два раза в сутки. Так как, в отличие от сейсмического шума, данные регистрации концентрации подпочвенного радона являются низкочастотными, то применение для сегментации алгоритма, основанного на вейвлет-пакетах, является нецелесообразным. Ниже приведен пример сегментации данного сигнала (рис. 4). Класс сегмента в данном случае определяется квантованным на восемь уровней значением показателя Гельдера.



а б

Рис.3. Зависимость классов сегмента (K) и их длительности (L) от времени *(а),* матрица переходов *(б)*.

Таблица 1

A	-->B	Характеристики правила
2	-->3	supp=0.84514 conf=0.88221
3	-->2	supp=0.86194 conf=0.89975
2 3	-->2	supp=0.7503 conf=0.88778
2	-->3 2	supp=0.7503 conf=0.78321
3 2	-->3	supp=0.78391 conf=0.90947
3	-->2 3	supp=0.78391 conf=0.8183
2 3 2	-->3	supp=0.68908 conf=0.9184
2 3	-->2 3	supp=0.68908 conf=0.81534
2	-->3 2 3	supp=0.68908 conf=0.7193
3 2 3	-->2	supp=0.68307 conf=0.87136
3 2	-->3 2	supp=0.68307 conf=0.79248
3	-->2 3 2	supp=0.68307 conf=0.71303
2 3 2 3	-->2	supp=0.60264 conf=0.87456
2 3 2	-->3 2	supp=0.60264 conf=0.8032
2 3	-->2 3 2	supp=0.60264 conf=0.71307
2	-->3 2 3 2	supp=0.60264 conf=0.62907
3 2 3 2	-->3	supp=0.61945 conf=0.90685
3 2 3	-->2 3	supp=0.61945 conf=0.7902
3 2	-->3 2 3	supp=0.61945 conf=0.71866
3	-->2 3 2 3	supp=0.61945 conf=0.64662

Применяя описанный выше подход к анализу последовательности сегментов, были получены ассоциативные правила, приведенные в таблице 2.

Таблица 2

A	-->B	Характеристики правила
4	-->5 4	supp=0.5419 conf=0.55429
4 5	-->4	supp=0.5419 conf=0.70163
5	-->4	supp=0.78771 conf=0.90385
4	-->5	supp=0.77235 conf=0.79
4	-->3	supp=0.7514 conf=0.76857
3	-->4	supp=0.74441 conf=0.894314

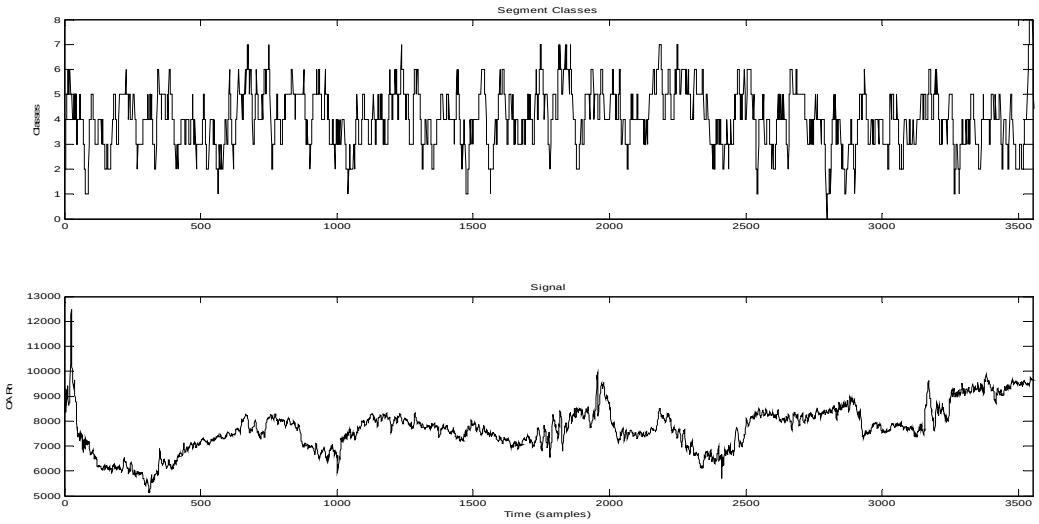


Рис. 4. Пример сегментации сигнала объемной концентрации радона.

Отметим, что полученные в обоих примерах правила зависят от параметров алгоритма сегментации, т. е. от параметров структурной модели сигнала.

Заключение

В результате выполненной работы предложен комплексный подход к исследованию геофизических сигналов с применением технологии Data Mining. Предложенный метод позволяет частично уйти от экспертного метода на этапе первичного анализа и интерпретации сигналов.

Применение методов сегментации и сиквенциального анализа позволяет обнаружить скрытые закономерности сигнала, недоступные непосредственному визуальному анализу.

Список литературы

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

2. Бокс Дж., Дженкинс Г. Анализ временных рядов: прогноз и управление. М.: Мир, 1974. 408 с.

3. Геппенер В.В., Тристанов А.Б., Фирстов П.П. Применение методов сегментации к обработке геофизических данных // Материалы ежегодной конференции, посвященной Дню вулканолога. Петропавловск-Камчатский: Изд-во Наука – для Камчатки, 2005. С. 183-187.

4. Любушин А.А. (мл.) Геофизический мониторинг: шумы, сигналы, предвестники. // Проблемы геофизики 21 века. Книга 2. М.: Наука, 2003. С. 70-94.

5. Малла С. Вейвлеты в обработке сигналов. М.: Мир, 2005. 671 с.

6. Моттль В.В., Мучник И.Б. Скрытые марковские модели в структурном анализе сигналов. М.: Физматлит, 1999. 352 с.

7. Никифоров И.В. Последовательное обнаружение изменения свойств временных рядов. М.: Наука, 1983. 200 с.

8. Тристанов А.Б. Обнаружение изменений в сигнале методом вейвлет-анализа // Труды 2-ой Всеросс. научн. конф. «Проектирование инженерных и научных приложений в среде MATLAB». М.: ИПУ РАН, 2004. С. 1798-1821.

9. Тристанов А.Б., Руленко О.П., Фирстов П.П. Некоторые особенности сейсмического шума в скважине НИС-1 в июле-августе 2003 года // Матер. конф., посвященной дню вулканолога. Петропавловск-Камчатский, 2004. С.82-89.

10. Page E.S. Continuous inspection schemes // Biometrica. 1954. V. 41. P. 100-114.

11. Feng D. Y., Quchi T. New indexes and methods in earthquake prediction research // Asta. Seismol. Sin. 1994. № 4. P. 331-342.