ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 519.92

ПРИНЯТИЕ РЕШЕНИЙ В ТРУДНОФОРМАЛИЗУЕМЫХ ЗАДАЧАХ РАСПОЗНАВАНИЯ ОБРАЗОВ
I. Постановка задачи и подготовка статистического материала

Т.И. Чачхиани

Нижегородский госуниверситет им. Н.И. Лобачевского

ms@nounb.sci-nnov.ru

Поступила в редакцию 27.10.2011

Разработан научно-исследовательский программный комплекс для решения задач распознавания образов в прикладных областях. На примере задачи медицинской диагностики рассмотрен первый этап работы комплекса. Он включает в себя постановку задачи, разработку карты обследования, создание базы данных, предварительный анализ статистического материала.

Ключевые слова: распознавание образов.

Введение

Методы теории распознавания образов находят свое применение в самых различных прикладных областях и научных исследованиях. Работы по развитию теории и методов распознавания велись коллективом математиков кафедры теории управления и динамики машин и НИИ ПМК под руководством академика РАЕН Ю.И. Неймарка с середины 60-х годов прошлого столетия [1–3, 5–7]. За это время был создан целый ряд алгоритмов и разработаны методики решения различных прикладных задач.

Предлагаемый научно-исследовательский программный комплекс обобщает этот опыт и показывает пути решения трудноформализуемых задач распознавания образов.

В качестве примера работы комплекса предлагается рассмотреть задачу медицинской диагностики. Проблемы, возникающие в медицине и требующие применения математических методов и компьютерного моделирования, связаны с постановкой диагноза и его достоверностью, выбором метода лечения, показанием к оперативному вмешательству и прогнозированием его исхода [1, 2].

Постановка диагноза или прогноза представляет собой сложную задачу, и до сих пор очень часто получаемый результат зависит от интуиции и опыта врача. Эти задачи являются трудноформализуемыми.

1. Составление карты обследования, кодирование признаков и статистическогоматериала

Основой для постановки диагноза или прогноза является описание состояния человеческого организма, окружающих его условий и воздействий.

1.1. Карта обследования и кодирования признаков

Будем называть картой обследования набор признаков, который характеризует состояние человека при определенном его заболевании. В зависимости от поставленной задачи в карту обследования включаются признаки, которые являются наиболее важными по данным медицинской литературы и по опыту и наблюдениям врачей.

Обычно врачами отбирается большая группа признаков, которые достаточно подробно описывают характер и локализацию болей, жалобы больного, его анамнез (возраст, пол, перенесенные заболевания в прошлом и т.д.). В карту обследования включаются результаты лабораторных исследований, данные рентгеноскопии, электрокардиографии и другие показатели.

Ведущими принципами при отборе признаков для карты обследования являются следующие [1]:

1. Карта обследования должна быть информативной, лаконичной, пригодной в условиях массового обследования.

2. Набор признаков, включенных в карту обследования, должен быть достаточным для решения поставленной задачи. Карта должна содержать большое число признаков, поскольку малый набор их может оказаться недостаточным для характеристики состояния больного с нерезко выраженной симптоматикой. В то же время избыточность признаков может привести к тому, что уровень помех при небольшом статистическом материале превысит уровень ценной информации.

3. Признаки должны быть четко определены, иметь одинаковое толкование различными врачами.

Все признаки можно подразделить на две группы. В первой группе они характеризуются количественными показателями. К ней относятся такие признаки, как возраст, температура, частота пульса, кровяное давление, жизненный объем легких, дозы назначаемых лекарств и т. д. Вторую группу признаков составляют сведения о больном в медицинских терминах и понятиях, которые не имеют количественного выражения. К этой группе относятся различные симптомы заболевания (наличие и характер болей, вид кожного покрова, заболевания в прошлом и др.).

Признаки, отобранные в карту обследования, подразделяются на градации. Например, признак «одышка» имеет три градации: нет одышки, есть одышка при физической нагрузке, есть одышка в покое.

Непосредственное кодирование признаков состоит в присвоении каждой градации признака некоторого числового показателя (кодового числа или веса).

Для перевода всех градаций признаков в цифровые значения разрабатываются специальные кодировочные таблицы. В основу кодирования градаций признаков положен принцип нарастания кодового числа по мере монотонного нарастания выраженности признака.

При кодировании признаков первой группы область изменения каждого признака разбивается на несколько интервалов (градаций) в зависимости от требуемой степени детализации описания признака и решаемой задачи. Так, в случае диагностики гипертрофии левого и правого желудочков сердца по возрасту выделены следующие градации:

менее 30 лет – код 0,
от 31 до 40 – код 1,
от 41 до 50 – код 2,
более 50 лет – код 3.

Признаки, входящие во вторую группу, также подразделяются на градации, которым затем присваиваются те или иные числовые значения. В случае, когда речь идет о признаке или симптоме, в отношении которого имеет смысл понятие больше или меньше или сильнее и слабее, естественно присваивать градациям признака числовые значения в порядке его монотонного изменения. Например, признак «головная боль» можно характеризовать следующими градациями:

нет – 0,
слабая – 1,
сильная – 2.

При кодировании обеих групп градациям присваиваются кодовые числа 0,1,…,k, где k – максимальное число градаций в данной задаче. В некоторых случаях для усиления значимости градации целесообразно увеличить разрыв между градациями, пропуская некоторые кодовые числа.

Иногда в карту обследования включаются признаки, которые имеются не у всех больных. Например, данные анализа крови могут отсутствовать, если у больного анализ крови не проводился. Числом k+1 кодируется отсутствие данных в соответствующем признаке. В дальнейшем о таких случаях мы будем говорить «непроверенный признак».

Признак, указывающий диагноз заболевания больного или оценку его состояния, обозначим через π. Он также кодируется целыми числами 0,1,2,…,k. Например, в задаче дифференциальной диагностики заболеваний сердца, органов дыхания и почек у детей по данным приемного покоя для диагноза были приняты следующие кодовые числа:

$\pi = 0$ – заболевания органов дыхания,
$\pi = 1$ – заболевания сердца,
$\pi = 2$ – заболевания почек.

В случае прогноза признак π имеет непрерывный характер и лишь условно может быть разбит на градации. Кодирование его производится в соответствии с тяжестью заболевания: монотонному изменению состояния больного должно соответствовать монотонное нарастание кодовых чисел 0,1,2,…,k.

1.2. Кодирование статистического материала

Таким образом, карта обследования содержит: признак π, обозначающий диагноз или прогноз заболевания, названия признаков и их градаций, принятых для описания состояния больного, и кодовые числа градаций.

В соответствии с этой картой каждая история болезни представляется в виде набора целых чисел.

Обозначим через M общее число признаков, включенных в карту обследования, а через x_j – j-й признак в ней. Тогда история болезни i-го больного представится в виде набора $\{\pi_i, x_1^i, x_2^i, ..., x_M^i\}$.

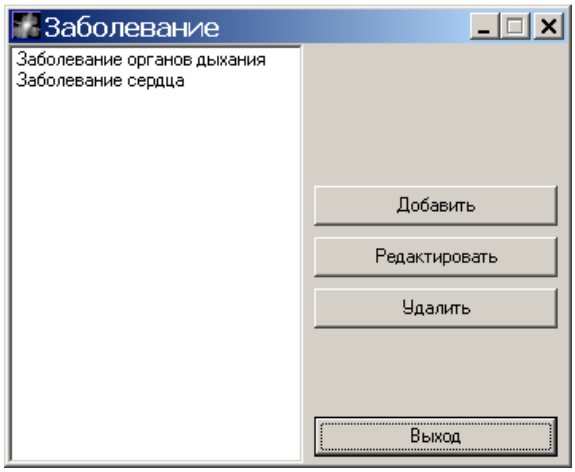

Рис.1. Ввод заболеваний

Множество таких наборов составляет статистический материал для решения задачи.

Перейдем к рассмотрению конкретной задачи медицинской диагностики: к дифференциальной диагностике заболеваний у детей по данным приемного покоя детской больницы.

Как уже ранее упоминалось, для диагноза были приняты следующие кодовые числа:

$\pi = 0$ – заболевания органов дыхания, $\pi = 1$ – заболевания сердца, $\pi = 2$ – заболевания почек.

Для характеристики состояния ребенка при данных заболеваниях в карту обследования был включен 31 признак (рис. 2, 3).

Статистический материал в данной задаче включает в себя 47 историй болезни детей с заболеваниями органов дыхания, 50 историй болезни детей с заболеваниями сердца и 49 – с заболеваниями почек.

1.3. Формирование баз данных

Решение задачи реализуется с помощью научно-исследовательского программного комплекса «Принятие решений в трудноформализуемых задачах распознавания образов» [4, 6, 7].

Для этого на первом этапе необходимо создать базы данных, ввести карту обследования и кодирования признаков, а затем статистический материал, образующий обучающую выборку для построения решающих правил.

При формировании обучающей выборки для каждого заболевания отбираются больные, имеющие достоверный диагноз одного из заболеваний и не имеющие признаков другого заболевания из рассматриваемых в задаче.

Эти задачи можно выполнить с помощью первого модуля комплекса «Подготовка обучающего материала» [6]. Он позволяет использовать единую базу данных для хранения всех данных по признакам и градациям, входящим в таблицу, и формировать на основе введенной таблицы статистический материал. При работе с новым статистическим материалом сначала определяются списки заболеваний и на основе карты обследования создается кодировочная таблица.

В окне ввода списка заболеваний (рис. 1) предоставлены возможности по добавлению, редактированию, а также удалению заболевания (диагноза).

Признаки и градации из карты обследования и кодирования добавляются последовательно в конец соответствующего списка (рис. 2). По необходимости можно редактировать уже внесенные данные.

Для того чтобы добавить данные больного в статистический материал, последовательно задаются:

– Ф.И.О. текущего больного;

– соответствующий диагноз больного, который выбирается из списка заболеваний;

– соответствующие истории болезни больного градации признаков, которые выбираются из кодировочной таблицы.

Градации присваиваются последовательно по порядку представления в списке, от каждого признака по одной градации.

На рис. 3 представлено главное окно программы, содержащее список больных с присвоенными диагнозами и градациями признаков.

На основе введенной кодировочной таблицы и рабочего статистического материала создается также экзаменационная база в целях ее использования при построении решающего правила.

2. Анализ статистического материала

Подготовленный статистический материал необходимо подвергнуть анализу, по результатам которого можно сделать вывод о возможности использования этого материала для решения задачи медицинской диагностики.

Первый этап анализа статистического материала включает в себя следующие действия: выделение дифференцирующих признаков, выявление и коррекция непроверенных признаков, коррекция статистического материала.

2.1. Подсчет частот встречаемости градаций признаков

Если в одном классе определенная градация признака встречается в несколько раз чаще, чем в другом, то эта градация может быть названа дифференцирующей. Наоборот, если некоторая градация встречается в разных классах с одина-

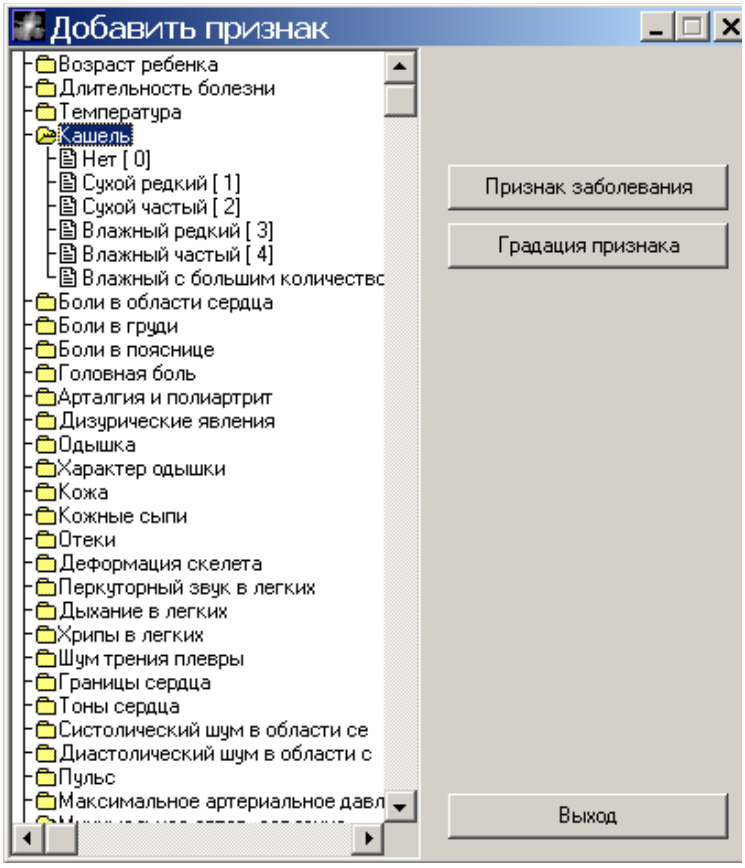

Рис. 2. Ввод градаций признаков

ковой частотой, то, как правило, она не является дифференцирующей и не может быть использована для разделения классов. Такие градации обозначим как незначащие для данного статистического материала. За счет исключения незначащих признаков можно в ряде случаев упростить задачу, уменьшив размерность пространства признаков.

Чтобы судить о дифференциальной значимости признаков и их градаций, необходимо иметь данные о частотах встречаемости градаций признаков в каждом классе больных с различными значениями признака π. Может оказаться, что у всех больных с $\pi = 0$ признак $x_{10} = 1$, а у всех больных с $\pi = 1$ он равен нулю. Тогда признак x_{10} является дифференцирующим, он разделяет больных на два класса: X_0 ($\pi = 0$) и X_1 ($\pi = 1$).

Для получения чисел и частот встречаемости градаций признаков в разных классах предназначен модуль подсчета частот встречаемости градации. С его помощью ведется подсчет количества градации каждого признака в классе больных с различными значениями π и вычисляются частоты встречаемости градаций признаков:

$$p_{ij}^{\pi} = \frac{a_{ij}^{\pi}}{I_{\pi}},$$

где a_{ij}^{π} – число больных класса X_{π} с градацией j в i-м признаке, I_{π} – число больных в классе X_{π}.

Кроме поиска дифференцирующих градаций, в статистическом материале следует найти истории болезней с непроверенными признаками, т.е. такие случаи, когда нет данных о градациях некоторых признаков. Для этого по договоренности отсутствие данных по признаку в карте больного и кодировочной таблице кодируется максимальным из кодовых чисел всех градаций и названием градации «непроверенное». Истории болезней с непроверенными признаками не могут быть использованы в качестве обучающего материала для построения решающих диагностических правил. Такие истории должны быть дополнены недостающими анализами или, если это невозможно, исключены из статистического материала.

2.2. Демонстрация работы по анализу статистических данных

Подсчет частот встречаемости градаций признаков

Для подсчета частот встречаемости используется модуль «Анализ статистического материала» [7]. Результат подсчета формируется в виде таблицы, в которой для каждой градации признака выводятся:

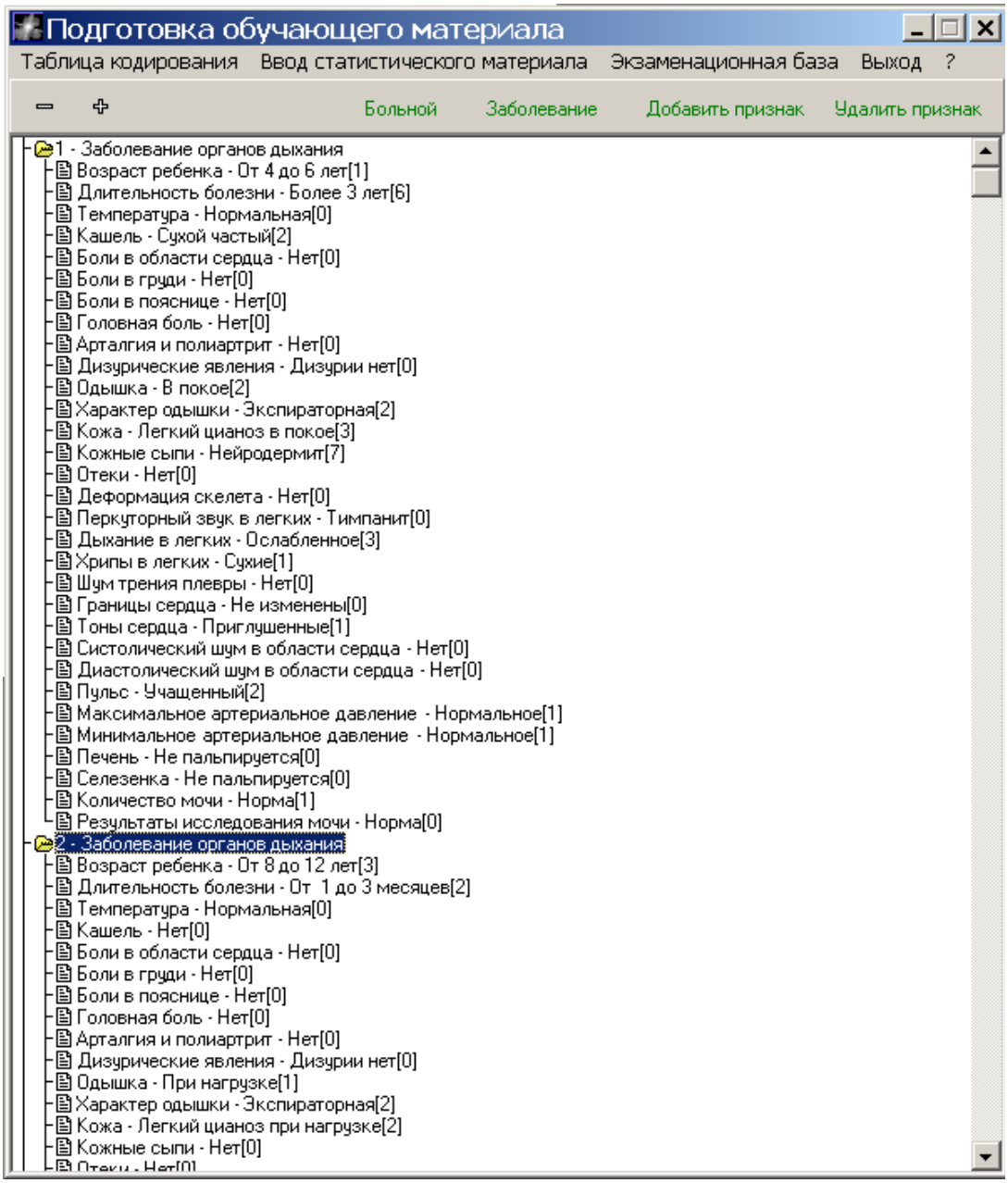

Рис. 3. Статистический материал в главном окне программы

1. Количество больных каждого класса с данной градацией.

2. Процент количества больных с данной градацией в массе больных класса (частота встречаемости градации).

Эти данные выводятся для каждого класса и для всего массива больных последовательно (форма представления по классам) или одновременно (форма представления по признакам).

В случае вывода по классам (рис. 4) происходит упорядочивание частот встречаемости градаций признаков по убыванию их значений.

То есть наиболее характерные градации признака для рассматриваемого класса оказываются на первом месте. Это позволяет оценить выраженность признаков и представительность обучающей выборки, получить ее общую характеристику.

На рис. 4 видно, что у признака № 11 (одышка) градация 1 (одышка при нагрузке) встречается почти у 70 % больных в классе 1 (с $\pi = 0$).

Таким же образом можно получить оценку ведущих признаков каждого класса.

Представление частот встречаемости по признакам облегчает анализ данных, который заключается, прежде всего, в сравнении частот встречаемости градаций разных классов с целью поиска градации, частоты встречаемости которой диаметрально противоположны (т.е. максимальны в одном классе и минимальны в другом).

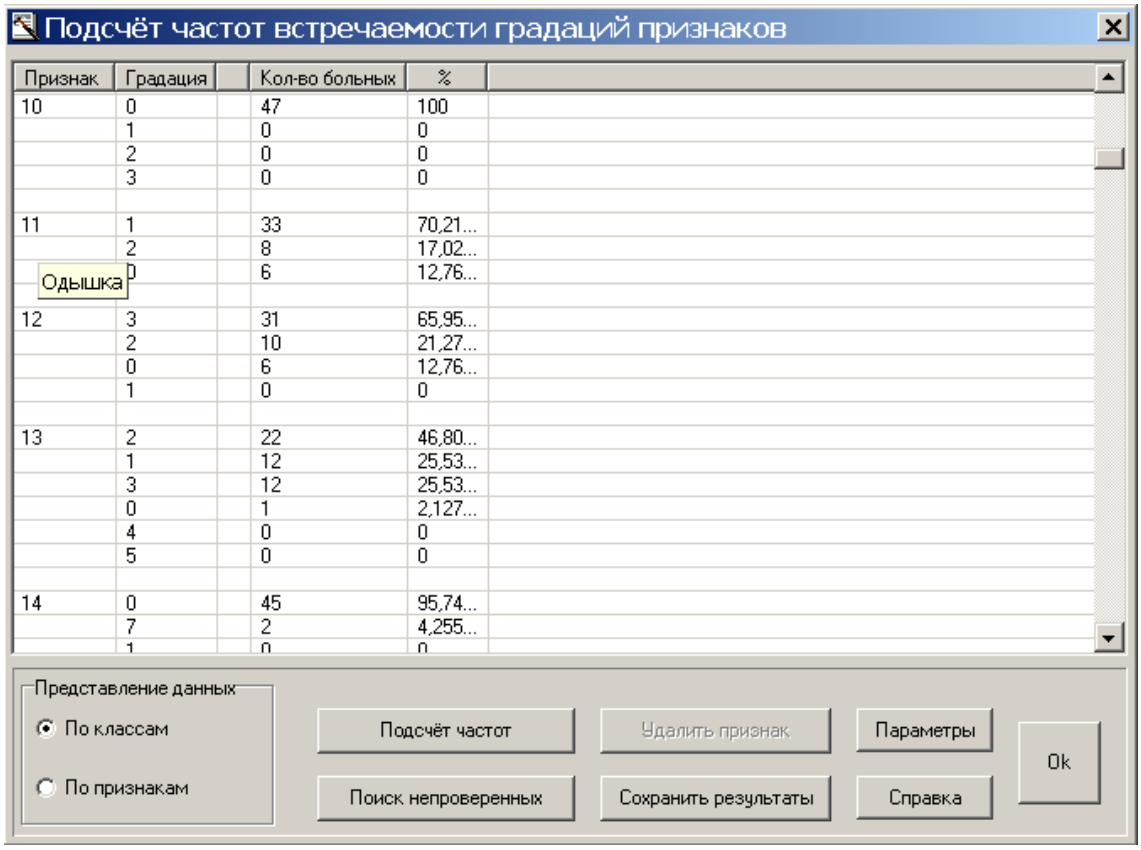

Рис. 4. Форма представления по классам для признака № 11

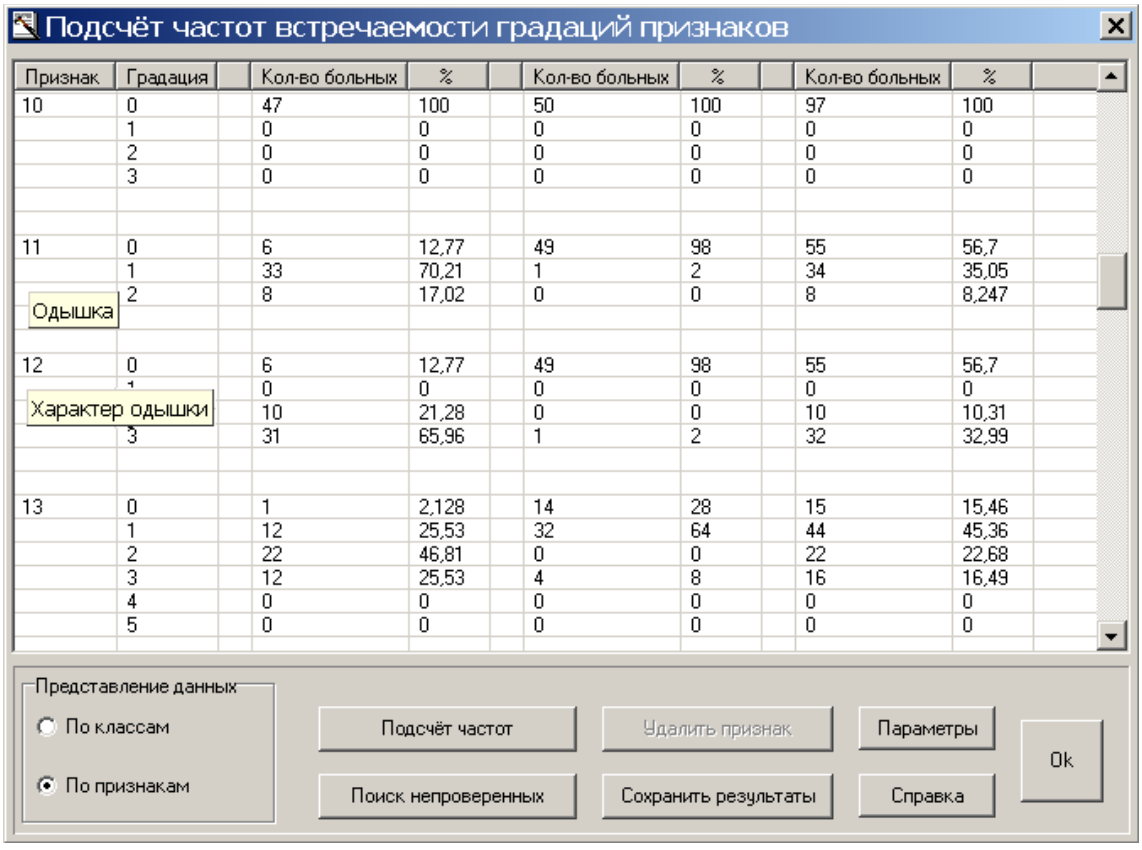

Рис. 5. Форма представления по классам для признаков № 11, 12

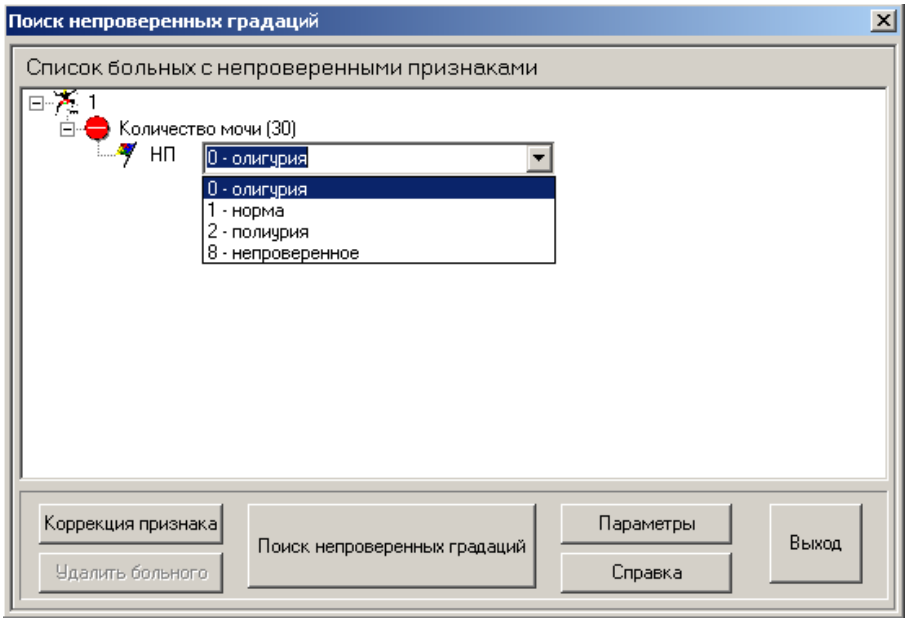

Рис. 6. Поиск непроверенных признаков

Таким образом, можно выделить дифференцирующие признаки. Например, признаки № 11, 12 будут дифференцирующими по градации с кодом 0 (нет одышки), так как частота встречаемости этой градации в классе 1 минимальна, а в классе 2 (с $\pi = 1$) – максимальна (рис. 5).

Незначащие признаки (с максимальными значениями частот встречаемости в обоих классах), такие как признак № 10 (рис. 5), после обсуждения с врачами можно отметить и удалить. Это позволит снизить размерность пространства признаков.

Для поиска непроверенных градаций используется соответствующий по названию модуль. При запуске модуля «Поиск непроверенных признаков» (из окна подсчета частот градаций признаков) все фамилии больных с непроверенными признаками (НП) будут найдены и выведены в виде раскрывающегося списка при соответствующем больном (рис. 6) для коррекции таких признаков или удаления фамилий больных.

Также при необходимости можно провести анализ частот встречаемости непроверенных признаков. Признаки с большими частотами встречаемости непроверенных градаций необходимо удалить. Если у достаточно большого количества больных какой-либо признак оказывается непроверен, то нужно удалить такой признак из карты обследования.

В случае, если какой-либо больной имеет много непроверенных признаков, следует удалить такого больного из статистического материала. Вообще все непроверенные признаки (НП) должны быть найдены и обработаны.

Анализ результатов работы на этом этапе решает несколько задач:

1. Прежде всего, выделение дифференцирующих и незначащих признаков позволяет сделать вывод о возможности решения поставленной задачи по построению диагностических решающих правил. Если дифференцирующие признаки отсутствуют или их мало, то кодировочная таблица и статистический материал не пригодны для дальнейшей работы, так как у больных, принадлежащих разным классам, признаки заболевания одинаковы или мало отличаются. Следовательно, необходимо вернуться к уточнению постановки задачи, разработке карты обследования и подготовке статистического материала.

2. В ряде случаев удается упростить задачу, исключив незначащие признаки, и тем самым уменьшить размерность пространства признаков.

3. Кроме того, определяется процент больных, имеющих непроверенные признаки. Этих больных нельзя использовать при построении диагностических правил. В случае если таких больных немного, то следует исключить больного. Если же признак оказывается непроверен у большинства больных, должен быть удален сам признак.

На этом этапе подготовка статистического материала для построения решающих правил заканчивается.

Список литературы

1. Распознавание образов и медицинская диагностика / Неймарк Ю. И. и др. М., 1972. 328 с.

2. Трошин М.В., Образцова Н.Д., Скорнякова Б.Л., Чачхиани Т.И. Математическая модель фазности течения черепно-мозговой травмы // В кн.: Клинико-кибернетические подходы к проблеме диагноза и прогноза черепно-мозговой травмы. Горький: ГИТО, 1982. С. 88–110.

3. Чачхиани Т.И. Опыт использования математических методов в анализе социологической информации // Межвуз. сб.: Социальные аспекты перестройки управления обществом. Горький: Горьковский ун-т, 1989. С. 116–119.

4. Чачхиани Т.И., Ивковская О.В., Макушина Н.С., Зиновьева Н.А. Принятие решений в задачах распознавания образов // Труды VI Международной конференции женщин-математиков. Н. Новгород: ННГУ, 1999. Т. 6. Вып.1. С. 86–88.

5. Чачхиани Т.И., Кузмичев Ю.Г., Вахрушева Е.А. Диагностический комплекс ЭКГ-показателей поражения сердца при дифтерии у детей // Материалы Межд. конф. «Математика. Образование. Экология». М.: Прогресс-Традиция, 2001. Т. 2. С. 198–203.

6. Диалоговая система принятия решений в задачах распознавания образов. Часть 1.: Методическая разработка / Сост. Т.И. Чачхиани, М.Г. Серова. Н.Новгород: ННГУ, 2004. 16 с.

7. Диалоговая система принятия решений в задачах распознавания образов. Часть 2.: Методическая разработка / Сост. Т.И. Чачхиани, М.Г. Серова. Н.Новгород: ННГУ, 2006. 14 с.

DECISION MAKING IN POORLY FORMALIZABLE PROBLEMS OF PATTERN RECOGNITION.
Part 1. Problem statement and preparation of statistical material

T.I. Tchatchiani

A research software package for pattern recognition in different applied areas has been developed. The first stage of the package performance is considered by the example of medical diagnostics. It includes the statement of the problem, the development of a medical record card, creation of a database, and a preliminary analysis of statistical material.

Keywords: pattern recognition.