РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2009, том 45, № 1

УДК 519.95, 519.254

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ СПЕКТРАЛЬНЫХ ДАННЫХ*

А. Б. Богданов[1], И. А. Борисова[2], В. В. Дюбанов[3], Н. Г. Загоруйко[2], О. А. Кутненко[2], А. В. Кучкин[1], М. А. Мещеряков[1], Н. Г. Миловзоров[4]

[1]*Институт криминалистики ФСБ РФ, Москва*
[2]*Институт математики им. С. Л. Соболева СО РАН,*
630090, г. Новосибирск, просп. Академика Коптюга, 4
E-mail: zag@math.nsc.ru
[3]*Новосибирский государственный университет,*
630090, г. Новосибирск, ул. Пирогова, 2
[4]*ОАО «ГМК "Норильский никель"»,*
125593, Москва, Тверской бульвар, 13

Описывается программная система «Спектран», предназначенная для автоматизации процессов анализа данных, представленных таблицами типа «объект–свойство», в которой реализованы алгоритмы интеллектуального анализа данных, основанные на функции конкурентного сходства (FRiS-функции). Показано применение системы «Спектран» на примере решения задач анализа множества объектов (микрочастиц исследуемого вещества), описанных спектральными характеристиками. Решаются следующие базовые задачи интеллектуального анализа данных: кластеризация частиц по похожести их спектров, выбор подмножества наиболее информативных каналов спектра, распознавание принадлежности частиц и их смесей к заданным классам и ряд других.

Ключевые слова: интеллектуальный анализ данных, распознавание образов, информативность признаков, функция конкурентного сходства.

Введение. При анализе веществ физическими методами используются различные способы воздействия на образцы и фиксируются зависящие от химического состава вещества реакции образцов на эти воздействия. Примером такого анализа может служить исследование микрообъектов и их совокупностей по данным рентгеноспектрального микроанализа. Исследуемое вещество представляет собой множество из нескольких десятков или сотен микрочастиц (объектов). Реакция каждой микрочастицы при этом методе воздействия отображается спектром, состоящим из 1024 линий (каналов). Амплитуда сигнала в спектральном канале может изменяться от нуля до нескольких тысяч условных единиц. Спектр одного и того же микрообъекта ме-

* Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант № 08-01-00040).

няется в зависимости от контролируемых и неконтролируемых условий эксперимента.

При анализе набора объектов требуется узнать, является ли данная микрочастица представителем некоторого класса веществ. Для этого достаточно определить, укладывается ли значение сигнала заданных каналов спектра в допустимые для этого класса «коридоры» значений. Задача в этом случае состоит в выборе информативных каналов спектра (задача типа X) и построении решающего правила (задача типа D), по которым классы веществ надежно отличаются друг от друга. Если оказывается, что анализируемое множество микрочастиц не принадлежит ни одному из известных классов веществ, то возникает задача формирования новых классов, которая решается методами автоматической кластеризации или таксономии (задача типа S).

Более интересные результаты получаются, если решать задачи не базовых типов (X, D и S), а комбинированного типа, например одновременного выбора признаков и решающих правил (задача типа DX), выбора признаков и таксономии в пространстве заданных признаков (задача типа SX), выбора таксономии и построения решающего правила (задача типа DS).

В программной системе «Спектран» представлены алгоритмы решения задач всех этих типов. Система успешно используется для решения задач анализа данных различных прикладных областей. Кроме спектральных данных анализировались данные, характеризующие свойства набора машинных программ, генетические данные об экспрессии генов, измеренной на здоровых пациентах и пациентах, больных раком и диабетом.

При решении этих задач нужно дать возможность эксперту использовать свои неформализованные знания и интуицию, чтобы корректировать решающие правила, менять состав классов, придавать признакам веса и т. д. При этом система выполняет ту или иную программу и выдает количественную оценку качества разных вариантов решений, что позволяет эксперту выбрать наиболее приемлемый вариант. Интерпретация получаемых результатов облегчается развитым графическим сопровождением.

Далее рассмотрим методы и алгоритмы, реализованные в системе «Спектран».

Методологическая база алгоритмов. В качестве единой основы для построения алгоритмов интеллектуального анализа данных, реализованных в системе «Спектран», используется функция конкурентного сходства или FRiS-функция (Function of Rival Similarity). Эта функция имитирует человеческие способности оценивать меру сходства между объектами и явлениями. Чтобы оценить сходство объекта Z с эталоном образа S_i, нужно знать не только расстояние r_i до образа, но и расстояние r_j до эталона образа S_j, который является ближайшим конкурентом образа S_i. Учет этих расстояний в шкале отношений можно осуществлять с помощью FRiS-функции

$$F_{i/j} = (r_j - r_i)/(r_i + r_j).$$

Здесь величина $F_{i/j}$ характеризует сходство объекта Z с эталоном образа S_i в конкуренции с эталоном образа S_j. Значения этой функции меняются в пределах от –1 до +1. Если контрольный объект Z совпадает с эталоном образа S_i, то $r_i = 0$ и $F_{i/j} = 1$, а $F_{j/i} = -1$, что свидетельствует об абсолютном сходстве объекта Z с эталоном i-го образа и о максимальном его отличии от эталона j-го образа. При расстояниях $r_i = r_j$ значения $F_{i/j} = F_{j/i} = 0$, что указывает на

границу между образами. В точках границы объект в равной степени похож и не похож на эти конкурирующие образы.

Опыт работы с FRiS-функцией показал, что она может служить базовым элементом для построения методов решения различных задач интеллектуального анализа данных (Data Mining). Эти методы описаны в [1, 2]. Здесь приведем лишь их краткое изложение.

Методы построения решающих правил. Построение решающих правил или обучение программы распознаванию образов состоит в поиске среди объектов обучающей выборки таких объектов, которые могут играть роль эталонов своих образов. Сходство с ними затем используется для распознавания контрольных объектов. Если распределения унимодальны и нормальны, эталоны должны располагаться в центрах тяжести образов. Если распределения полимодальны и образы линейно неразделимы, эталоны должны стоять в центрах мод. С ростом сложности распределения число эталонов k должно увеличиваться.

Алгоритм FRiS-Stolp [3] обладает именно такими адаптивными свойствами. Он нацелен на выбор минимального числа эталонов (столпов), которые не только защищают самих себя, но и обеспечивают заданную надежность защиты всех остальных объектов обучающей выборки. Поясним основную идею этого алгоритма на примере распознавания двух образов.

Сначала все объекты первого образа по очереди играют роль его столпа. Для всех остальных объектов вычисляются два расстояния: r_1 до этого столпа и r_2 до ближайшего объекта второго образа. По даным расстояниям определяется мера сходства F. Если она выше некоторого порога F^*, то объект считается защищенным этим столпом. Объект, который защищает наибольшее количество объектов, назначается первым эталоном первого образа. Для объектов, оказавшихся незащищенными, делается та же процедура. Процесс поиска столпов первого образа заканчивается тогда, когда все его объекты оказываются защищенными. Те же процедуры делаются и для объектов второго образа.

После завершения всех этих процедур вычисляется среднее значение функции конкурентного сходства F_S всех объектов обучающей выборки со своими столпами. Эта величина может служить мерой качества обучения.

Процесс распознавания контрольного объекта Z очень прост. Оцениваются его расстояния r_1 и r_2 до двух ближайших столпов, принадлежащих разным образам, и выбирается образ S_i, $i = 1, 2$, со столпом, значение F_i которого максимально.

Выбор подсистемы информативных признаков. Успех решения данной проблемы зависит от того, как организована процедура направленного перебора вариантов признаковых подсистем и по каким критериям оценивается их информативность. Перебор вариантов в системе «Спектран» делается с помощью алгоритма GRAD [4], а в качестве критерия информативности используется FRiS-критерий [5].

Алгоритм GRAD работает следующим образом. Вначале из N исходных признаков формируются вторичные признаки в виде «гранул» мощности 1, 2 и 3. Гранулы мощности 1 – это отдельные наиболее информативные признаки. Из них методом полного перебора выбираются наиболее информативные пары и тройки признаков – гранулы мощности 2 и 3.

Затем гранулы подаются на вход итеративной процедуры AdDel, в которой чередуются операции добавления (Addition) к имеющейся подсистеме n_1 наиболее информативных гранул с операциями исключения (Deletion) из

подсистемы n_2 ($n_2 < n_1$) наименее информативных гранул. По мере роста количества признаков в подсистеме ее информативность растет, достигает максимума и затем начинает уменьшаться. Наличие перегиба кривой информативности позволяет автоматически выбрать подсистему с наилучшими количеством и составом признаков.

Информативность признака тем выше, чем сильнее его значения у объектов одного образа отличаются от значений у объектов другого образа. Это отличие хорошо улавливается средним значением FRiS-функции (F_S) похожести объектов обучающей выборки на свои эталоны. Исследования подтвердили существенное преимущество данного критерия перед широко используемым критерием, который основан на числе ошибок распознавания объектов обучающей выборки в режиме скользящего экзамена (One-Leave-Out) или перекрестной проверки (Cross-Validation). Величина F_S позволяет более точно предсказывать надежность распознавания будущей контрольной выборки.

Автоматическая классификация (кластеризация). Трудность применения FRiS-функции в процессе автоматического разделения M неизвестных микрообъектов на группы (кластеры, таксоны) состоит в том, что вначале все объекты как бы принадлежат одному классу и нет возможности оценить расстояние до конкурентного класса. По этой причине в алгоритм FRiS-Cluster [6] вводится виртуальный класс, объекты которого удалены от каждого реального объекта на расстояние r_2^*. При известном расстоянии r_1 от объекта a_j до столпа a_i появляется возможность оценивать функцию

$$F = (r_2^* - r_1)/(r_2^* + r_1).$$

Если $F > F^*$, объект a_j считается принадлежащим кластеру со столпом a_i. Объект, который в роли столпа набирает в свой кластер наибольшее число объектов, становится столпом кластера C_1.

Повторением таких процедур удается разделить всю выборку на k кластеров. Среднее значение функции сходства F_S всех объектов со своими столпами характеризует качество кластеризации.

Меняя пороговое значение F^*, можно изменить число кластеров k. Если задать диапазон приемлемых значений k, то программа найдет вариант с максимальным значением F_S в этом диапазоне. Обычно выбирается вариант, который совпадает с вариантом, предлагаемым экспертом.

Если кластеры близко расположены друг к другу, их целесообразно объединять в один таксон. В программе предусмотрена процедура проверки условий объединения. После ее работы могут получаться классы, в состав которых входит несколько кластеров (подклассов). Эти классы могут иметь произвольную форму и не обязательно разделяться линейными границами.

Использование FRiS-функции в логических решающих правилах. Логические решающие правила (ЛРП) в виде деревьев [7] отличаются простотой использования и наглядностью получаемых результатов, но не обладают некоторыми из свойств, имеющихся у FRiS-функции. В частности, ЛРП не указывают типичных представителей классов, не оценивают качество обучения и надежность принятого решения. По этой причине в систему «Спектран» включен вариант ЛРП, дополненный некоторыми свойствами функции сходства, – алгоритм FRiS-LRP. Эти дополнения начнут работать после того, как алгоритм ЛРП построит решающее правило в виде допусти-

мых значений сигнала в выбранных частотных каналах (т. е. построит последовательность допустимых коридоров). Для объектов обучающей выборки, попавших в коридор данного признака, создается виртуальный объект, играющий роль столпа на этом признаке. Затем находится реальный объект, сходство которого со столпами на всех признаках максимально. Этот объект является типичным представителем класса. Среднее значение функции сходства всех объектов со своими столпами (F_S) служит оценкой качества обучения. По значению FRiS-функции распознаваемого объекта со своим столпом можно определить надежность принятого решения о принадлежности этого объекта к данному классу.

Алгоритмы комбинированного типа. В системе «Спектран» реализованы следующие алгоритмы комбинированного типа [8]:

1. Алгоритм FRiS-DX представляет собой комбинацию алгоритма FRiS-Stolp для построения решающего правила (D) и алгоритма FRiS-GRAD для выбора информативных признаков (X). Результатом работы такого алгоритма является набор из заданного количества наиболее информативных подпространств с указанием их информативности, при этом для каждой подсистемы определяется список столпов и перечень объектов, защищаемых ими.

2. Алгоритм FRiS-LRP также решает задачу типа DX и выдает логическое решающее правило для распознавания каждого класса. Если контрольный объект уложился в допустимые коридоры для более чем одного класса, решение принимается по максимальной мере его сходства с типичными представителями конкурирующих классов.

3. Алгоритм FRiS-Class решает задачу типа SD – автоматической таксономии (S) с одновременным построением решающего правила (D) в виде набора столпов кластеров.

Интерфейс с пользователем. Первой функциональной частью системы, с которой сталкивается пользователь, является окно «Мастер» интерактивного диалога с пользователем, служащее для быстрого выполнения последовательности типовых действий. «Мастер» (рис. 1) знакомит пользователя с

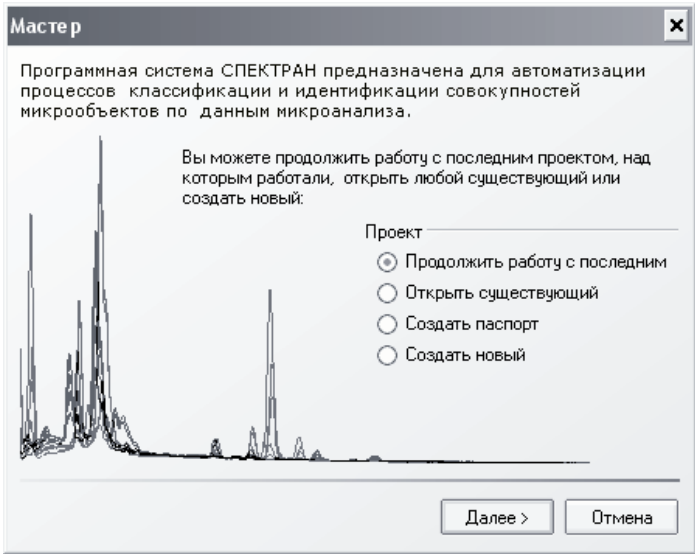

Рис. 1

системой «Спектран» и предлагает базовый набор действий с проектами: «Продолжить работу с последним», «Открыть существующий», «Создать новый проект или создать паспорт».

Последовательность открывающихся в диалоге слайдов дает возможность пользователю задавать основные параметры создаваемого проекта, строить решающие правила, выполнять процедуры таксономии и распознавания. При выборе опции «Открыть существующий» открывается окно, в котором показывается список проектов и информация о выбранном проекте. Это облегчает пользователю навигацию среди массы проектов, созданных к текущему моменту. При выборе опции «Создать новый» открывается последовательность окон, позволяющая выбрать один из способов хранения исходных данных (все спектры в одном файле или каждый спектр в своем файле), имя файла, в котором хранятся данные, вид сглаживания и нормировки. Здесь же можно указать, какие линии спектра нужно использовать: весь спектр или заданное подмножество каналов. Это подмножество задается заранее заготовленным списком или установкой «птичек» на номерах выбираемых каналов.

Если классы анализируемых веществ определены, то для решения задачи распознавания появляется окно, в котором можно указать нужный вид работы (построить решающие правила или распознать контрольные объекты) и отметить, для каких классов строятся решающие правила. В следующем окне (рис. 2) выбирается тип решающих правил: FRiS-GRAD, FRiS-LRP или экспертные правила. Здесь же можно указать параметры настройки алгоритмов (сколько решающих правил выдать, максимальное число признаков в правиле и пр.).

После нажатия кнопки Старт, например, строятся решающие правила и появляется результат в виде экрана с окнами, в которых указаны имена объектов распознаваемых классов. Распознавание может проводиться как на основании обучающей выборки и созданным по ней решающим правилам, так и по паспортам – накопленным в базе описаниям веществ с указанием но-

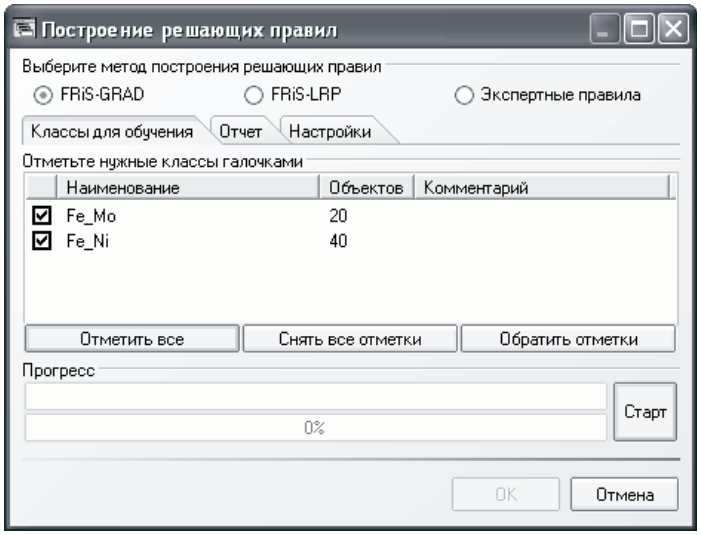

Рис. 2

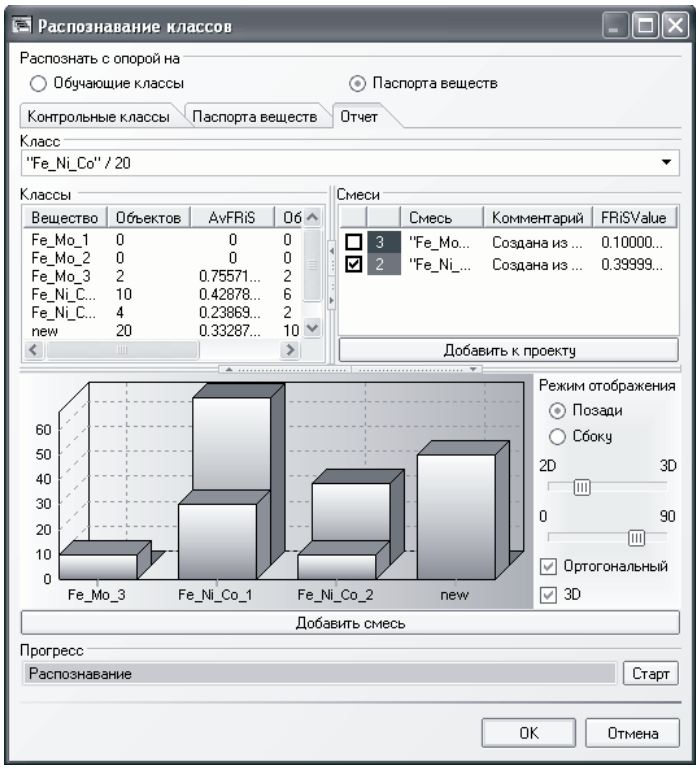

Рис. 3

меров каналов и коридоров допустимых значений сигнала в этих каналах. Результат распознавания представляется как в виде количественных показателей, так и в виде графиков (рис. 3).

Отобранные по результатам распознавания паспорта могут быть непосредственно добавлены к текущему проекту. Далее к ним применяются методы автоматического построения решающих правил FRiS-GRAD или FRiS-LRP, на основе результата работы которых осуществляется более точная процедура распознавания контрольной выборки.

В окне «Редактирование смесей» (рис. 4) отображаются результаты следующего уровня распознавания – распознавания смесей, на основании которых возможно автоматически создать новую смесь и, внеся необходимые поправки и уточнения, пополнить имеющуюся базу смесей.

Помимо разветвленного инструментария создания экспертных описаний исследуемого вещества пользователю доступна возможность проведения автоматической кластеризации. В результате работы алгоритма FRiS-Cluster предлагаются не только различные варианты таксономии (для разного числа кластеров), но также и количественная оценка каждого из вариантов разбиения в наглядной для человека форме – значение локального максимума на графике качества кластеризации (рис. 5). Заметим, что эксперт может внести любые изменения в результаты кластеризации. Исходя из своих собственных соображений, он может породить любое число новых классов путем дробления или объединения имеющихся кластеров.

Одной из важнейших задач разработчиков системы «Спектран» было создание удобного функционального и наглядного рабочего места эксперта.

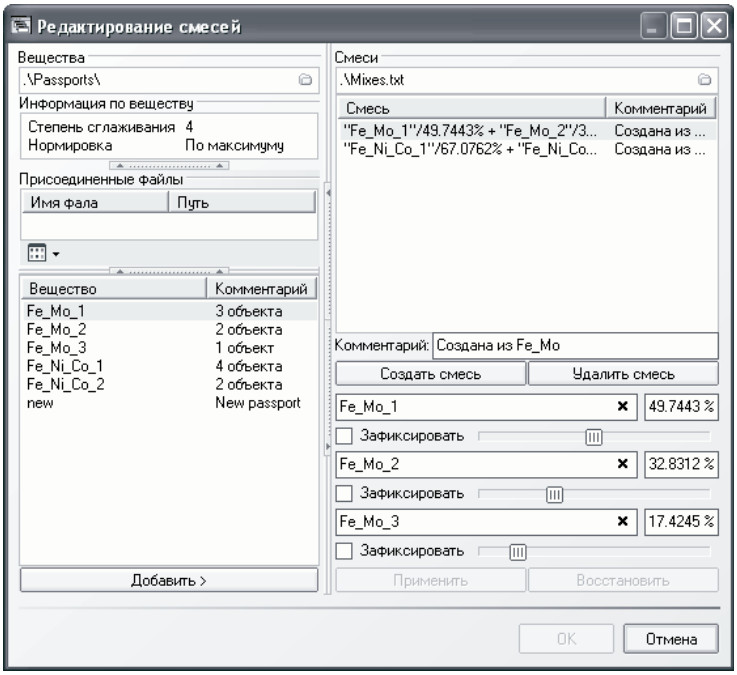

Рис. 4

Данная концепция подразумевает не просто большой набор разумных переборных алгоритмов, работающих за приемлемое время. Преследовалась цель – организовать единое рабочее пространство как с точки зрения общности настроек и форматов данных, так и с точки зрения общности интерфейс-

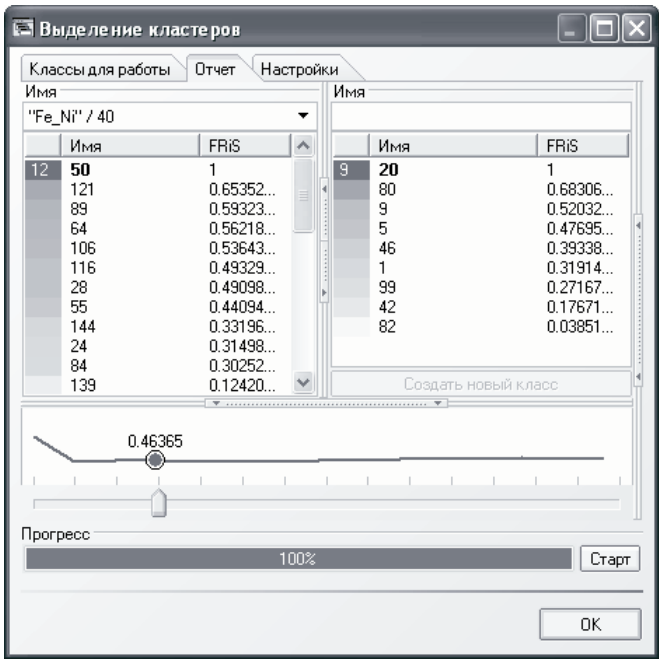

Рис. 5

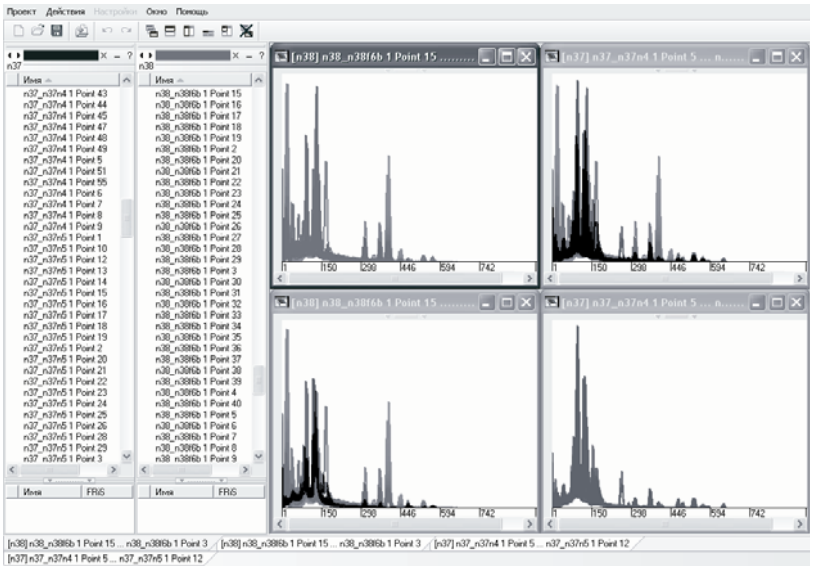

Рис. 6

ных решений, визуальных представлений данных и результатов работы. Это обеспечивается широким набором визуальных приемов, контекстных меню, возможностью передачи данных из одной части системы в другую простым переносом мышью.

Визуальное представление спектральных данных включает в себя большой набор инструментов, начиная с масштабирования и прокрутки диаграмм и заканчивая выделением цветом и подсветкой спектров или отдельных спектральных каналов. Немаловажной является возможность одновременного просмотра на экране нескольких групп спектров (рис. 6).

Несмотря на обилие методов работы система несложна в освоении, так как большинство интерфейсных решений давно и прочно являются стандартом в графических системах типа Windows и интуитивно понятны любому опытному пользователю. Кроме того, наряду с обширным и гибко настраиваемым набором функций она имеет целый ряд типовых решений, которые выполняются по умолчанию и не требуют от пользователя углубления в тонкости настроек.

Заключение. Система «Спектран» представляет собой функционально насыщенный и дружественный инструмент, позволяющий автоматизировать все основные процессы анализа объектов любой природы, описанных количественными признаками. Сейчас ведутся разработки методов анализа таблиц, в которых одновременно присутствуют признаки, измеренные в количественных, порядковых и номинальных шкалах.

СПИСОК ЛИТЕРАТУРЫ

1. **Загоруйко Н. Г.** Интеллектуальный анализ данных, основанный на функции конкурентного сходства // Автометрия. 2008. **44**, № 3. С. 31–40.

2. **Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Methods of recognition based on the function of rival similarity // Pattern Recogn. and Image Analisys. 2008. **18**, N 1. P. 1–6.

3. **Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.** Использование FRiS-функции для построения решающего правила и выбора признаков (задача комбинированного типа DX) // Тр. Всеросс. конф. «Знания – Онтологии – Теории» (ЗОНТ-07). Новосибирск: Изд-во ИМ СО РАН, 2007. Том 1. С. 37–44.

4. **Загоруйко Н. Г., Кутненко О. А.** Алгоритм GRAD для выбора признаков // Тр. VIII Междунар. конф. «Применение многомерного статистического анализа в экономике и оценке качества». М.: Изд-во МЭСИ, 2006. С. 81–89.

5. **Борисова И. А., Загоруйко Н. Г., Кутненко О. А.** Критерии информативности и пригодности подмножества признаков, основанные на функции сходства // Заводская лаборатория. 2008. **74**, № 1. С. 68–71.

6. **Борисова И. А., Загоруйко Н. Г.** Функции конкурентного сходства в задаче таксономии // Тр. Всеросс. конф. «Знания – Онтологии – Теории» (ЗОНТ-07). Новосибирск: Изд-во ИМ СО РАН, 2007. Том 2. С. 67–76.

7. **Лбов Г. С., Старцева Н. Г.** Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Изд-во ИМ СО РАН, 1999. 212 с.

8. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний. Новосибирск: Изд-во ИМ СО РАН, 1999. 273 с.

Поступила в редакцию 2 июня 2008 г.