Глава 3. Навигация по картам

3.1. Описание программы ViDa Expert 1.0

3.1.1. Внутренняя структура объектов.

Программа ViDa Expert имеет внутреннюю иерархию объектов. Некоторые из них соответствуют тем объектам, с которыми оперирует исследователь на практике, другие объекты являются контейнерами, содержащими и упорядочивающими объекты исследования. Знакомство с внутренней структурой объектов необходимо пользователю для осмысленного использования программы. На рис. 36 изображены сами объекты системы и отношения между ними.

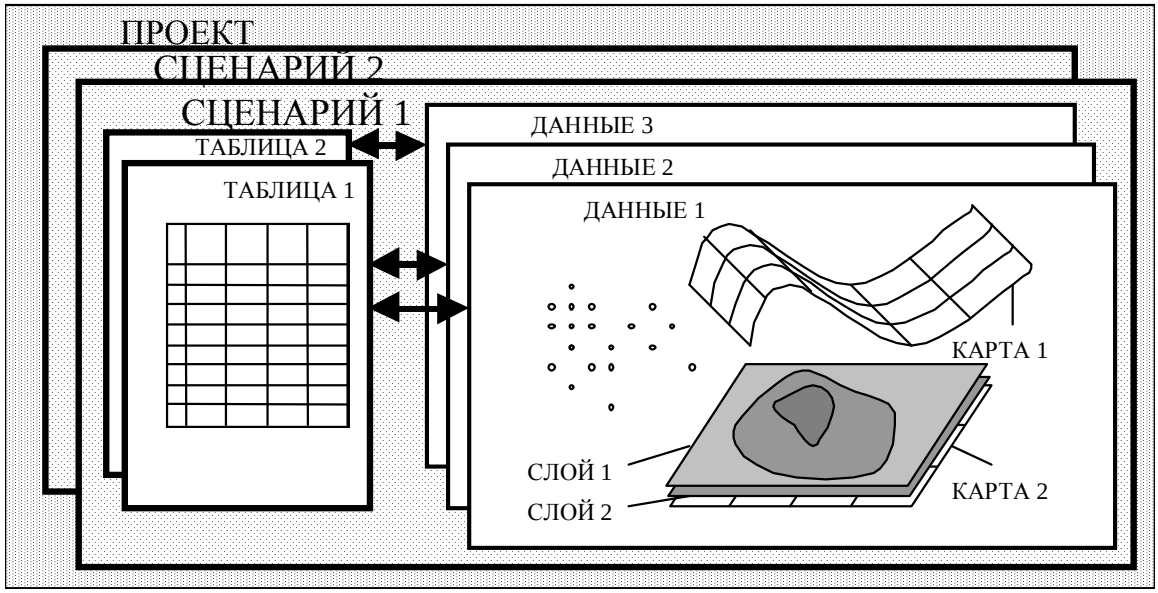


Рис.37. Внутренняя структура объектов программы ViDa Expert

Объектом-контейнером верхнего уровня является ПРОЕКТ, который содержит в себе несколько СЦЕНАРИЕВ. СЦЕНАРИЙ – это совокупность определенным образом настроенных наборов данных и карт. Каждый сценарий как объект-контейнер содержит в себе набор объектов ТАБЛИЦА и набор объектов ДАННЫЕ.

Объект типа ТАБЛИЦА предназначен для хранения исходной табличной информации. В программе ViDa Expert реализовано три способа заполнить таблицу данных – через стандартные файлы баз данных (Paradox и DBase), через текстовые файлы баз данных, и с помощью так называемого vet-файла (внутренний формат табличных данных системы ViDa).

Загрузив таблицу данных, пользователь на основе объекта ТАБЛИЦА, выбирая необходимые числовые поля и указывая способ их нормировки, создает объект ДАННЫЕ, который содержит числовой массив всех значений выбранных признаков. В дальнейшем объект ДАННЫЕ сохраняет связь с объектом ТАБЛИЦА, на основе которого он был создан. Используя один и тот же объект ТАБЛИЦА, можно создавать различные объекты типа ДАННЫЕ, выбирая разные наборы строк, признаков и способы их нормировки.

На основе объекта ДАННЫЕ пользователь создает различным образом настроенные объекты типа КАРТА, которые в дальнейшем хранятся в нем как в объекте-контейнере. Объекты типа КАРТА содержат всю необходимую информацию о положении узлов сетки в пространстве и о способе ее доопределения до многообразия.

Для визуализации данных на основе объекта КАРТА создается набор объектов типа СЛОЙ. Объект типа слой содержит всю необходимую информацию для отрисовки на экране информационного слоя.

В программе ViDa Expert 1.0 встроено 4 вида слоев, это: Слой точек данных, Слой сетки, Слой раскрасок, Слой объектов.

Каждый СЛОЙ имеет характеристику «Вид». В программе ViDa Expert 1.0 реализовано 4 варианта Видов: это

1) вид на координатные плоскости;
2) вид на плоскость главных компонент;
3) вид во внутренних координатах карты (простая развертка карты);
4) вид во внутренних координатах карты (нелинейная развертка карты).

В 3-ем варианте карта предстает в виде равномерной сетки узлов, точки данных размещены в соответствии с их проекциями на карту. В 4-ом – виде карта изображается в виде криволинейной сетки. Отдельный диалог позволяет настраивать криволинейную развертку несколькими способами: а) так, чтобы расстояние между узлами сетки на плоскости как можно точнее соответствовали расстоянию между узлами в исходном пространстве; б) так, чтобы криволинейная развертка соответствовала определенной раскраске; например, чтобы координатная сетка была более «густой» в светлых областях раскраски и более «разреженной» в темных областях.

Карта может быть различным образом раскрашена. За вариант раскраски отвечает свойство «Тип раскраски» Слоя раскрасок. В программе ViDa Expert 1.0 реализовано 9 вариантов Раскрасок:

1) раскраска по значению выбранного признака;
2) раскраска по двумерной плотности;

3) раскраска по двумерной плотности выделенного подмножества;

4) раскраска по относительной двумерной плотности выделенного подмножества;

5) раскраска по многомерной плотности;

6) раскраска по многомерной плотности выделенного подмножества;

7) раскраска по относительной многомерной плотности выделенного подмножества;

9) раскраска по расстоянию от точки карты до ближайшей точки данных;

3.1.2. Различные варианты работы с программой ViDa Expert

Основные идеологические принципы работы в программе ViDa Expert – это

а) принцип «красной кнопки»;

б) принцип «торчащих хвостиков смысла».

Что это означает?

Принцип «красной кнопки» состоит в том, чтобы пользователь нажимал минимальное количество кнопок (клавиш) для получения «рядового результата», то есть результата, не учитывающего конкретные особенности задачи, но служащего началом для дальнейшей работы. Согласно этому принципу пользователь должен иметь возможность получить результат, не владея всеми тонкостями методов настройки. Иными словами, все уже должно быть максимально настроено для получения результата, который бы как-то удовлетворил пользователя, не желающего или не имеющего времени вникать в детали работы программы. Желательно еще, чтобы этот результат был в некотором отношении «неплохим» (пусть не оптимальным).

Принцип «хвостиков смысла» заключается в том, чтобы, тем не менее, пользователь видел, что за «красной кнопкой» стоит целый массив разнообразных настроек, «рычажков», с которыми он может экспериментировать и видоизменять результат. Пользователь должен ощущать, что он совершает не бездумные действия и при желании может добиться оптимального результата.

3.1.3. Некоторые типовые задачи

Опишем типовые задачи, которые пользователь может решать с помощью программы ViDa Expert.

Создание задачника и линейный анализ данных

1. Открыть новый проект (пункт *Новый* в меню *Проект*).
2. Добавить новый сценарий (пункт *Добавить* в меню *Сценарий*)
3. Добавить новую таблицу (пункт *Добавить таблицу* в меню *Сценарий*)
4. Отметить в таблице поля-признаки для анализа и выбрать способ нормировки.
   4.1. При желании пользователь может с помощью диалога «Выбор объектов из таблицы» выбрать отдельные строки-объекты из таблицы и указать цвета раскраски, с помощью которых будет задаваться изначальное деление точек на классы.
5. Добавить новый набор данных (пункт *Добавить задачник* в меню *Данные*)
6. С помощью диалога «Линейная статистика» провести простейший анализ данных.
7. При необходимости сохранить набор данных (пункт *Сохранить данные* в меню *Данные*). В файле *ved* сохраняются выбранные поля и варианты нормировки. В дальнейшем файл *ved* может быть открыт в самом начале работы программы (Пункт *Загрузить данные* в меню *Проект*). Таблица автоматически сохраняется в одноименном файле с расширением *vet*.

Автоматическое создание карты и простая визуализация

1. Загрузить данные или создать задачник.
2. Создать карту с помощью кнопки “*See It!*”.
3. Далее данные и карту можно рассматривать в разных пространствах:
   3.1. Трехмерные линейные подпространства, натянутые на отдельные координатные оси (Вид *На плоскость координат*).
   3.2. Трехмерные линейные подпространства, натянутые на главные компоненты (Вид *На главные компоненты*).
   3.3. Двумерное пространство карты (если карта двумерная, вид *Во внутренних координатах*).
   3.4. Трехмерное пространство карты (если карта трехмерная, вид *Во внутренних координатах*).
4. Карту можно раскрашивать, выбирая раскраски в окне «Раскраска».
5. Точки во всех видах автоматически снабжаются всплывающей аннотацией. Текст аннотации задается полем таблицы, задаваемом в списке «Поле-метка» в панели *Объекты*.

Кластерный анализ с визуальным контролем

1. Загрузить данные или создать задачник.
2. Создать карту с помощью кнопки “*See It!”*.
3. В диалоге “Анализ Данных” провести кластерный анализ, результаты которого синхронно отображаются графически и в таблице. Результат классификации можно запомнить в таблице с помощью кнопки «*Номера в таблицу*».

Аннотирование точек данных

1. Загрузить данные или создать задачник.
2. Создать карту с помощью кнопки “*See It!”*.
3. В диалоге “Аннотирование данных” произвести аннотирование данных.

3.2. Применение методов визуализации данных к картографированию экономических таблиц

В качестве примера применения технологии визуализации данных нами была предпринята попытка применить методы визуализации произвольных данных к картографированию таблицы крупнейших российских предприятий, взятой из журнала «Эксперт-200» [50]. Файлы исходных данных были получены с официального сайта журнала http://www.expert.com.

Исходная таблица содержала информацию об экономическом положении двухста крупнейших российских предприятий, ранжированную в поряке убывания валового объема производства продукции. Изначально таблица содержала следующие поля-признаки (часть из них является независимыми признаками, часть рассчитывается по явным формулам):

1) Название предприятия;
2) Регион местонахождения предприятия;
3) Отрасль, к которой относится предприятие;
4) Валовый объем производства продукции в 1998 году;
5) Валовый объем производства продукции в 1997 году;
6) Темпы роста предприятия
7) Валовый объем производства в 1998 году, выраженный в долларовом эквиваленте;
8) Балансовая прибыль предприятия;
9) Прибыль предприятия после налогообложения;
10) Прибыльность предприятия;
11) Число работающих на предприятии;
12) Производительность труда.

Шумским С.А. [49] уже была предпринята попытка визуализации таблицы предприятий, взятой из журнала «Эксперт» за 1997 год. В этой работе были использованы традиционные самоорганизующиеся карты Кохонена и диаграммы Хинтона. Там же было предложено использовать в качестве координат пространства данных отношения некоторых независимых признаков из таблицы. Было предложено четыре таких координаты.

Нами было решено расширить пространство исходных данных еще одним измерением, в результате чего был получен следующий набор независимых признаков:

N	Обозначение признака	Значение
1	LG_VO1998	Логарифм валового объема производства продукции в 1998 году
2	TEMP	Валовый объем производства продукции в 1998 году / Валовый объем производства продукции в 1997 году
3	PROFIT_BAL	Балансовая прибыль предприятия / Валовый объем производства продукции в 1998 году
4	PROFIT_NAL	Прибыль предприятия после налогообложения / Валовый объем производства продукции в 1998 году
5	PRODUCTIV	Прибыль предприятия после налогообложения / Число работающих на предприятии

В результате была составлена таблица из двухсот записей с пятью полями. Часть записей содержала неполную информацию (по отдельным признакам информация отсутствовала).

Данные были предварительно нормированы по формуле $\tilde{x}_i = th\left(\frac{x_i - M}{\sqrt{D}}\right)$, где $\tilde{x}_i, x_i$, M, D – новое, старое значения признака, среднее значение и дисперсия признака соответственно.

Карта, с помощью которой осуществлялась визуализация множества данных, была построена по алгоритму построения упругих карт. Первоначальная сетка содержала 10 узлов по вертикали и 10 по горизонтали. Для нахождения локального минимума функционала применялся метод отжига. Параметры μ и λ медленно (так чтобы при каждом изменении карта успевала перейти в близлежащий локальный минимум) менялись от значений $\mu = 5$, $\lambda = 5$ до $\mu = 0.1$, $\lambda = 0.1$.

После построения упругой карты данные из пространства признаков были спроецированы на карту с помощью процедуры нахождения ближайшей точки карты в случае кусочно-линейной интерполяции между узлами.

В качестве иллюстрации анализа экономических данных ниже приведены раскраски полученной карты по координатным полям, а также слой рассчитанной плотности данных в точках карты. На раскрасках большими точками с номерами выделена группа предприятий, принадлежащих нефтегазовой промышленности. Такое выделение позволяет проанализировать место той или иной отрасли промышленности среди других предприятий.

1) Раскраска по признакам

На рисунке 38а изображено значение признака LG_VO1998 в точках карты. При этом более светлым участкам соответствуют более высокие показатели признака. Самый яркий цвет соответствует первым 10% предприятий с самым большим валовым объемом производства. Для примера кружками с цифрами выделены предприятия нефтегазовой промышленности. Цифрам соответствуют следующие названия предприятий:

1 – ОАО «Газпром»; 2 – Нефтяная компания «ЛУКойл»; 3 – Башкирская топливная компания; 4 – Нефтяная компания «Сургутнефтегаз»; 5 – Тюменская нефтяная компания; 6 – «Татнефть»; 7 – Нефтяная компания «Славнефть»; 8 – Нефтяная компания «Роснефть»; 9 – Оренбургская нефтяная компания «Онако»; 10 - Центральная топливная компания; 11 – Нефтяная компания «КомиТЭК».

Рисунок 38б изображает раскраску по показателю TEMP. Как видно из рисунка 38б, область крупнейших предприятий не пересекается с областью наиболее высоких темпов роста. В правом нижнем углу карты, например, располагаются предприятия пищевой промышленности, цветной металлургии и другие быстро развивающиеся отрасли.

На рисунках 38в, 38г, 38д показаны раскраски по признакам PROFIT_BAL, PROFIT_NAL, PRODUCTIV. Эти раскраски схожи, что указывает на корреляцию последних трех признаков. Вместе с этим различия в раскраске позволяют выделить предприятия, которые выпадают из корреляционной зависимости.

2) Раскраска по плотности данных

На рисунках 38е),38ж),38з) показана раскраска карты по плотности данных, оцененной с помощью какой-либо непараметрической оценки. Существует два способа оценить плотность данных. Во-первых, можно рассматривать двумерное распределение точек на карте. Во-вторых, можно рассчитать плотность точек в исходном n-мерном пространстве, и изображать на карте значения этой плотности в точках расположения карты. На рисунках изображено применение первого способа. Более темным участкам соответствуют более высокие значения плотности.

Рисунок 38е) изображает двумерное распределение общей плотности данных. На рисунке 38ж) – распределение плотности предприятий

нефтегазовой промышленности. Рисунок 38з) отражает удобную для оценок относительную плотность предприятий нефтегазовой промышленности (то есть отношение первых двух плотностей).

На рисунке 38и) отражено расстояние от каждой из точек карты до ближайшей точки данных. Более темным участкам соответствуют большие расстояния. Видно, что в целом данные достаточно плотно прилегают к карте, за исключением участка в левом верхнем углу (впрочем, точки данных там отсутствуют и темный цвет указывает на то, что точки в левом верхнем углу карты расположены в многомерном пространстве достаточно далеко от основного массива данных).
Беглый взгляд на рисунки позволяет сделать, например, такие выводы. Предприятия нефтегазовой промышленности являются лидерами по объему валового производства, но темпы роста этой области промышленности невелики по сравнению, например, с пищевой промышленностью. Предприятия нефтегазовой промышленности распадаются на две группы, которые существенно отличаются по прибыльности производства. В целом, набор таких рисунков могут служить удобным средством анализа для специалистов в макроэкономике.

3.3. Нейроинформатика – наука или фантастика?

Попробуем ответить на этот вопрос с помощью приема картографирования текстовых коллекций на основе представления текстов в виде *частотных словарей*.

По 50-ти текстам, представляющих собой научные статьи, доклады конференций и книги был составлен словарь из 800 наиболее употребляемых слов. Аналогичный словарь объемом 700 слов был составлен для коллекции фантастических произведений различных авторов и жанров. Поскольку в русском языке одно и то же слово может быть представлено в нескольких формах, то для идентификации корня брались лишь первые значащие буквы слова. В словари не включались слова-связки и незначащие слова (местоимения, общеупотребительные слова и др.).

Далее оба словаря были объединены. Поскольку словари оказались частично перекрывающимися, в результирующем словаре оказалось 1375 слов. Слова были пронумерованы в алфавитном порядке.

Для того, чтобы представить текст в виде многомерного вектора ему сопоставлялся набор частот $w_i = n_i / N$, i = 1…1375, где n_i – число встретившихся форм i-ого слова из словаря, N – общее число слов в тексте. В результате была получена таблица из 1376 столбцов (первый из них содержал название текста, остальные – частоты), в которую были занесены частотные словари 113 текстов, в которые вошли фантастические произведения С.Лема, Р.Желязны, С.Кинга, А.Азимова, А.Кларка, Р.Брэдбери, К.Булычева и др., книга А.Н.Горбаня «Демон

Дарвина», книга Е.М.Миркеса «Нейрокомпьютер. Проект стандарта», некоторые научные статьи красноярской группы исследователей «Нейрокомп», тезисы докладов, представленных на конференции «Нейроинформатика и ее приложения - 2000», проводившиейся в г.Красноярске в октябре 2000 года, главы этой книги и некоторые другие тексты.

На рис.39а) приведен вид полученного многомерного облака точек на плоскость первых двух главных компонент. Как видно, частотные словари фантастических и научных текстов хорошо разделяются вдоль первой главной компоненты. Рассмотрим те признаки, которые оказались наиболее значимыми для такого разделения. Эти признаки-словоформы вошли в вектор первой главной компоненты с наибольшими по абсолютной величине весами. Первый десяток таких словоформ показан в табл.1. Наоборот, те признаки, которые имеют близкие к нулю веса, оказались незначимыми для разделения.

Для более подробного анализа текстов по набору данных была построена карта. Данные спроецированные на карту, показаны на рисунке 39б). Видно, что тексты в той и другой группе распадаются на подгруппы. Анализ названий текстов, входящих в подгруппы позволяет произвести классификацию текстов следующим образом:

1. Фантастика гуманитарной направленности;
2. Фантастика технической направленности;
3. Биологические и медицинские приложения нейросетей;
4,5. Технические приемы по созданию нейросетей;
6. Моделирование данных (в эту подгруппу входит и эта книга – объекты «Глава 1», «Глава 2», «Глава 3»).

Эта классификация не охватывает таких выделяющихся текстов, какими являются книги А.Н. Горбаня «Демон Дарвина», С.Лема «Сумма технологий» и некоторых других.

На примере картографирования коллекций текстов рассмотрим возможности автоматического *аннотирования* точек данных. На рис.39б) показан самый простой способ аннотирования данных – на точки повешены названия тех признаков, значения которых оказались для данного объекта максимальными. В нашем случае это означает, что соответствующие словоформы оказались в тексте наиболее часто встречающимися.

Другой способ аннотирования состоит в выделении для данного объекта тех признаков, значения которых наименее вероятны по ансамблю объектов. Поясним, что это означает на нашем примере. Для каждого признака-слова может быть построена гистограмма распределения значений по всем объектам. В некоторых интервалах гистограммы окажется большое количество объектов, в некоторых – малое. Выберем объект (или точку пространства признаков) для аннотирова-

ния. Каждому из конкретных значений признаков у выбранного объекта можно сопоставить вероятность его появления (взяв ее из построенной по всем объектам гистограммы). Если вероятность окажется высока, то это значение признака является в некотором роде «типичным» для данной системы объектов. Если вероятность мала – значит такие значения признак принимает лишь на небольшом количестве объектов, то есть выбранный объект по данному признаку является «нетипичным».

Табл.1. Слова-признаки, самые значимые и самые малозначимые для разделения текстов на научные (нейроинформатика) и фантастические

Слова, обладающие отрицательными весами		Слова, обладающие положительными весами		Слова с близкими к нулю весами	
Слово	Вес	слово	вес	слово	вес
Верн	-0.09101	использ	0.076008	идеал	0.00129
чувств	-0.08864	основ	0.073729	прост	0.000917
добр	-0.08147	задач	0.068003	резк	0.000833
земл	-0.07893	данн	0.065691	парадокс	0.000689
люд	-0.07666	нейро	0.064884	высказ	0.000601
странн	-0.07657	определ	0.063517	кислот	0.000266
холод	-0.0736	функц	0.057877	ремонт	0.00017
трудн	-0.07148	анали	0.057309	констр	0.000062
страх	-0.06715	модел	0.056592	окрестност	-0.00004
осторожн	-0.0666	параметр	0.055749	вопрос	-0.00019
сомнен	-0.0665	сет	0.053742	груб	-0.00043
мысл	-0.06584	резуль	0.053426	скобк	-0.00052
чуж	-0.06563	обуч	0.052626	сформулир	-0.00079
зло	-0.06536	алгоритм	0.052238	мишен	-0.00083
ужас	-0.0644	решен	0.051107	вражд	-0.00102
тревог	-0.06423	выбор	0.050932	свидет	-0.00112

В случае текстовых коллекций маловероятный признак – это то слово, которое отличает данный текст от остальных. Так, например, слово «данные» может стабильно часто употребляться во всех текстах по нейроинформатике. Поэтому, информация о том, что в тексте часто употребляются слова «анализ», «данные» никак не выделяет его среди остальных текстов. Информация же о том, что слово «лимфоцит» оказалось маловероятным в данном тексте для данного ансамбля текстов, определенным образом его характеризует. Наоборот, если маловероятным оказалось слово «данные», то и это означает, что текст «выпадает» из общей направленности собрания текстов по нейроинформатике.

В случае если коллекция текстов исходно разнородна, как в нашем примере – тогда тоже имеет смысл выделять маловероятные признаки, но вероятности лучше рассчитывать по объектам того класса, в который входит объект.

На рис.40а) представлен способ аннотирования текстов по маловероятным словам.

Источник иной информации о исследуемой системе текстов – визуализация транспонированной задачи, описанная в разделе 2.8 (см. рис.40б). В этом случае роль объектов играют слова из частотного словаря, а признаков – тексты. Если на плоскости первых главных компонент два слова оказались рядом, то это позволяет предположить, что на данной совокупности текстов они коррелируют, то есть примерно одинаково часто встречаются в одних текстах и одинаково редко – в других. Если слова оказались сильно разнесены (на противоположных «полюсах» карты) это указывает на обратную корреляцию – если одно слово встречается часто в каком-либо тексте, то другое в этом тексте встречается, скорее всего, редко.

Аннотирование по маловероятным признакам применимо и здесь. В данном случае маловероятный признак – текст – выделяется для аннотируемого слова. Это означает, что среди выбранных текстов маловероятно встретить заданное слово именно в указанном тексте.

Может показаться, что поставленный в заголовке вопрос не имеет большого «научного» и практического значения. Однако, на примере картографирования текстовых коллекций демонстрируются методы, применимые для задач более серьезных наук. Пример использования частотных словарей – *анализ генетических последовательностей*. В этом случае текст – последовательность «букв» генетического алфавита A, C, G, T. Последовательные буквы можно объединять в слова разной длины. Таким образом, генетический текст можно представлять в виде частотного словаря. Генетическому коду каждой отдельной особи сопоставляется свой частотный словарь. Сравнение и визуализация генетических частотных словарей внутри выборки особей одного вида, а также сравнение генетических частотных словарей особей разных видов может служить источником ценной информации для специалистов в этой области.

Приведем пример. По 1800 реальным генетическим последовательностям бактерий, принадлежащих семействам Proteobacteria, Firmicutes, Acidobacterium, Aerobic bacillus, Cyanobacteria была составлена таблица частот встречающихся в тексте слов длины 1, 2 и 3 (синглеты, дуплеты и триплеты). На рис.41а) показан вариант картографирования таблицы. На карте точками разной формы выделены три отдельных рода бактерий (Proteobacteria a-sd, Proteobacteria b-sd, Firmicutes Actinomycetes). Видно, что биологическая классификация в многомерном пространстве частотных словарей задает достаточно компактные и отделенные друг от друга группировки объектов.

3.4. Визуализируем выборы

Аннотирование полезно не только для частотных словарей. На рис.42а) показана карта выборов американских президентов, построенная на основе известной таблицы [12,56,68]. В таблице содержатся ответы на 12 вопросов, расшифровка обозначений которых приведена на рисунке. На карте явно выделяются выборы 1880 года, попадающие в область, где преобладают объекты противоположного класса. В своем классе (победа правящей партии) для выборов 1880 года оказались крайне маловероятны значения признаков CONC (*Была серьезная конкуренция при выдвижении от правящей партии?*) и PREZ (*Кандидат от правящей партии был президентом в год выборов?*). Действительно, среди всех побед правящей партии признак CONC был равен единице только для выборов 1880 года. С другой стороны, этот признак оказывается весьма значимым для решения вопроса о победе на выборах.

Чтобы подтвердить последнее утверждение, рассмотрим транспонированную задачу (рис. 42б). В группу визуализируемых признаков включено поле «Ответ», в котором содержится результат выборов. Видно, что признак, наиболее связанный с результатом выборов – CONC. С другой стороны, признаком, наиболее удаленным от «Ответ» является признак PREZ, что может указывать на обратную корреляцию. Признаки DEPR, O_HERO, MIST, THIRD, CHANGES, WAVE образуют группу взаимосвязанных признаков.

3.5. Осложнения инфаркта-миокарда[1]

Инфаркт миокарда – распространенное и грозное заболевание. Бурное распространение этого заболевания за последние полвека сделало его одной из наиболее острых проблем современной медицины.

Заболеваемость инфарктом миокарда остается высокой во всех странах. Особенно это касается городского населения высокоразвитых стран, испытывающего стремительный ритм современной жизни и подвергающегося хроническому воздействию стрессовых факторов, нерегулярного и не всегда сбалансированного питания. В США ежегодно около 1,5 миллионов человек заболевают инфарктом миокарда.

Несмотря на то, что внедрение современных лечебно-профилактических мероприятий несколько снизило смертность от инфарктов, она продолжает оставаться довольно высокой. Около 15-20% больных острым инфарктом миокарда погибают на догоспитальном этапе, еще 15% в больнице, т.е. общая летальность при остром инфаркте миокарда 30-35%.

Течение заболевания у пациентов с инфарктом миокарда различно. Инфаркт миокарда может протекать без осложнений или с осложнениями не ухудшающими долгосрочный прогноз. В тоже время око-

[1] Описание ситуации с осложнениями инфаркта миокарда взято из [40,54,55].

ло половины пациентов в острый и подострый периоды имеют осложнения, приводящие к ухудшению течения заболевания и даже летальному исходу. Предвидеть развитие этих осложнений может не всегда даже опытный специалист.

Для решения задачи прогнозирования осложнений инфаркта миокарда с целью своевременного проведения необходимых профилактических мероприятий сотрудниками кафедры внутренних болезней № 1 Красноярской государственной медицинской академии была собрана информация о течении заболевания у 1700 больных инфарктом миокарда, проходивших лечение в 1989-1995 годах в Кардиологическом центре городской больницы № 20 г.Красноярска. Информация получена из историй болезни пациентов и сконцентрирована в 128 полях электронной таблицы. База данных содержит сведения о данных анамнеза каждого больного, клинике настоящего инфаркта миокарда, электрокардиографических, лабораторных показателях, лекарственной терапии и особенностях течения заболевания в первые дни инфаркта миокарда.

В результате получилась большая таблица, анализ информации в которой имеет, с одной стороны, практическое значение, с другой – данные в ней имеют весьма сложную структуру. По утверждениям некоторых специалистов в области нейроинформатики: «Красноярская таблица по осложнениям инфаркта миокарда содержит почти все известные сложности, с которыми может столкнуться исследователь при анализе реальных данных. Любой метод анализа, претендующий на практическое применение, должен быть апробирован и на этой таблице.».

На рис.42 приведен пример визуального представления таблицы по осложнениям инфаркта миокарда. На рис. ??? а) приведена оценка плотности распределения объектов. Поскольку в таблице содержится большое количество признаков, необходимо провести предварительный анализ признаков на значимость для того, чтобы предоставить пользователю наиболее значимые признаки. Был проведен самый простой анализ признаков на значимость с помощью первой главной компоненты, и на рис.42б)-42д) показаны раскраски по нескольким признакам, которые имеют наибольшие по абсолютной величине веса в векторе главных компонент. На рис.42е) большими треугольниками обозначены летальные исходы заболевания.

Подробный анализ таблицы по осложнениям инфаркта миокарда требует отдельного и весьма обширного исследования. Информационные раскраски и применение методов визуализации могут играть в этом исследовании вспомогательную роль иллюстративного материала к другим методам, а также имеют самостоятельную ценность.

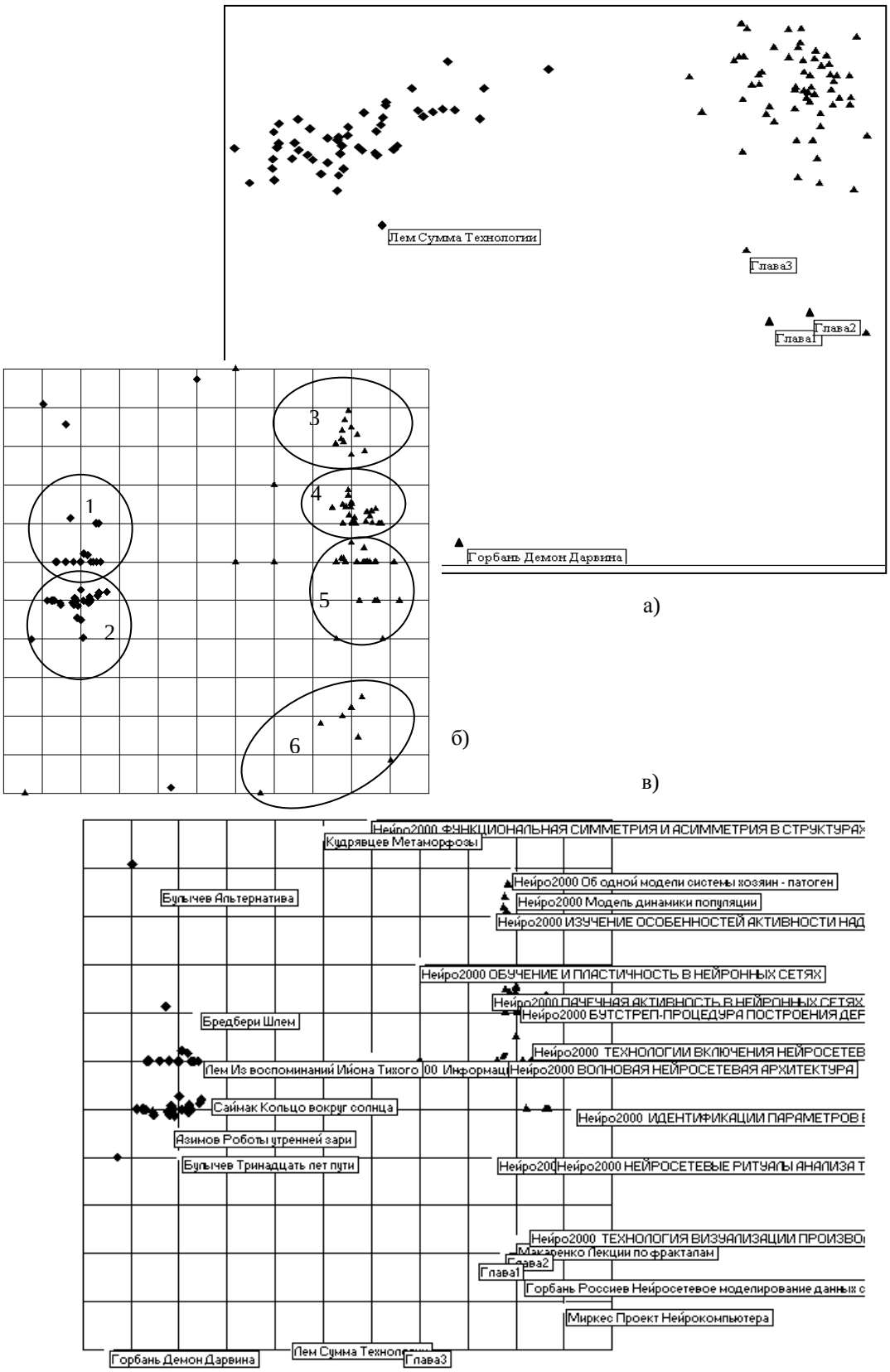


Рис. 39. Визуализация собрания из 116 текстов. Ромбами обозначены фантастические тексты, треугольниками – «научные».

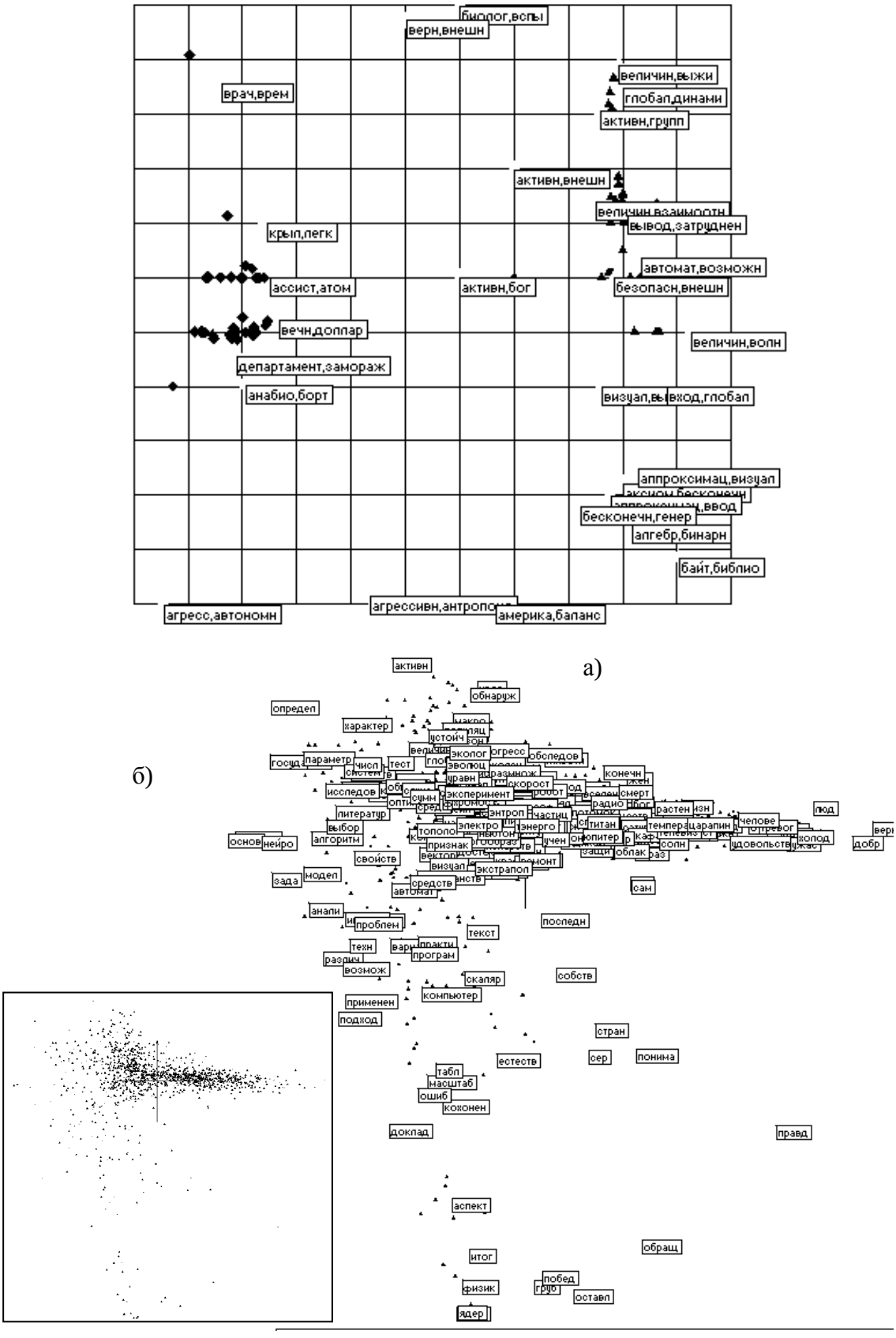
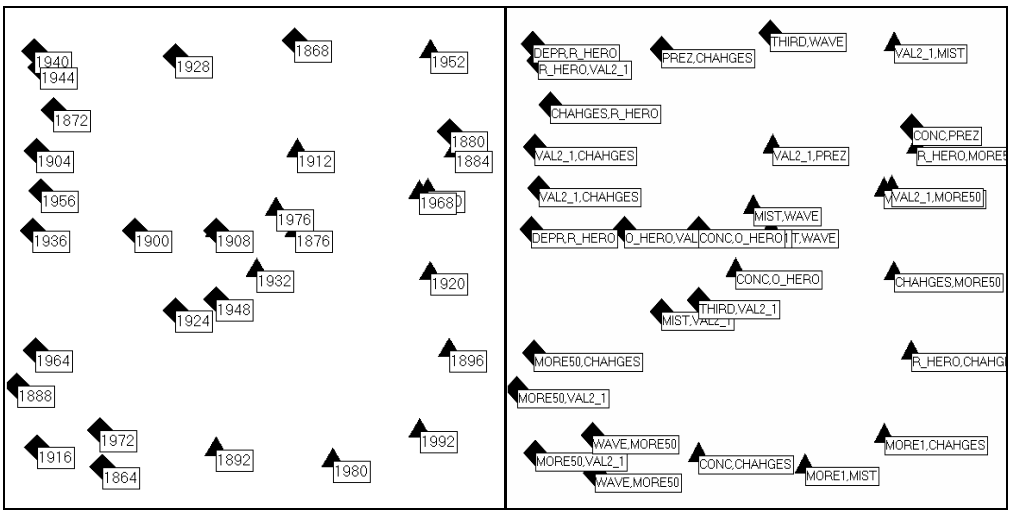
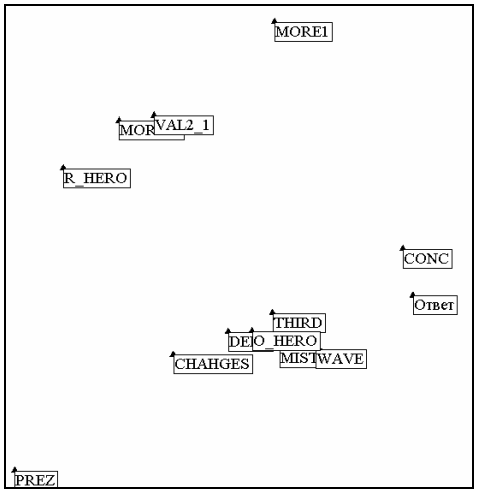


Рис. 40. а) аннотирование текстов по маловероятным словам;
б) визуализация транспонированной задачи.



а) б)



в)

Рис. 41. Карта выборов американских президентов. Ромбами отмечены выборы, на которых победу одержала правящая партия, треугольниками – победы оппозиции.

а) аннотация по году выборов;

б) аннотация по двум самым маловероятным признакам в своем классе;

в) картографирование транспонированной задачи (близкие признаки – коррелированы);

Расшифровка названий признаков:

MORE1	Правящая партия была у власти более одного года?
MORE50	Правящая партия получила больше 50% на прошлых выборах?
THIRD	В год выборов была активна третья партия?
CONC	Была серьезная конкуренция при выдвижении от правящей партии?
PREZ	Кандидат от правящей партии был президентом в год выборов?)
DEPR	Был ли год выборов временем спада или депрессии?
VAL2_1	Был ли рост среднего национального валового продукта на душу населения >2,1%
CHANGES	Произвел ли правящий президент существенные изменения в политике?
WAVE	Во время правления были существенные социальные волнения?
MIST	Администрация правящей партии виновна в серьезной ошибке или скандале?
R_HERO	Кандидат правящей партии – национальный герой?
O_HERO	Кандидат оппозиционной партии – национальный герой?

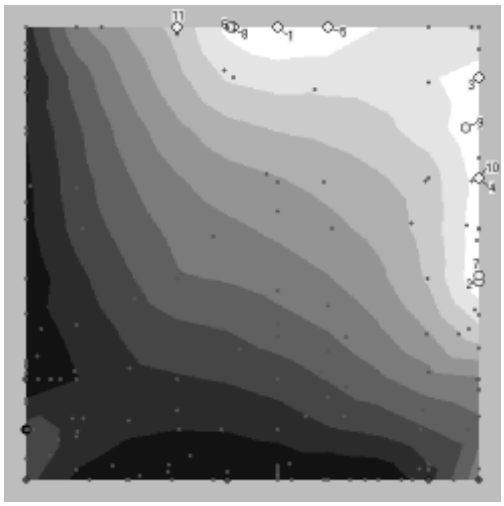


Рис. 38а

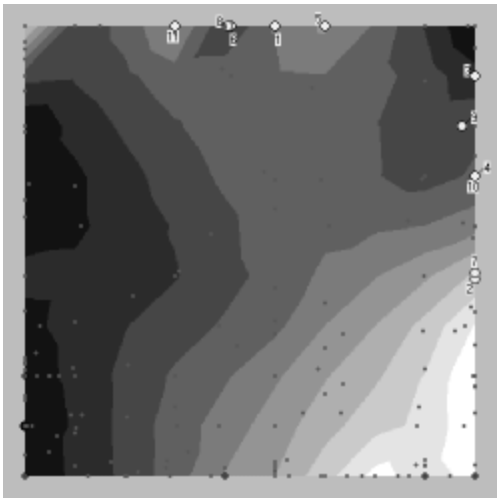


Рис. 38б

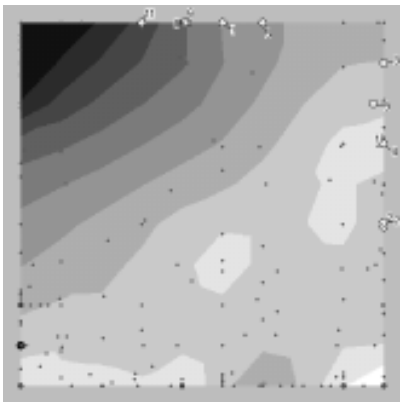


Рис. 38в

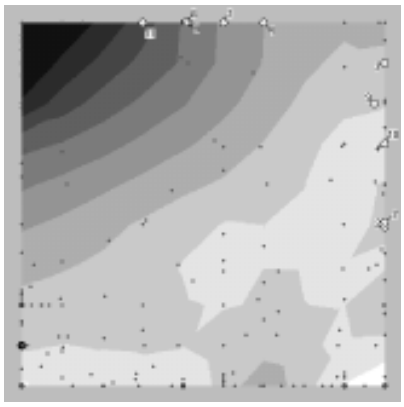


Рис. 38г

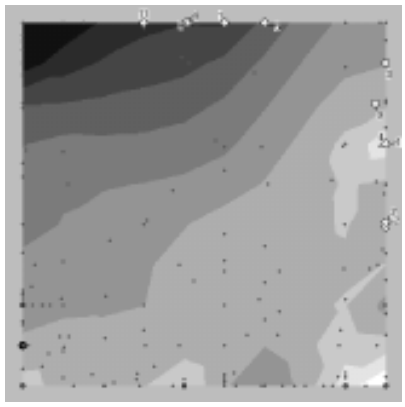


Рис. 38д

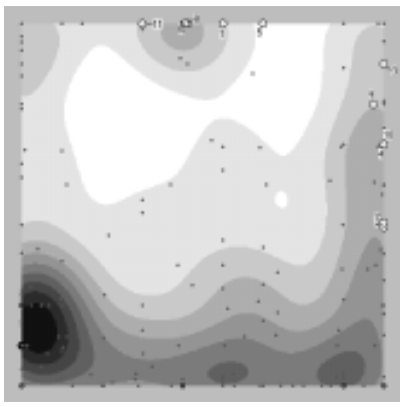


Рис. 38е

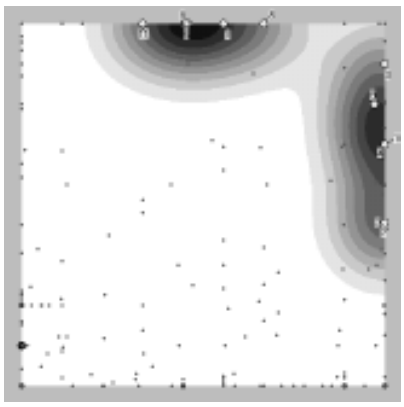


Рис. 38ж

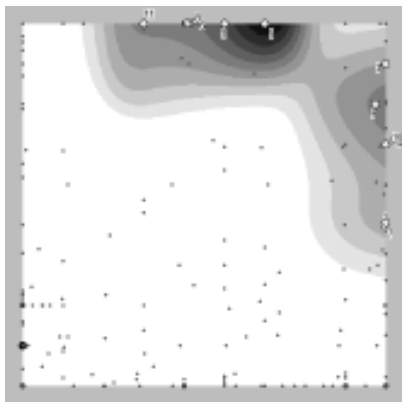


Рис. 38з

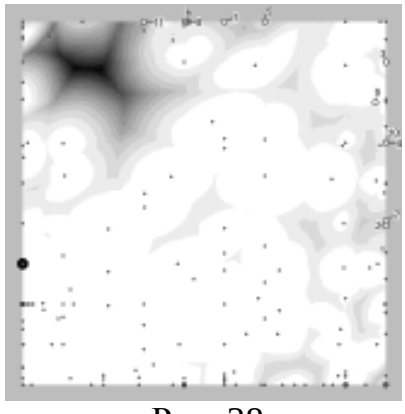


Рис. 38и