ВЫСОЦКАЯ ГАЛИНА СТЕПАНОВНА

НЕПАРАМЕТРИЧЕСКИЕ СИСТЕМЫ КЛАССИФИКАЦИИ В ЗАДАЧАХ ИССЛЕДОВАНИЯ МЕДИКО-БИОЛОГИЧЕСКИХ ПРОЦЕССОВ

05.13.14 - Системы обработки информации и управления

А в т о р е ф е р а т
диссертации на соискание ученой степени
кандидата технических наук

Красноярск - 1998

Общая характеристика работы

Актуальность темы. Разработка эффективных систем обработки информации, на основе создания мощных банков данных, и возможность более свободного обмена информацией через сети приводит к многократному увеличению ее объемов. В этих условиях возрастает потребность создания математических средств и разработки программ для структуризации и анализа больших массивов данных с целью обнаружения скрытых закономерностей и представления их в удобном для человека виде. Несмотря на это большое количество информации все-таки остается не востребованной.

Когда пользователь оперирует большой, постоянно увеличивающейся числовой информацией, важным средством исследования систем в условиях исходной неопределенности становятся методы классификации и распознавания образов. Они позволяют создать представление о структуре этих данных, дифференцируя и объединяя их в классы. Если в результате решения задачи классификации получены компактные группы, однородные по характерным признакам, то в дальнейшем анализе мы можем использовать такие группы, как структурные единицы.

Классификация данных обеспечивает обход проблемы сложности и априорной неопределенности при моделировании систем позволяет получить более точные оценки и распространить полученные результаты на множества объектов. Актуальной для классификации и распознавания образов, особенно в области медицины, является также задача о сокращении объема и размерности обучающей выборки.

В настоящее время с различных теоретических позиций разработано большое число способов решения задачи классификации и распознавания образов. Установлено, что трудоемкость сложных методов классификации пропорциональна квадрату объема выборки N, и в лучшем случае имеет порядок O(N ln N). Требуемый объем памяти зачастую также пропорционален квадрату объема выборки.

Известны примеры программных реализаций методов классификации и распознавания в таких коммерческих пакетах, как ER DAS, STATGRAF, STATISTICA, IDRISI и др. Но, как правило, в этих пакетах используется метод k - ближайших соседей и аналогичные методы, основным достоинством, которых является относительно низкая трудоемкость.

Из российских разработок наиболее популярны пакеты ОТЭКС и КВАЗАР. В то же время опыт работы в области классификации и распознавания образов показывает необходимость

создания быстродействующих алгоритмов классификации, хорошо работающих не только в ситуации хорошо разделимых классов, но и тогда, когда границы между соответствующими классами "размыты".

Работа выполнялась в рамках научной темы Института Вычислительного моделирования СО РАН "Создание теории многоуровневых непараметрических систем принятия решений" (1.13.5.3), грантов РФФИ N93 - 012 - 0486, N97 - 01 - 01043.

Цель работы состояла в разработке и исследовании непараметрической системы классификации статистических данных в условиях больших выборок и ее применении при анализе медико - экологических процессов.

Цель достигается путем решения следующих задач:

- Разработка и исследование быстродействующих непараметрических алгоритмов и комплекса программ решения задач автоматической классификации и распознавания образов.
- Разработка системы классификации океанических вод по спектральным данным.
- Разработка информационных средств автоматизации исследования и прогнозирования состояний комплекса "сердечно - сосудистая система углеводный обмен".

Методы исследования. Для решения поставленных задач использовались методы теории вероятностей и теории сложных систем, непараметрические алгоритмы автоматической классификации и распознавания образов, средства программирования.

Научная новизна работы состоит в разработке быстродействующего непараметрического алгоритма автоматической классификации, позволяющего исследовать структуру статистических выборок в условиях априорной неопределенности. Это стало возможным при решении задачи автоматической классификации с позиций теории вероятности.

При этом впервые проблема автоматической классификации реализована в рамках задачи распознавания образов с помощью итерационной процедуры последовательного восстановления непараметрической оценки уравнения разделяющей поверхности между классами, соответствующими одномодальным фрагментам плотности вероятности. Количество классов априори не задается. Такой подход позволяет существенно снизить трудоемкость классификации.

Для повышения эффективности алгоритмического обеспечения пакета используется интегральная непараметрическая оценка плотности вероятности, которая по сравнению с классической процедурой Розенблатта - Парзена обладает повышенными аппроксимационными свойствами, что обеспечивается введением дополнительного сглаживающего оператора.

Практическая ценность. Разработанные непараметрические алгоритмы классификации и распознавания образов реализованы в виде диалогового пакета программ "NPCL". Пакет является составной частью программного обеспечения для статистического моделирования сложных развивающихся систем при неполной информации.

Ориентация предложенных моделей и алгоритмов на обнаружение скрытых закономерностей при малом уровне исходной информации придает пакету универсальный характер и позволяет исследовать объекты различной природы.

Разработанный комплекс программ был использован при решении следующих практических задач:

- Построение статистической модели взаимодействия сердечно - сосудистой системы и системы углеводного обмена с целью синтеза критериев диагностики нарушений толерантности к глюкозе по состоянию сердечно - сосудистой системы и данным анамнеза, что позволяет снизить затраты на диагностику и лечение сахарного диабета.
- Моделирование гидробиоценозов поверхностных вод океана при стационарных и нестационарных условиях по обобщенным биооптическим показателям. Разработанные классифика-

ционные модели и программы были использованы при исследовании поверхностных вод по физическим и биологическим параметрам, полученным в 36 рейсе НИС "Ак. Вернадский" в Западной части тропической Атлантики (май - август 1987г.).

- Автоматизация исследований в медицине, экологии, лесном хозяйстве.

Автор защищает:

1. Методику синтеза и быстродействующие непараметрические алгоритмы автоматической классификации больших массивов статистических данных.
2. Диалоговый комплекс программ **NPCL**, обеспечивающий решение задач автоматической классификации, распознавания образов, минимизации описания и визуализации результатов обработки информации.
3. Статистическую модель взаимодействия параметров сердечно - сосудистой системы и системы углеводного обмена.
4. Статистическую модель взаимосвязи между обобщенными биооптическими показателями поверхностных вод океана при стационарных и нестационарных условиях.

Реализация результатов работы. В результате исследования создан диалоговый пакет программ **NPCL**, на основе которого разработаны системы медицинской диагностики, внедренные в Институте медицинских проблем Севера СО РАМН. Информационная система классификации и анализа спектральных данных используется в Институте биофизики СО РАН при автоматизации научных исследований биоценозов океанических вод.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на международных, всесоюзных и всероссийских конференциях: симпозиум "Машинные методы обнаружения закономерностей" (Минск, 1985), 4-й съезд кардиологов (Москва, 1986), симпозиум "Имитация систем в биологии и медицине" (Прага 1986), Всероссийская научно - практическая конференция "Рискометрия и адаптация в медицине" (Иваново, 1995), Всероссийская конференция "Распознавание образов и анализ изображений. Перспективные информационные технологии" (Ульяновск, 1995), Международный симпозиум "Распространение радиоволн в городе" (Томск 1997), Всероссийская конференция "Проблемы защиты населения и территории в чрезвычайных ситуациях".

Публикации. Результаты проведенных теоретических и экспериментальных исследований опубликованы в 15 печатных работах.

Структура о объем работы. Диссертационная работа состоит из введения, пяти глав, заключения, библиографии (84 наименования), содержит 102 страницы машинописного текста и 16 рисунков.

Автор считает своим долгом выразить глубокую благодарность сотрудникам Института медицинских проблем Севера СО РАМН профессору Поликарпову Л.С., к.м.н. Хамнагадаеву И.И., к.м.н. Шусту Г.М. к.м.н. Пироговскому Н.В. и сотруднику Института биофизики СО РАН д.т.н. Шевырногову А.П. за предоставление данных для обработки и их интерпретацию.

Содержание работы.

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследования, выделены основные положения, имеющие новизну и практическую ценность.

В первой главе изложена формальная постановка задачи автоматической классификации и приведен анализ существующих программных средств и ряда существующих вероятностных алгоритмов классификации.

Дан обзор программных реализаций методов классификации и распознавания в таких коммерческих пакетах, как ER DAS, STATGRAF, STATISTICA, IDRISI, ОТЭКС, КВАЗАР и др.

Несмотря на наличие этих пакетов и разнообразие представленных в них алгоритмов, применение их при исследовании сложных медико - биологических систем затруднено.

Сформулированы основные требования, предъявляемые к алгоритмам классификации медико - биологических данных:

- обнаружение "размытых" классов, количество которых априори не определено;
- устойчивость результатов классификации к изменению объема статистических выборок и параметров алгоритма;
- высокая вычислительная эффективность классификации больших массивов статистических данных;
- затраты памяти пропорциональные объему выборки;
- выделение классов, согласующихся с вероятностной их природой.

Во второй главе изложены теоретические основы непараметрической системы автоматической классификации и рассматривается ее реализация.

Предлагаемая система классификации состоит из подсистем, выполняющих следующие функции:

- предварительная обработка данных;
- вычисление непараметрической оценки плотности вероятности;
- выделение компактных групп точек (классов), соответствующих одномодальным фрагментам плотности вероятности;
- поиск оптимальных параметров непараметрического алгоритма распознавания обнаруженных образов в выбранном подмножестве признаков обучающей выборки;
- организация процедуры распознавания вновь поступающих точек;
- визуализация результатов классификации;
- статистический анализ классов;
- представление результатов классификации, в виде удобном для обработки стандартными пакетами программ.

В результате:

- разработано методическое и алгоритмическое обеспечение непараметрической системы классификации, позволяющее автоматизировать процесс обработки больших массивов данных.
- предложенный непараметрический алгоритм автоматической классификации позволяет обнаруживать классы с несимметричным характером плотности вероятности и снижает требования к оптимальному выбору коэффициента размытости.
- трудоемкость модифицированного алгоритма непараметрической классификации пропорциональна квадрату объема выборки умноженному на коэффициент размытости интегральной оценки плотности вероятности.

Для классификации выборок относительно небольшого объема используется непараметрическая оценка плотности вероятности

$$\overline{p}(x) = \frac{1}{n\prod_{j=1}^{k} 2c_j\beta_j} \sum_{i=1}^{n}\prod_{j=1}^{k}\overline{\Phi}\left(x_j - x_j^i\right), \quad (1)$$

где $\overline{\Phi}(x-u)=\begin{cases} 0, & при \quad |x-u|\ge c+\beta \\ \beta, & при \quad |x-u|<c-\beta \\ (c+\beta-|x-u|, & при \quad c+\beta>|x-u|\ge c-\beta \end{cases}$

Пусть $X=\{\overline{x_i}=(x_{i,1},x_{i,2},\ldots,x_{i,K}), i=\overline{1,n}\}$ -таблица "объект - признак", где n - объем выборки, K - количество признаков.

Для классификации и распознавания образов используется не вся таблица, а некоторое подмножество таблицы X - $X=\{\overline{x_i}=(x_{i,j_1},x_{i,j_2},\ldots,x_{i,j_k}), i=\overline{1,n}\}$, где $j_1<j_2<\ldots<j_k$, k - количество признаков.

Для упрощения дальнейших вычислений, проводится нормировка значений $X_j=x_i^j, i=\overline{1,n}$ для каждого j.

Поскольку ядерная функция в приведенной выше оценке плотности (1) имеет компактный носитель, то подавляющее число слагаемых равны 0. С помощью предварительной сортировки часть таких слагаемых отсеивается, что позволяет существенно снизить количество необходимых вычислений.

Трудоемкость вычислений на этапе вычисления оценки плотности вероятности пропорциональна $n^2(c+\beta)$.

При этом известно, что $c \to 0$, $\beta \to 0$, хотя $nc \to \infty$.

Для данных большого объема предлагается следующая процедура. Область определения разбивается на N непересекающихся гиперкубов с ребром 2β. Пусть P^j - частоты попадания случайной величины x в j -й интервал. Тогда плотность вероятности p(x) можно оценить статистикой

$$\overline{p}(x)=\frac{1}{\prod_{v=1}^{k} c_v}\sum_{i=1}^{T} P^j \prod_{v=1}^{k} c_v \Phi\left(\frac{x_v - z_v^i}{c_v}\right),$$

где z^i, $i=\overline{1,N}$ - центры интервалов. Очевидно, что c = qβ, где q - целое число.

Поскольку при преобразовании непрерывного сигнала происходит разбиение области определения на N непересекающихся элементов, то оценка удобна при обработке спутниковой информации. Данный метод применим к анализу спутниковой информации, когда нас интересуют классы, элементы которых не образуют на поверхности связного множества.

Поскольку для реализации метода необходимо подсчитать P^j - частоты попадания случайной величины x в j -й элемент, воспользуемся методом хэш-кодирования.

Пусть элементы вектора $x^i=(x_1^i,x_2^i,\ldots x_k^i)$ являются номерами интервалов, в которые попадают исходные значения после разделения. Данные в оперативной памяти ЭВМ хранятся в M хэш-таблицах, где M - количество различных значений x_k^i. Это позволяет облегчить поиск точек, лежащих в окрестности точки x^i. Очевидно, что такие точки могут находится только в соседних хэш-таблицах. По сравнению с алгоритмом, использующим одну хэш-таблицу, время просмотра предлагаемой процедурой уменьшается в M/q раз. Кроме того такое использование оперативной памяти связано с ограничениями на максимальный сегмент данных в Delphi.

Выбор алгоритма вычисления плотности вероятности зависит от величин k - количество используемых признаков и q=c/β. Если их значения невелики, то программа осуществляет просмотр окрестности точки $x^i=(x_1^i,x_2^i,\ldots x_k^i)$ в виде гиперкуба объемом q^k. В противном случае,

просматриваются q хэш-таблиц, соседних с таблицей, в которой находится точка x^i. Выбор того либо иного варианта происходит в зависимости от сравнения величин T_1 и T_2. Здесь $T_1=t_1q^k$, t_1 - время вычисления адреса точки x_i в хэш-таблицах, а

$$T_2 = t_2 \sum_{i=1}^{q} H_i$$

где t_2 - время проверки принадлежит ли элемент хэш-таблицы окрестности точки x^i.

H_i - количество непустых элементов i -ой хэш-таблицы. Эти величины подсчитываются в процессе заполнения хэш-таблиц.

Величина t_1 прямо пропорционально зависит от размерности признакового пространства, кроме этого реальное значение этой величины зависит от времени, затрачиваемого на возведение в степень и вычисление модуля. Величина t_2 прямо пропорционально зависит от вероятности попадания проверяемой точки в гиперкуб и от времени, затрачиваемого на сравнение 2-х чисел. Поскольку абсолютные значения t_1 и t_2 существенно зависят от тактовой частоты процессора, естественно вычислять их отношение t_1/t_2. Величины этих соотношений t_1/t_2 при одних и тех же k и q существенно не меняется для одной и той же марки процессора. Для случая равномерного распределения точек выборки и процессоров Pentium величина t_1/t_2 изменяется в пределах от 1.6 до 1.8.

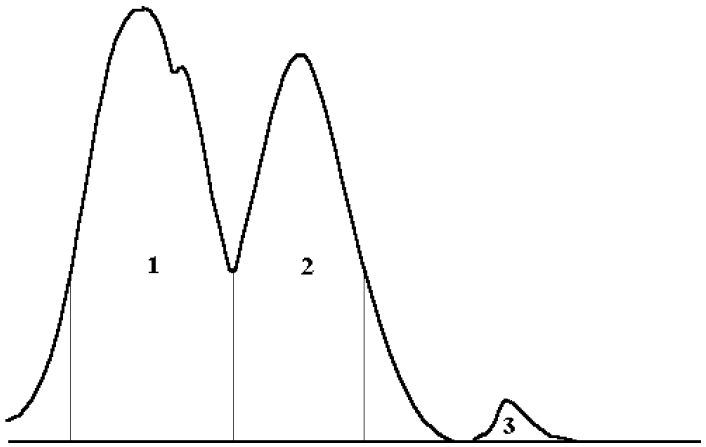

Рис. 1. Иллюстрация выделения классов, соответствующих одномодальным фрагментам плотности вероятности

Следующим этапом реализации непараметрического алгоритма автоматической классификации является выделение групп точек, соответствующих одномодальным фрагментам плотности вероятности. В качестве примера, иллюстрирующего отличие предлагаемого алгоритма от стандартного, рассмотрим плотность вероятности на рис. 1. В результате применения традиционного алгоритма выделяется 4 класса, помеченные цифрами 1, 2, 3, 4. Несмотря на улучшенные аппроксимационные свойства применяемой оценки, класс 2 вполне может быть расценен, как результат дефекта восстановления плотности. При небольшом изменении параметров c и **β** в (1) этот класс исчезает. Класс 4 может быть интерпретирован, как группа, образованная случайными помехами, которые так часто имеют место, когда классифицируемая выборка формируется при аппаратном сборе информации.

Определение 1. Класс C^t считается "значимым по высоте" относительно уровня $h_1>0$, если

$$\max\left\{\left|\overline{p}(x^i)-\overline{p}(x^j)\right|,\forall x^i,x^j\in C^t\right\}\geq h_1.$$

Введем отношение "сходства", между элементами выборки.

Определение 2. Пусть

$$\rho(i,j)=\max_{l=1,k}\left|x_l^i-x_l^j\right|$$

Будем считать, что точки схожи, если $\rho(i,j)\leq c$.

Тогда алгоритм обнаружения одномодальных классов может быть представлен в виде последовательности действий:

1. Провести сортировку элементов выборки со значениями плотностей большими h в порядке убывания плотностей. Обозначим через n' - количество таких элементов. Номера элементов выборки будем хранить в массиве `Addr3`.
2. Пусть i=1, K=0 , где K - количество классов.
3. Определить множество $\overline{a}$ ={x^j: $\overline{p}$ (x$^{\text{Addr3[i]}}$) = $\overline{p}$ (x^j),j=$\overline{1,L}$ }.
4. Сформировать множество точек a' , с которыми близки точки из $\overline{a}$.

```
for l:=1 to L do
if (NumClass[Addr3[l]] = 0) then
begin
    for j:=LowLimits[Addr3[l]] to UpperLimits[Addr3[l]] do
    if (NumClass[Addr3[l]] ≠ 0) (ρ(Addr3[l],Addr3[j]) < c )then
    begin
        {пополняется множество a'}
 end{цикла по j};
end{цикла по l};
```

5. Проверяем точки множества $\overline{a}$, сравнивая их с точками из a'. При этом возможны следующие варианты:

5.1. Точка сходна с точками только из одного класса. Тогда точка помечается, как кандидат на присоединение к этому классу.

5.2. Точка связана с элементами классов $l_1< l_2<... < l_t$. Проводим проверку "значимости по уровню разделения классов".

Пусть t_s и t_q вершины классов s и q (s < q) , если $\overline{p}$ (t_q)- $\overline{p}$ (x^i) < h_2 , то присоединить точки из класса q к множеству $\overline{a}$.

Если классы хорошо разделены, то точка помечается, как кандидат на присоединение к классу с наиболее близкой вершиной.

6. После проверки всех точек из $\overline{a}$, старые элементы множества a' удаляются, помеченные точки присоединяются к классам, исключаются из множества $\overline{a}$ и помещаются в множество a'.

Если остались не расклассифицированные точки, то перейти к п.5.

Если помеченных точек не было, то оставшиеся точки разделяем на m связных подмножеств и объявляем их центрами классов. Увеличить K=K+m.

7. Положить i=i+L. Если i<n' , перейти к 3.
8. Провести распознавание точек выборки со значениями плотностей меньшими h_2 с использованием непараметрического алгоритма распознавания образов.

Предложенный алгоритм отличается тем, что трудоемкость пропорциональна $c_n n^2$.

В результате работы этого алгоритма для примера на рис. 1 выделяется 2 класса, помеченные цифрами I, II.

Предложенный алгоритм менее чувствителен к изменению параметров c и β. Появление или отсутствие класса 4 зависит от выбора параметра h_2.

В свою очередь выбор этого параметра зависит от целей преследуемых при классификации. Если пользователя интересуют большие классы, то рекомендуется принять $h=0.05\,\overline{P}$, где

$$\overline{P} = \max_{i=\overline{1,n}} \overline{p}(x^i).$$

При исследовании, например, небольших объектов на снимках поверхности Земли, порог h уменьшается.

После появления первой точки, принадлежащей границе между классами, классы не считаются сформированными, а продолжают пополняться. Это связано с тем, что двойной просмотр точек, лежащих ниже границы, приводит к увеличению трудоемкости алгоритма.

Предложенный алгоритм ориентирован не на теоретическую плотность, а на ее непараметрическую оценку. При этом c и β могут быть выбраны меньше оптимальных, но результат классификации не изменится.
Поскольку трудоемкость всех этапов алгоритма классификации зависит от c и β такая возможность приводит к сокращению объема вычислений.

В третьей главе приведено описание разработанного пакета программ "NPCL".
Разработанное программное обеспечение предоставляет исследователю следующие возможности:
<u>при решении задач самообучения:</u>
- автоматическая классификация статистических данных, составленных из непрерывных признаков с симметричными и несимметричными законами распределения;
- агрегирование результатов классификации;

<u>при решении задач распознавания образов</u>:
- обучение распознаванию образов в условиях непараметрической неопределенности с позиций различных критериев оптимальности;
- распознавание образов при ограниченном объеме обучающей выборки;
- статистическое оценивание вероятности ошибки распознавания образов;

<u>при решении вычислительных проблем классификации</u>:
- формирование наборов информативных признаков;
- оптимизация непараметрических алгоритмов классификации;
- восстановление плотностей вероятности на основе интегральной непараметрической оценки и ее оптимизация;
- оформление многомерных результатов классификации в виде последовательности таблиц;
- отображение многомерных результатов классификации в пространство признаков размерностью более чем два;
- картирование результатов классификации при наличии соответствующих данных;

- корректировка исходных данных, параметров алгоритма, результатов расчета по требованию специалиста;
- ознакомление со справочным аппаратом, содержащим общие сведения о пакете, информацию об используемом математическом аппарате и категориях.

В четвертой главе рассматривается применение разработанных непараметрических методов и программ для типизации океанических вод по спектральным данным. Предлагается подход к построению статистических моделей взаимосвязи между обобщенными биооптическими показателями поверхностных вод океана при стационарных и нестационарных условиях.

Исходную информацию составили данные по физическим и биологическим параметрам поверхностных вод, полученные в 36 рейсе НИС "Ак. Вернадский" в Западной части тропической Атлантики (май август 1987г.).

Исследуемый район характеризуется весьма большим диапазоном изменений оптических свойств поверхности океана, так как для западной части полигона характерно наличие вод Амазонки, выходящих в океан, а северо-восточная и юго-восточная часть полигона это, соответственно, наличие северной субтропической водной массы с большим количеством взвешенного и растворенного вещества.

Используемые при анализе данные представляют собой набор спектров яркости, измеренных через шахту корабля в дневное время по ходу судна и на станциях. Спектральный диапазон измерения от 395 до 587 нм., содержащий 10 участков с шириной спектральных каналов 12-15 нм. Максимумы пропускания спектральных каналов находится на длинах волн 395, 420, 461, 503, 534, 556, 575 и 587 нм. Пространственное разрешение при скорости корабля 13 узлов составляло 30-80 метров.

Так как свет, попавший в шахту корабля, а затем в объектив спектрофотометра проходит через большую водную толщу, то динамика взвешенного биологического вещества и растворенной органики вызывает изменение спектрального состава принимаемого излучения.

Для устранения влияния условий освещенности параметры спектра излучения нормировались относительно яркости на длине волны 478 нм.

В связи со сложностью и динамичностью процессов в океане, определяемой сложной структурой течений, взаимодействием атмосферы и океана, изменчивой биологической структурой, как по составу так и в пространстве, зависимостью этой структуры как от внешних условий, так и от внутренних особенностей ее развития, исследуемая система имеет нестационарный характер.

Была разработана обобщенная статистическая модель, устанавливающая соответствие между макросостояниями системы, соответствующим некоторым областям в пространстве ее входных и выходных переменных. Например, между типами спектральных оптических характеристик океанических вод и интервалами содержания в них хлорофилла.

Пусть $y = y(v,t) \in Y$, $x = x(v,t) \in X$ векторы параметров, определяющие соответственно биологические и спектральные оптические характеристики океанических вод в конкретных временных $t \in T$ и пространственных $v \in V$ координатах.

В общем случае, пространственно - временная взаимосвязь $M(x,y,u)$ является сложной, зависящей от начальных значений (x_0, y_0) и от внешних условий U (температура, соленость характер пространственного распределения составляющих компонент и т.д.).

За вектор x примем значения индексов цвета в выбранных спектральных диапазонах

$$x = \left\{ x_1 = \frac{J_1}{J_5}, x_2 = \frac{J_2}{J_5}, \ldots, x_9 = \frac{J_9}{J_5} \right\}.$$

Под j-м типом S_x^j индексов цвета будем понимать "компактную" область $X^j \subset X$, соответствующей одномодальному фрагменту плотности вероятности P(x). Таким образом, множество S_x^j является обобщенным показателем спектральных характеристик наиболее характерных для конкретных условий. Данный обобщенный показатель является вероятностным, что обеспечивает его устойчивость и позволяет значительно снизить размерность модели исследуемой системы.

Если между x и y существует взаимосвязь, то каждому типу спектральных характеристик $S_x^j \in S_X$, $j=\overline{1,M}_X$ в пространстве биологических показателей соответствует некоторая область значений $S_y^j \in S_Y$, $j=\overline{1,M}_Y$. При этом не исключается условие $M_X > M_Y$, то есть некоторые S_y^v могут соответствовать нескольким типам S'_X из S_X. Это возможно, если их вероятностные характеристики достоверно не отличаются в рассматриваемом подмножестве S'_X.

Модель изучаемой системы представляется в виде последовательной смены во времени обобщенных показателей $\{S_X(t), S_Y(t), t \in T\}$, что может быть представлено в виде следующей логической схемы:

$$M(X'(t)) : \overset{1}{\downarrow} B(x'(t)) R(S_X, S_Y, x'(t)) (t = t+1) \gamma \overset{1}{\uparrow},$$

где B(·) - оператор ввода частично наблюдаемого вектора x(t);

R(·) - решающее правило, предназначенное для оценивания обобщенного показателя S_Y по значениям x'(t) ;

γ - логическое условие выхода из процесса прогноза S_Y.

Операторы алгоритма срабатывают слева направо, при выполнении условия процесс вычислений заканчивается, в противном случае осуществляется переход по стрелке.

Изменение условий U в системе M(x,y,u) влечет изменение закономерностей взаимосвязи между элементами множеств S_x, S_y.

Предположим существование взаимосвязи между изменяющимися условиями и некоторым набором компонент x" наблюдаемого вектора x'. Тогда, некоторой компактной области условий $U^i \subset U$, $i = \overline{1,N}$ в пространстве x" соответствует конкретная область $Q^i_{x"}$. Пусть $M_{Q^i_{x"}}(x'(t))$ - модель системы для сложившихся условий U^i. Тогда допустимо построение коллектива моделей $\{M_{Q^i_{x"}}(x'(t))\}$ для различных условий $U^i \subset U$ в режиме реального времени проведения полевого эксперимента (режим мониторинга). В данном случае $Q^i_{x"}$ представляет собой область компетентности i -ой модели.

По поступающим экспериментальным данным x , на основе имеющихся на данный момент моделей, оценивается последовательность обобщенных показателей S_y , и проводится их сравнение с экспериментальными значениями y в соответствии с принятым критерием $I(S_y,y)$. Если $\min_i I(S_y,y) > I^*$, то существующий набор моделей не является достаточным и соответствующие значения x', y накапливаются для последующего построения новых моделей, а x" из x' запоминается для построения их областей компетентности. Здесь I* заданные значения крите-

рия, а S_y^j - обобщенный показатель, определенный по i -ой модели из имеющегося набора.

При достижении достаточного объема наблюдений x', y , не согласующихся с существующими моделями, по предложенной методике, строится новая, которая дополняет имеющийся их набор. Тогда, обобщенная модель динамики взаимосвязи между x'(t), S_y представляется следующей логической схемой

$$M(X'(t)) : \overset{1}{\downarrow} B(x'(t)) R^i (X''(t)) M_{Q_x^i} (x'(t)(t = t+1) \gamma \overset{1}{\uparrow},$$

где B(x'(t)) - оператор ввода частично наблюдаемого вектора x'(t) ;
R^i(x"(t)) - алгоритм выбора i -ой модели;
$M_{Q_{x"}^i}(x'(t))$ - модель системы в условиях Q_x^i ;
γ - логическое условие выхода из процесса прогноза S_γ.

$$R^i(x''(t)) : x''(t) \in n\ Q_x^i, \text{ если } \overline{P}_i \overline{p}_i(x'') > \overline{P}_\nu \overline{p}_\nu(x''),\ \nu = \overline{1, N},\ \nu \neq i.$$

N - количество моделей в коллективе $\{M_{Q_{x"}^i}(x'(t))\}$, а $\overline{p}_i(x'')$ - непараметрическая оценка плотности вероятности компонент вектора $x" \in Q_{x"}^i$ в области компетентности i - ой модели, $\overline{P}_i$ - оценка априорной вероятности распределения $x" \in Q_{x"}^i$. Решающее правило реализуется с помощью непараметрических алгоритмов распознавания образов.

Предложенные методы были применены:

- для разработки критериев дифференциации и оценивания спектральных оптических характеристик поверхностных вод океана,
- картирования поверхностных вод океана по результатам классификации спектральных оптических характеристик,
- исследования взаимосвязи типов спектральных характеристик с содержанием хлорофилла.

В пятой главе рассматривается применение разработанных методов и программ для построения классификационной модели комплекса "сердечно - сосудистая система - углеводный обмен". Рассматриваются критерии диагностики нарушений толерантности к глюкозе по состоянию сердечно-сосудистой системы и другим косвенным признакам. Значение этой проблемы определяется большими затратами при традиционных методах диагностики сахарного диабета.

В качестве исходной информации для построения статистической модели комплекса "сердечно-сосудистая система углеводный обмен" использовались данные эпидемиологических исследований 1319 жителей Крайнего Севера. У обследуемых проводились стандартизованное измерение артериального давления (АД), уровня гликемии натощак и через 30, 60, 120 мин. после нагрузки 50г. глюкозы, а также измерение антропометрических показателей и анкетирование.

Была разработана статистическая модель комплекса систем с односторонним характером взаимодействия. Пусть комплекс систем $<S \rightarrow Q>$ характеризуется выходными (y^S,y^Q) и входными переменными (x,z) , отражающими взаимодействие комплекса с внешней средой. Вектор x=(ω,u) состоит из управляемых u и контролируемых ω воздействий. Причем переменные z не контролируются.

Система Q под воздействием y^S системы S в условиях (x,z) может находиться в одном из макросостояний Q_i, $i \in I_Q$, которые определяются с помощью соотношения

$$M_Q(y^Q): y^Q \in Q_i, \text{ если } F_{i\bar{i}}(y^Q, \alpha^Q) > 0,$$

где $F_{i\bar{i}}(y^Q, \alpha^Q)$ - уравнение разделяющей поверхности между Q_i и $Q_{\bar{i}} = \bigcup_{i \neq j} Q_j$, $j \in I_Q$, априори заданное с точностью до набора параметров $\alpha^Q \in \Delta_Q$.

Существует стохастическая зависимость $\varphi_Q : y^S \times X \to y^Q$. Дополнительно задана обучающая выборка (x^i, $y^Q(i)$, $y^S(i)$, $V^i(\alpha^Q)$), $i = \overline{1,n}$, составленная из наблюдений переменных (x, y^Q,y^S) и указаний $V^i(\alpha^Q)$ об их принадлежности к конкретному макросостоянию Q_i, $i \in I_Q$, формируемых на основе решающего правила.

С учетом стохастической зависимости $\varphi_Q : y^S \times X \to y^Q$ определим решающее правило

$$\overline{M}_Q(y^S, x) : (y^S, x) \in \overline{Q}_i, \text{ если } F_{i\bar{i}}(y^S, x, \overline{\alpha}, \alpha^Q) > 0$$

для оценивания в пространстве ($y^S \times X$) областей соответствующих состояниям системы Q.

Обозначим через $\rho_{\overline{F}}(\alpha^Q, \overline{\alpha})$ выражение ошибки прогноза макросостояний $\overline{Q}_i$, $i \in I_Q$, статистическая оценка $\overline{\rho}_{\overline{F}}(\cdot)$ которой может быть вычислена в режиме "скользящего экзамена". Пусть при некотором значении $\alpha^Q \in \Delta_Q$ определен вид уравнения разделяющей поверхности $\overline{F}_{i\bar{i}}(\cdot)$, $i \in I_Q$.

Тогда оптимизация $\overline{m}_Q(y^S, x)$ осуществляется в результате решения задачи

$$\min_{\alpha} \overline{\rho}_{\overline{F}}(\alpha^Q, \overline{\alpha}), \ \forall \alpha^Q \in \Delta_Q.$$

Основываясь на вышеизложенном, модель комплекса систем в терминах макросостояний запишется в виде семейств решающих правил

$$\left\{\overline{m}_i^Q(y^S, x), i \in I_Q\right\}\left\{\overline{m}_j^S(y^S, x), j \in I_S\right\},.$$

Вид $\varphi_Q(\cdot)$ априори не определен. При построении решающего правила используются методы непараметрической статистики.

Основываясь на предложенном подходе синтезирована структура комплекса $<S \to Q>$, проведен ее анализ. Установлена неоднородность состояний систем комплекса. В частности по показателям гликемической кривой выделяется 4 состояния углеводного обмена у женщин и 5 состояний у мужчин, достоверно отличающиеся между собой.

Основным дифференцирующим свойством системы углеводного обмена является возраст, остальные функциональные признаки и факторы оказывают комплексное влияние на формирование состояний системы.

Подтверждена неоднородность обследуемой популяции по показателям сердечно-сосудистой системы. Обнаружены по три состояния как у мужчин, так и у женщин близкие по своим значениям к общепринятой дифференциации (норма, пограничное состояние, артериальная гипертония).

Для исследования взаимосвязи между системами комплекса восстановлены операторы сопряжения между его состояниями, обнаруженными на предыдущем этапе синтеза структуры. С этой целью из условия минимума ошибки прогноза выделены наборы информативных признаков.

Наибольшее влияние на прогноз состояний (норма, нарушение толерантности к глюкозе) углеводного обмена у мужчин при конкретных состояниях сердечно-сосудистой системы оказывают следующие сочетания признаков: количество употребляемого алкоголя, уровень систолического и диастолического артериального давления, частота сердечных сокращений и индекс массы тела. При этом следует отметить, что признак "возраст" очевидно оказывает влияние на прогноз опосредованно через возрастные изменения выше перечисленных параметров. Ошибка прогноза составляет при этом 8%. Использование дополнительных признаков вегетативный индекс, характер трудовой деятельности снижает ошибку прогноза до 2%. У женщин характерен такой же набор информативных признаков, только признак "количество употребляемого алкоголя" заменяется возрастом. Ошибка прогноза состояний составляет при этом 9%.

При известном состоянии углеводного обмена уровни артериального давления, соответствующие норме, артериальной гипертонии и пограничному состоянию, в значительной мере (ошибка прогноза 4%) определяются индексом массы тела, степенью тяжести физического труда, количеством употребляемого алкоголя и показателями гликемии натощак и через 30 минут после углеводной нагрузки. Курение, возраст, вес, вегетативный индекс оказывают меньшее влияние на состояние сердечно-сосудистой системы, что не исключает, однако, возможности их опосредованного влияния.

Небольшое количество косвенных признаков, определяющих процесс взаимодействия систем, позволяет практически использовать полученные результаты в виде критериев оценивания состояний комплекса систем и провести дальнейший анализ свойственных им закономерностей.

По результатам анализа комплекса "сердечно-сосудистая система углеводный обмен" разработаны наборы таблиц, которые являются доступным средством для решения вопросов индивидуальной профилактики и прогноза развития нарушений углеводного обмена. Проверка предложенных критериев оценки состояний углеводного обмена по косвенным признакам осуществлялась Г.М. Шустом в Якутском республиканском эндокринном диспансере. Точность оценивания составила 93.3%.

Основные результаты и выводы.

1. Разработано методическое, алгоритмическое и программное обеспечение непараметрической системы классификации, позволяющее автоматизировать процесс обработки больших массивов статистических данных.
2. Предложенный непараметрический алгоритм автоматической классификации позволяет обнаруживать классы с несимметричным характером плотности вероятности и снижает требования к оптимальному выбору коэффициента размытости.
3. Трудоемкость модифицированного алгоритма непараметрической классификации пропорциональна квадрату объема выборки умноженному на коэффициент размытости интегральной оценки плотности вероятности.
4. Разработан диалоговый пакет программ "NPCL", функциональные возможности которого позволяют выполнять комплексную обработку разнотипной медико-экологической информации.
Ориентация математического обеспечения пакета на экспериментальные данные, позволяет ис-

пользовать его при исследовании объектов различной природы.
5. Разработана статистическая модель взаимосвязи между спектральными оптическими характеристиками поверхностных вод океана и концентрациями хлорофилла. Разработанные критерии на основе непараметрических алгоритмов классификации могут использоваться для определения концентрации хлорофилла в поверхностном слое воды;
6. Разработаны классификационные модели оценивания и прогнозирования состояний комплекса "сердечно-сосудистая система - углеводный обмен" в условиях Севера. Выделен набор наиболее информативных признаков, позволяющий прогнозировать состояния системы углеводного обмена у мужчин и женщин с ошибкой менее 9%.

Основное содержание диссертационной работы изложено в следующих публикациях:

1. Высоцкая Г.С., Лапко А.В., Поликарпов Л.С., Пироговский Н.В. Диспансеризация больных артериальной гипертонией в условиях Крайнего Севера. // Тезисы докл. 4 Съезда кардиологов, Москва, 1986. - N306
2. Высоцкая Г.С., Лапко А.В., Орехов К.В. и др. Комплекс систем "Сердечно-сосудистая система - система углеводного обмена": принципы моделирования, алгоритмы управления, результаты исследований. // Имитация систем в биологии и медицине: Материалы 5-го Пражского симпозиума соц. стран - Прага 1986. - N717
3. Высоцкая Г.С., Лапко А.В., Каленюк Н.М. Имитация в задачах исследования медико-биологических систем // Имитация систем в биологии и медицине: Материалы 5-го Пражского симпозиума соц. стран - Прага 1986. - N717
4. Лапко А.В., Высоцкая Г.С. Simulation and control of a complex of discrete-time systems when information is not complete // Advances in Modelling Simulation, AMSE Press, Paris, 1987. - vol 7. - N 2. - pp.18-20.
5. Седов К.Р., Лапко А.В., Высоцкая Г.С., и др. Статистическая модель комплекса "сердечно-сосудистая система - углеводный обмен" // Препринт ВЦ СО АН СССР, 1989г. - N16. - 18с.
6. Седов К.Р., Лапко А.В., Высоцкая Г.С., и др. Статистическая модель взаимодействия сердечно-сосудистой системы и углеводного обмена в экологических условиях Севера // Биофизические и биотехнические аспекты гомеостаза, Красноярск ИФ СО АН СССР, 1989г. - с.40-48.
7. Высоцкая Г.С. Диалоговый пакет прикладных программ для моделирования развивающихся медико-биологических систем // Математические модели и алгоритмы в задачах обработки данных, Красноярск, КГУ, 1993г. - с.128-137.
8. Лапко А.В., Высоцкая Г.С., Ануфриева Н.К. и др. Распознающие системы в задачах исследования и прогноза динамики древостоев // Математические модели и алгоритмы в задачах обработки данных, Красноярск, КГУ, 1993г. - с.21-37.
9. Высоцкая Г.С., Шевырногов А.П. Статистические модели в задачах оценивания динамики океанических биоценозов и сопряженных океанологических характеристик // препринт ИФ СО АН СССР, 1991г. - 47с.
10. Высоцкая Г.С. Диалоговый пакет программ "NPCL" // Непараметрические методы классификации и их применение, Новосибирск, Наука, 1993г. - с.131 -134.
11. Лапко А.В., Высоцкая Г.С., Секурцева Т.Т. - Непараметрические системы классификации. // Известия высших учебных заведений. Физика. - 1995г. N9 - с.90-95.
12. Поликарпов Л.С., Соустин В.П., Ченцов С.В., Высоцкая Г.С., Лапко А.В. Информационная технология комплексного исследования процессов в системе "человек - окружающая среда" при неполной информации. // Информационные системы в науке.- М.: РФФИ, 1995г. - с.68-69.
13. Высоцкая Г.С., Поликарпов Л.С., Хамнагадаев И.И., Щербаков В.В., Лапко А.В., Шуст Г.М.

Прогностическая значимость факторов риска сердечно-сосудистых заболеваний среди жителей Крайнего Севера. // Рискометрия и адаптация в медицине: (Материалы Всесоюзной научно-практической конференции, Иваново). - Иваново: Ивановская государственная медицинская академия, 1995.г. - с.12 -13.
14. Высоцкая Г.С., Лапко А.В., Ченцов С.В. Непараметрические системы распознавания образов в условиях больших выборок. // Распознавание образов и анализ изображений. Перспективные информационные технологии. Материалы Всероссийской конференции с международным участием (РОАИ-95). Ульяновск, 27 августа - 3 сентября 1995 г. - Ульяновск: Гос. Техн. университет, 1995г. - с.59 -61.
15. Поликарпов Л.С., Хамнагадаев И.И., Лапко А.В., Высоцкая Г.С. Прогнозирование ишемической болезни сердца у мужчин сельского населения Севера (методические рекомендации). //Красноярск: Краевое управление здравоохранения, 1995г. - 15с. 16. Лапко А.В., Высоцкая Г.С., Секурцева Т.Т., Поликарпов Л.С., Ченцов С.В. Информационная технология моделирования и принятия решений в системе "человек - окружающая среда"// Экологические аспекты устойчивого развития регионов. Тезисы международной конференции. Новгород, 22-25 сентября 1995 г. - Новгород: Нов. ГУ, 1995г. - с.101-107.

Работа выполнена в Институте вычислительного моделирования СО РАН

Научный руководитель: доктор технических наук, профессор Лапко А.В.

Научный консультант: доктор технических наук, Шевырногов А.П.

Официальные оппоненты: доктор технических наук, профессор Шайдуров Г.Я.
доктор технических наук, Семенкин Е.С.

Ведущая организация: Институт математики СО РАН

Защита состоится 17 апреля 1998 г. в 14^{00} часов на заседании диссертационного совета Д.064.54.01 Красноярского государственного технического университета по адресу: 660074, г. Красноярск, ул. Киренского 26.

С диссертацией можно ознакомиться в библиотеке Красноярского государственного технического университета.

Автореферат разослан 14 марта 1998 г.
Ученый секретарь
диссертационного совета, д.т.н., профессор А.Н. Ловчиков