УДК 004.032.26 (06)

ИНФОРМАЦИОННАЯ НЕЙРОСЕТЕВАЯ СИСТЕМА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Н.В. Замятин, В.П. Максимов, Н.В. Платонов, М.Н. Тарасевич

Томский университет систем управления и радиоэлектроники
E-mail: zam@fet.tusur.ru

Изложена методика нейросетевого анализа данных (data mining). Показано, что применение нейронных сетей Кохонена позволяет эффективно выделять группы связанны данных и определять между ними закономерности. Разработана информационная система для геофизической предметной области.

Развитие технологии интеллектуального анализа данных (ИАД) вызвано рядом объективных факторов. Главные из них: накопление большого количества данных в различных предметных областях и скорость накопления информации намного превысила скорость ее обработки. Только научные учреждения за один день записывают информации примерно на 1 терабайт (по данным аналитического отдела американской компании GTE). При этом наука является не самым большим источником данных, и существуют огромные базы данных в сфере коммерции, энергетики, геологии, медицине, управляющих структурах оргсистем.

На сегодняшний день технология ИАД содержит множество различных подходов к обнаружению знаний. Каждый из них имеет свои преимущества и недостатки. При этом выбор конкретного подхода определяется спецификой предметной области и организацией данных [1]. Целесообразно применение ИАД для выявления знаний в большом количестве данных сейсмической разведки полезных ископаемых.

При обработке результатов в сейсморазведке для построения прогностических моделей традиционно используется аппарат многомерной линейной регрессии. Его использование имеет следующие сложности и ограничения:

- Ограничение классом линейных зависимостей. Искомая прогностическая модель сразу предполагается линейной. Хотя и возможно использование нелинейного преобразования независимого параметра, эта процедура носит достаточно произвольный характер, и итоговая модель все равно будет линейно зависеть от преобразованных параметров.
- Сложность выделения влияющих параметров. Из-за большого количества динамических параметров некоторые из них даже не рассматриваются как кандидаты на участие в модели. С другой стороны, в модель, согласно требованиям классической статистики, включаются только независимые параметры. Однако, зависимые параметры, в совокупности, могут также нести ценную информацию о целевом параметре.
- Зашумленность входной информации. Сейсморазведка района проводится в течение нескольких лет различными исследовательскими партиями, с использованием различного оборудования и т. п. Поэтому, зачастую в данных разведки встречаются выбросы – резко нетипичные значения. Эти выбросы значительно влияют на строимые линейные модели.

Аппарат нейронных сетей свободен от перечисленных недостатков: нейросети могут аппроксимировать любую непрерывную функцию, автоматически проводят анализ чувствительности влияния входных параметров на результат, устойчивы к шуму в исходных данных.

В качестве основных функциональных требований к интеллектуальным нейросетевым системам можно выделить:

- совместимость форматов хранения информации с наиболее распространенными средствами табличной обработки данных (MS Access, MS Excel).
- возможность обработки не только числовой, но и текстовой информации,
- возможность нормализации исходных обучающих данных различными способами,
- классификация обучающих данных посредством обучения нейросети.
- визуализация полученных карт различными способами.

Выделенные функциональные требования упорядочены в соответствии с жизненным циклом процесса ИАД, что позволяет создать информационную систему интеллектуального анализа данных (ИСИАД), предназначенную для решения задач кластеризации и классификации разнородной информации. Способом классификации данных, реализованным в системе, является нейронная сеть Кохонена.

Нейронная сеть Кохонена (самоорганизующаяся карта Кохонена) решает задачи классификации многомерных векторов. Достоинством сети, по сравнению с другими алгоритмами, является легкость визуализации и интерпретации полученных результатов. Обучение сети проходит без учителя, только на основе выборки входных данных (так называемое неуправляемое обучение).

Различные типы визуализации обученной сети позволяют легко выявить структуру входной информации: унифицированная матрица расстояний

отображает кластерную структуру данных, график компонентов позволяет установить форму зависимости входных параметров, плоскость компонентов и карта попаданий отражают распределение входных параметров.

Структурная схема ИСИАД приведена на рис. 1.

Модуль предобработки выполняет создание и хранение таблицы обучающих данных. Дополнительно модуль обеспечивает нормализацию таблицы обучающих данных. В качестве средства хранения была выбрана СУБД MS Access, являющаяся составной частью модуля.

Функция «Импорт данных» в ИСИАД реализуется средствами СУБД. Поэтому важным критерием выбора СУБД становится развитость ее средств обмена данными.

Функция «Создание классификаторов» сопоставляет текстовым значениям категориальных (перечислимых) данных числовой код. После такого сопоставления все атрибуты принимают только числовые значения, и данные являются подготовленными для использования в алгоритме обучения.

Нормализация данных осуществляется по указанию пользователя, масштабированием значений каждого параметра в диапазон [0;1]. Нормализуются уже подготовленные данные.

Модуль обучения нейросети реализует итеративный алгоритм обучения карты Кохонена.

Функция «Настройка параметров» данного модуля позволяет настроить конфигурацию самой карты (размеры карты и тип), а также параметры ее обучения (количество итераций, способ инициализации и др.).

Функция «Обучение» непосредственно позволяет обучить сеть Кохонена. Для доступности данной функции предварительно должны быть подготовлены исходные данные.

Расчет качества классификации может быть выполнен только после обучения нейросети.

После обучения карты Кохонена дополнительно может быть выполнена ее кластеризация алгоритмом *k*-средних. Алгоритм *k*-средних может применяться непосредственно к исходным данным. Однако, данный алгоритм имеет ряд недостатков и вычислительно сложен. К тому же, для классификации новых данных требуется новая прогонка алгоритма.

Модуль визуализации является одним из важнейших модулей, так как анализ полученной карты строится на различных способах ее графического отображения.

Данный модуль реализует четыре способа отображения карты Кохонена: унифицированная матрица расстояний, карта попаданий, плоскость значений некоторого компонента, график компонентов.

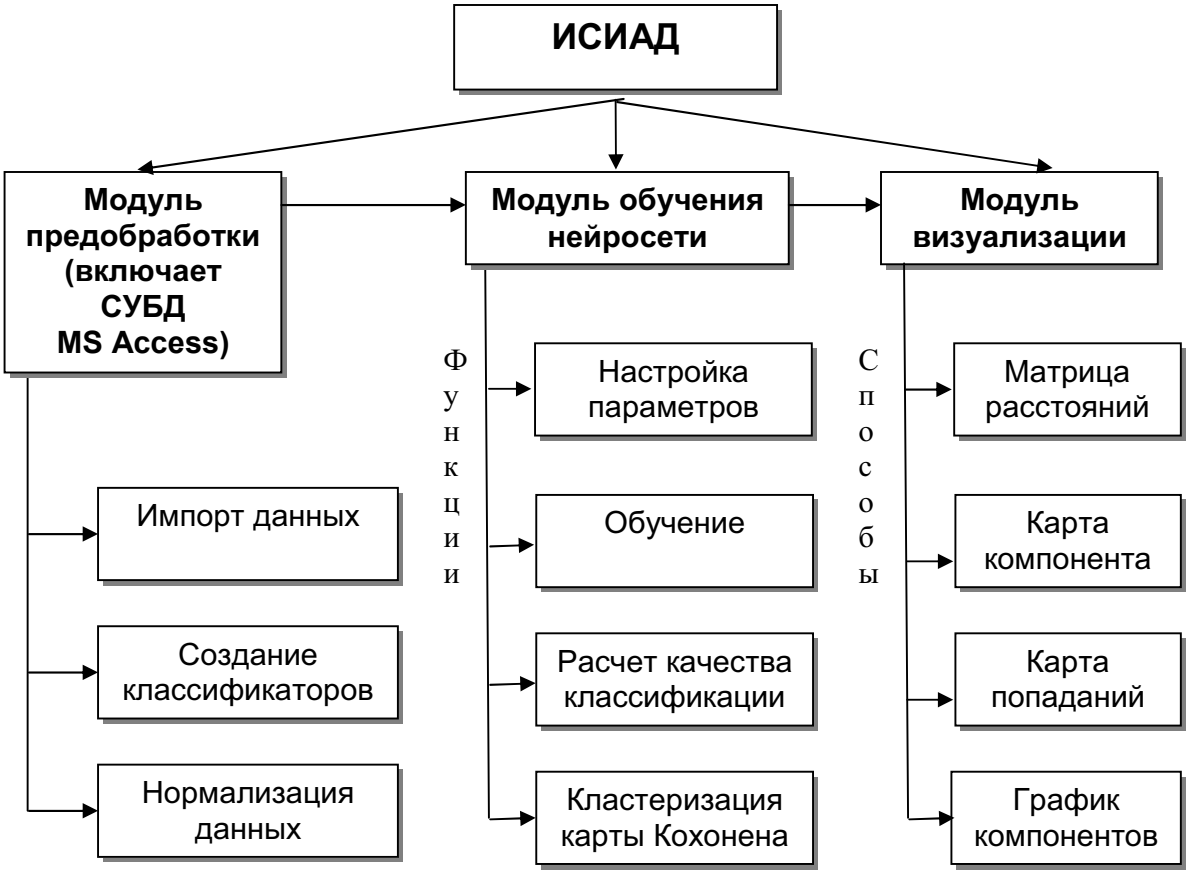

Рис. 1. *Структурная схема ИСИАД*

Визуализация карты требует дополнительных промежуточных вычислений. Результаты этих вычислений сохраняются в базе данных для последующего использования.

В качестве исходных данных для тестирования информационной системы использованы данные сейсмической разведки. Процесс сейсмической разведки заключается в проведении последовательных взрывов зарядов на местности через определенные расстояния (50, 100 и т. д. метров). После каждого взрыва, установленные датчики фиксируют параметры взрывной волны, отраженной от геологических пластов. Линия, вдоль которой проводятся взрывы, называется профилем.

Для построения прогностических моделей карты параметров дискретизируют и используют сетки параметров – значения параметра в узлах регулярной сетки (шаг сетки – 100 м).

Основной задачей сейсморазведки является прогнозирование значений параметров, влияющих на содержание нефти в пластах породы, вдоль карты района. Такими параметрами являются: пористость и проницаемость горизонта. Исходными данными для прогноза являются сетки структурных и динамических параметров, соотнесенные с пробами пластов.

Обработка полученной информации позволяет выделить различные акустические характеристики пластов (горизонтов). Причем количество этих параметров может быть достигать до 200 и более. Примеры параметров: время прихода отраженной от горизонта волны, средняя энергия, амплитуда волны, фаза и т. д.

Параметры делятся на структурные, относящиеся к профилю разведываемой поверхности, (время прихода волны, глубина) и динамические, описывающие отраженную взрывную волну (амплитуда, энергия).

Значения параметров измеряются вдоль взрывных профилей. Затем проводится интерполяция параметров между профилями. В итоге получается карта параметра – графическое изображение распределения значения параметра по равномерной сетке. Всего, в виде сеток, получены значения 26 параметров. Все они участвовали в обучении нейросети. Задачей являлось установление зависимостей между параметрами и степени их влияния на целевой параметр пористости.

Пример карты параметра приведен на рис. 2.

Пунктирными линиями на рисунке показаны профили взрывов. Значения входных параметров могут быть определены не в каждой точке сетки, например, из-за отсутствия вблизи данной точки

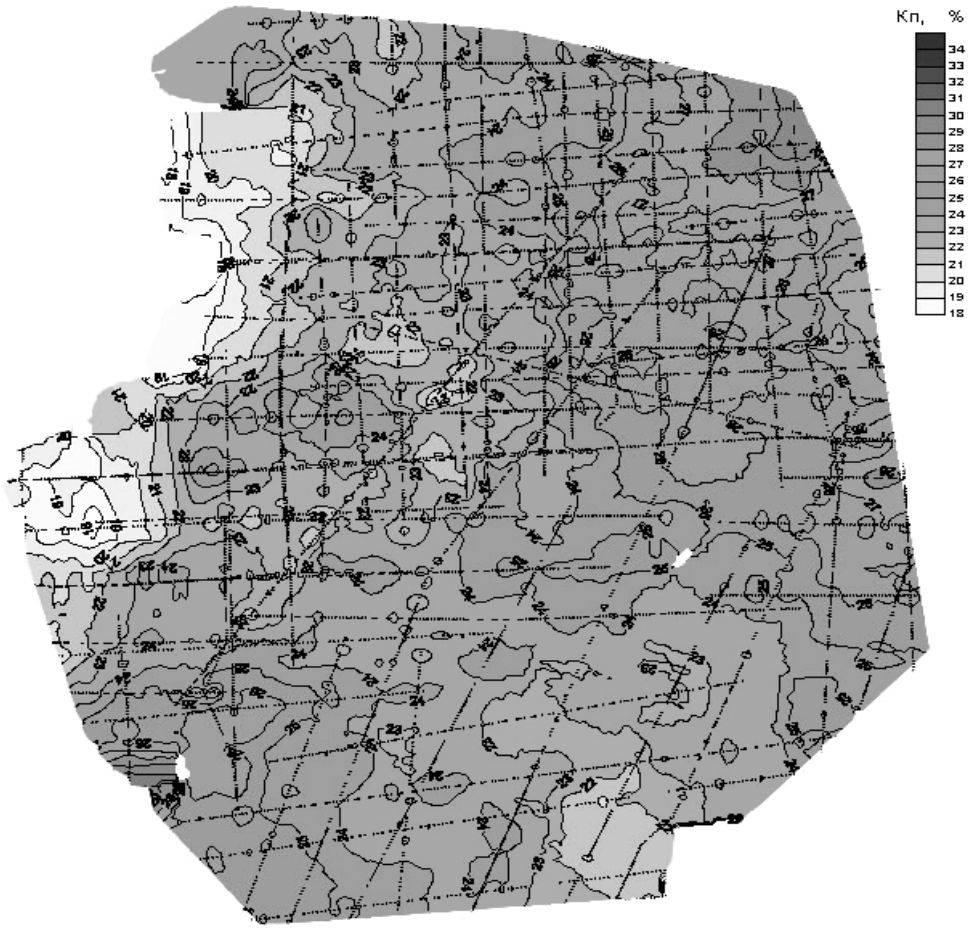

Рис. 2. *Карта параметра «Пористость горизонта b9», b9 – значения сетки*

взрывного профиля. Поэтому в исходной таблице данных содержатся пропущенные значения. Строки, содержащие пропущенные значения, были удалены из таблицы (очистка данных). После очистки осталось 54509 записей.

Перед обучением нейросети процесс подготовки данных заключается в присвоении числовых значений текстовым параметрам и (если задан соответствующий параметр нейросети) нормализации полученной таблицы.

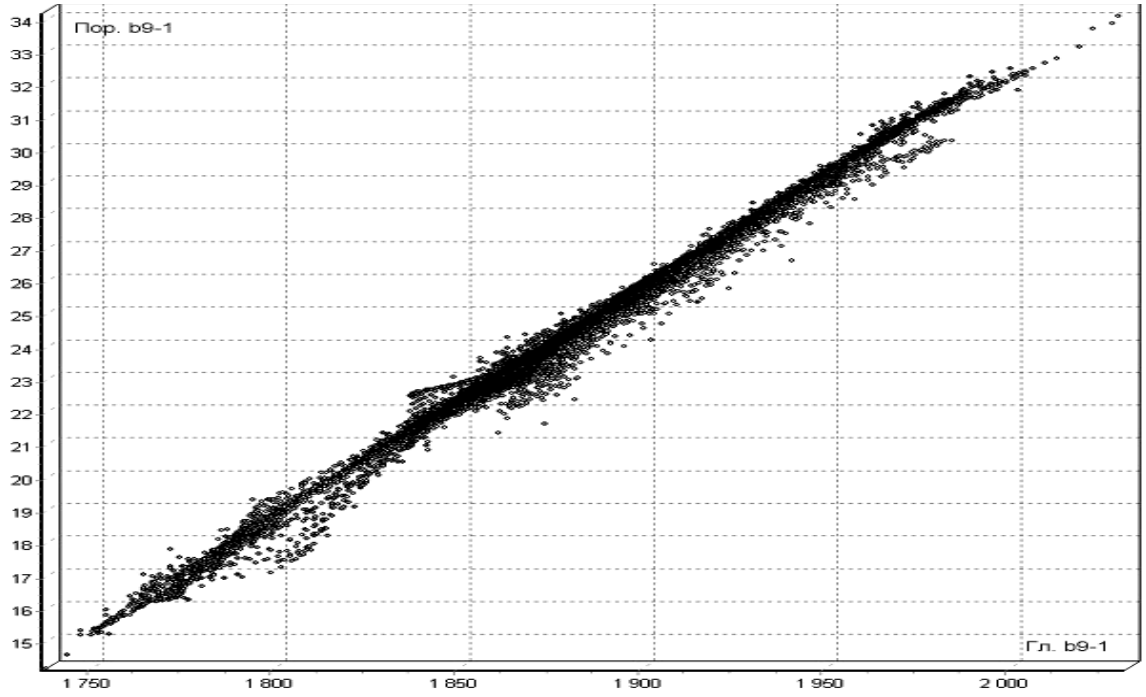

Рис 3. *График компонентов для пары (Гл. b9-1, Пор b9-1)*

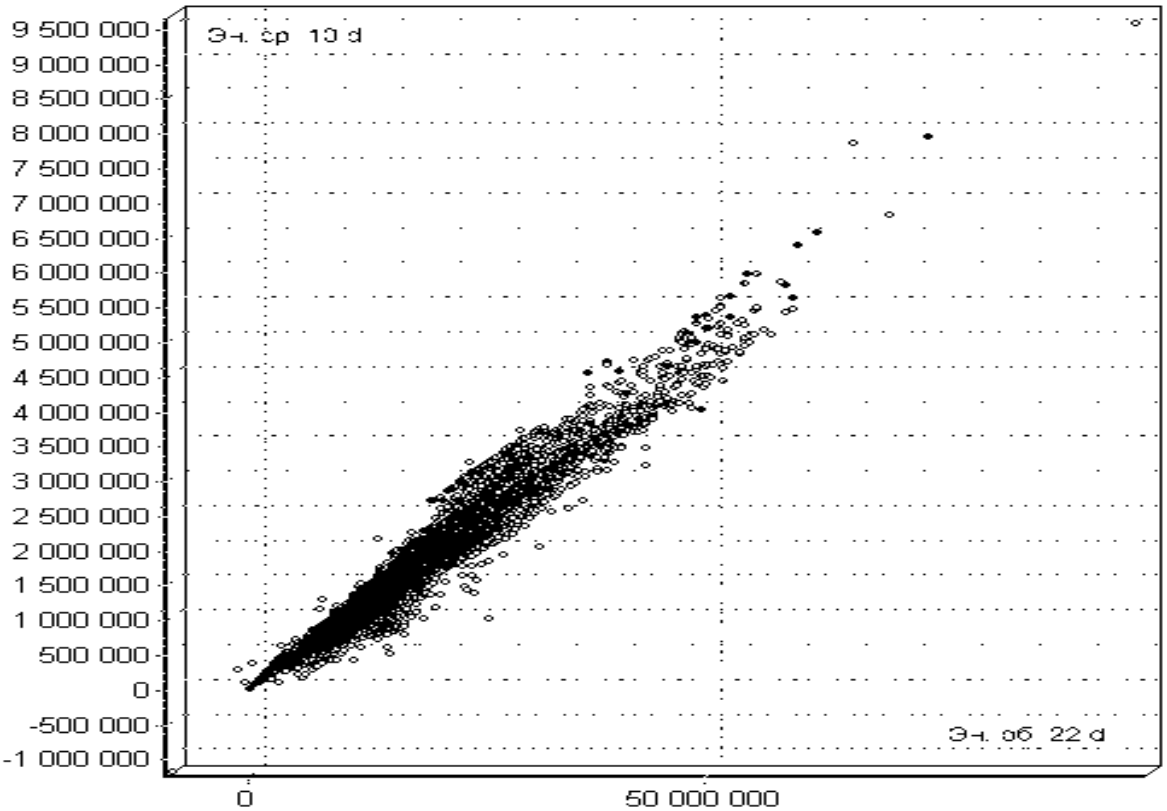

Рис 4. *График компонентов для пары (Эн. общ. 22 d, Эн. сред. 10 d)*

Для поиска зависимостей между параметрами используют визуализированные плоскости компонентов. Для этого в ИСИАД формируются плоскости всех компонентов и осуществляется их перегруппировка, располагающая сходные плоскости рядом. Такое расположение позволяет легко выделить зависимые компоненты входных векторов, поскольку для них в одинаковых местах карты будут находиться схожие шаблоны карты. Выявлено четыре группы коррелирующих компонентов. Установление конкретных форм зависимостей проводится при помощи графиков компонентов.

На рис. 3 приведен график компонентов для пары данных (Глубина b9-1, Пористость b9-1).

Анализ рисунка позволяет выявить четкую линейную зависимость между этими параметрами. Установление параметров данной зависимости может быть проведено при помощи любого из статистических методов. График компонентов для пары (Энергия. общая. 22 d, Энергия. средняя. 10 d) приведен на рис. 4.

Между данными параметрами также существует линейная зависимость. График компонентов для пары (Общая. абсолютная. амплитуда. b9-1, Энергия. общая. b9-1) приведен на рис. 5.

Видно, между параметрами энергии и амплитуды существует зависимость, близкая к квадратичной.

Анализ графиков других пар компонентов позволяет установить факты существования между ними линейных зависимостей. Данная группа параметров является сильно коррелирующей между собой. Поэтому, при построении модели численного прогноза в число значимых параметров следует включать только один из них.

Четвертая группа коррелирующих параметров интересна тем, что о данной зависимости не было ничего известно до начала процесса ИАД. На рис. 6 приведен график компонентов для пары (Дискрета 44 IIa, Проницаемость b9-12 16ob).

Зависимость между компонентами подобна линейной, однако разброс точек от основной линии достаточно велик. Это говорит о том, что на параметр проницаемости влияют также и другие параметры, не вошедшие в анализ.

Проведенный анализ не выявил во входной информации кластерной структуры. Это связано с тем, что все параметры являются численными характеристиками и имеют непрерывный числовой диапазон изменения, причем большинство из них независимо. В связи с этим, нельзя было сравнить распределение значений компонентов по различным кластерам, сравнить «населенность» кластеров (используя карту попаданий) и на этой основе дать описание типичных представителей кластера.

В то же время проведенный интеллектуальный анализ данных позволил:

- получить наглядное представление о структуре входной информации.
- выделить группы зависимых компонент и выявить тип зависимостей между компонентами.

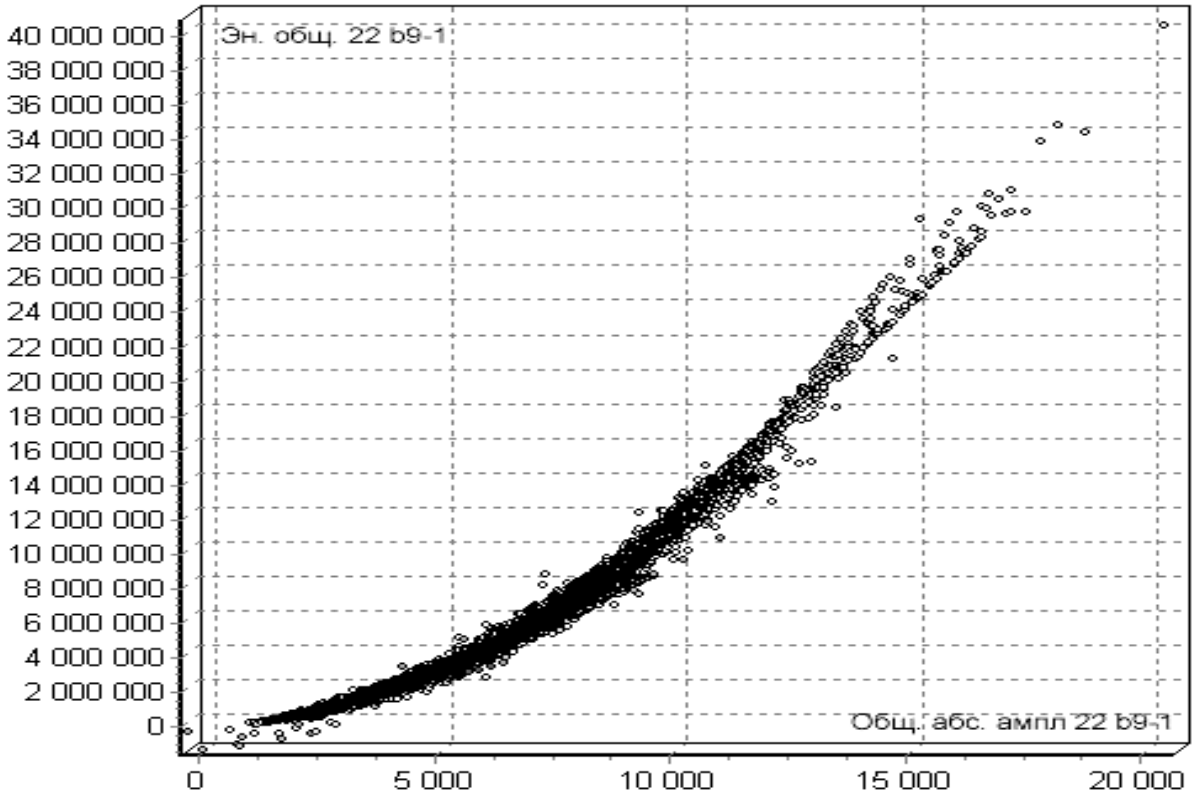

Рис 5. *График компонентов для пары (Общ. абс. ампл. 22 b9-1, Эн. общ. 22 b9-1)*

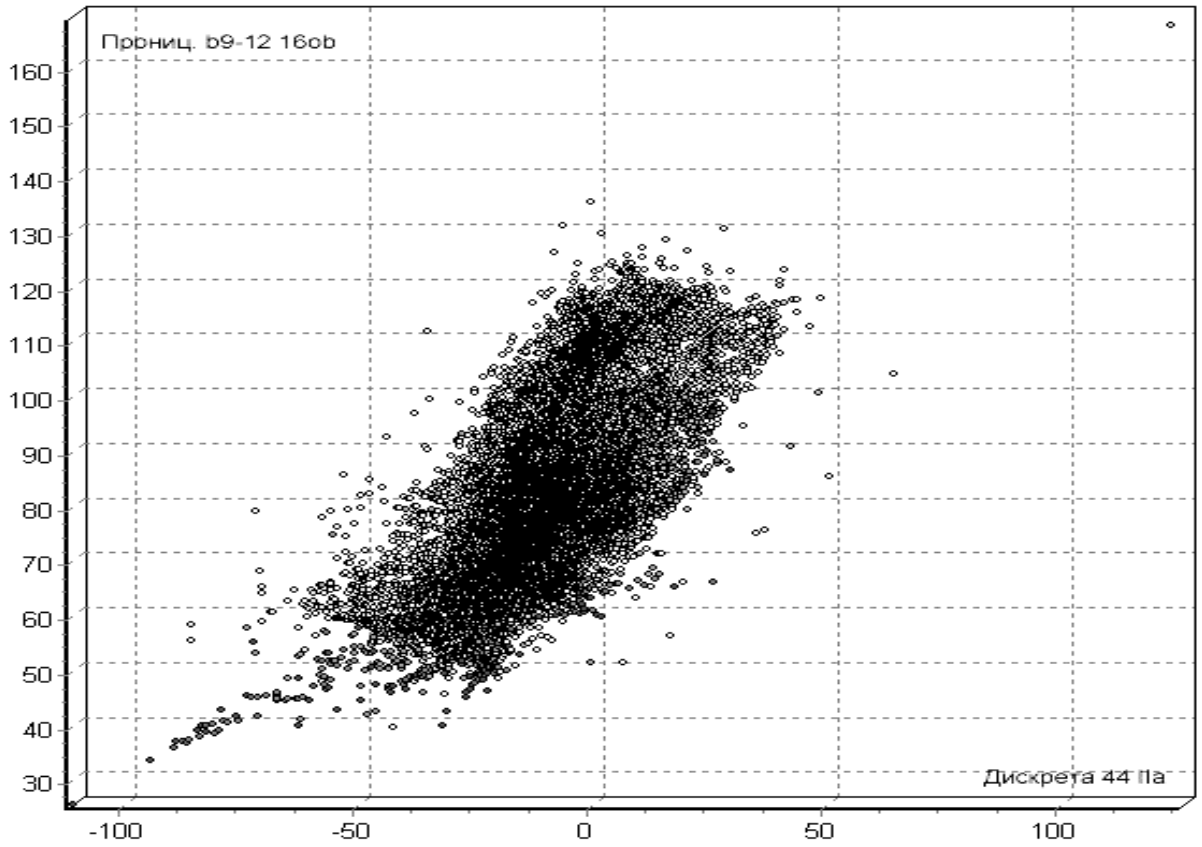

Рис 6. *График компонентов для пары (Дискрета 44 IIa, Прониц b9-12 16ob)*

- сделать вывод об отсутствии кластерной структуры входных данных, что обусловлено независимостью большей части параметров.
- выделить четыре группы коррелирующих параметров, содержащие от двух до восьми компонент.
- установить формы зависимостей параметров внутри групп.

Полученная информация может быть использована при построении числовых моделей прогнозирования целевых параметров, либо с помощью классических статистических методов, либо с помощью искусственных нейронных сетей с обратным распространением ошибки, что открывает большие возможности по интерпретации полученных результатов. Определение группы параметров, коррелирующих с целевой функцией, позволяет сразу выделить значимые компоненты, которые должны войти в числовую модель прогноза, и, тем самым, сократить размерность задачи.

Разработанная информационная система не привязана к конкретной предметной области. Это позволяет аналитику применять ее в любой области, после соответствующей предобработки накопленной в ней информации.

СПИСОК ЛИТЕРАТУРЫ

1. Корнеев В.В., Гареев А.Ф., Васютин С.В. Базы данных. Интеллектуальная обработка информации. – М.: Нолидж, 2000. – 352 с.
2. Уоссермен Ф. Нейрокомпьютерная техника: теория и практика. Пер. с англ. – М.: Мир, 1992. – 127 с.
3. Александров В.В. Интеллект и компьютер. – СПб.: Анатолия, 2004. – 285 с.
4. Львов В.В. Создание систем поддержки принятия решений на основе хранилищ данных // Системы управления базами данных. – 1997. – № 3. – С. 30–40.