На правах рукописи

Кузоваткин Андрей Николаевич

АЛГОРИТМЫ ПОСТРОЕНИЯ СМЕШАННЫХ ДИАГНОСТИЧЕСКИХ ТЕСТОВ И ПРИНЯТИЯ РЕШЕНИЙ

05.13.18 – математическое моделирование,
численные методы и комплексы программ

Автореферат диссертации
на соискание ученой степени
кандидата технических наук

Научный руководитель:
Доктор технических наук, профессор
Янковская А.Е.

Томск – 2005

Работа выполнена в Томском государственном университете.

Научный руководитель: доктор технических наук, профессор
Янковская Анна Ефимовна.

Официальные оппоненты: доктор технических наук, профессор
Матросова Анжела Юрьевна;

кандидат технических наук, доцент
Ходашинский Илья Александрович.

Ведущая организация: Институт "Кибернетический центр"
Томского политехнического
университета.

Защита состоится 24 февраля 2005 в 10 час. 30 мин. на заседании диссертационного совета Д 212.267.08 в 102 аудитории 2-го корпуса Томского государственного университета по адресу: г. Томск, пр. Ленина 36.

С диссертацией можно ознакомиться в библиотеке Томского государственного университета.

Автореферат разослан 21 января 2005 г.

Ученый секретарь
диссертационного совета
доктор технических наук, доцент Скворцов А.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Автоматизацию обработки информации невозможно представить без применения широкого спектра методов и моделей классической и дискретной математики, математического моделирования, распознавания образов, искусственного интеллекта и реализующих их программных средств, включая интеллектуальные системы.

Основополагающие результаты в области распознавания образов, искусственного интеллекта и интеллектуальных систем получены Журавлевым Ю.И., Ларичевым О.И., Осиповым Г.С., Поповым Э.В., Поспеловым Г.С., Поспеловым Д.А., Финном В.К. Большой вклад в эти области внесены также Берштейном Л.С., Вагиным В.Н., Гладуном В.П., Еремеевым А.П., Загоруйко Н.Г., Закревским А.Д., Зенкиным А.А., Кобринским Б.А., Кузнецовым О.П., Курейчиком В.М., Лбовым Г.С., Мелиховым А.Н., Нариньяни А.С., Переверзевым-Орловым В.С., Печерским Ю.Н., Рудаковым К.В., Рязановым В.В., Стефанюком В.Л., Тарасовым В.Б., Хорошевским В.Ф., Янковской А.Е. и др.

Тестовые методы распознавания образов и принятия решений являются основой создания интеллектуальных систем, находящих применение в различных проблемных областях, к которым относятся: медицина, генетика, экология, геология, психология, социология, экономика, криминалистика, архитектура и строительство, управление, образование, научные исследования и ряд других.

Процедуры построения безусловных и условных тестов, используемых для распознавания образов, являются весьма трудоемкими, и их оптимизации уделяется большое внимание в трудах Журавлева Ю.И., Дюковой Е.В., Мошкова М.Ю., Янковской А.Е. и ряда других ученых.

Таким образом, до сих пор весьма актуальными являются как разработка менее трудоемких алгоритмов тестового распознавания, приводящих к построению "хороших" тестов, то есть, обеспечивающих лучшее качество решений (по простоте решающих правил, точности принимаемых решений и минимизации затрат на их получение), так и создание на базе этих алгоритмов интеллектуальных систем, инвариантных к проблемным областям.

Именно на обеспечение лучшего качества принимаемых решений и направлена диссертационная работа, посвященная алгоритмам построения смешанных диагностических тестов (СДТ) и принятия решений. Обоснования целесообразности применения СДТ в интеллектуальных системах приводятся в работах Янковской А.Е.

Актуальность исследований подтверждается поддержкой Российского фонда фундаментальных исследований (проекты: № 98-01-00295 "Логико-вероятностный вывод на основе оптимальных смешанных диагностических тестов, частичной импликации и средств когнитивной графики в интеллектуальных системах"; № 01-01-0772 "Логические тесты, логико-комбинаторно-вероятностный вывод и средства когнитивной графики в интеллектуальной системе"; № 01-01-01050 "Развитие интеллектуальной системы логико-комбинаторного принятия решений, основанной на матричном представлении знаний"; № 04-01-00144 "Технология решения связанных по экспертному заключению задач, основанная на логических тестах и средствах когнитивной графики в интеллектуальной системе", в выполнении которых принимал участие диссертант).

Цель работы. Развитие методов построения СДТ при матричном представлении данных и знаний в пространстве двоичных, троичных, номинальных, целочисленных признаков и реализация этих методов в интеллектуальном ин-

струментальном средстве ИМСЛОГ (ИИС ИМСЛОГ), предназначенном для создания прикладных интеллектуальных систем выявления закономерностей в данных и знаниях и принятия решений для широкого круга проблемных областей.

Для достижения данной цели были поставлены следующие задачи:

1. Для матричного представления данных и знаний в пространстве двоичных, троичных, номинальных, целочисленных признаков разработать алгоритмы построения СДТ и принятия решений на их основе, а также алгоритмы принятия решений непосредственно в процессе построения смешанного диагностического теста в пространстве троичных признаков.
2. Разработать алгоритмы построения безызбыточной двоичной матрицы импликаций (для интервального и константного задания значений целочисленных признаков), задающей различимость объектов из разных образов.
3. Осуществить программную реализацию перечисленных выше алгоритмов в виде программных модулей ИИС ИМСЛОГ.
4. Создать на базе ИИС ИМСЛОГ прикладные интеллектуальные системы для решения задач в следующих областях: "Генетические изменения (оценка состояния здоровья ребенка по генетическим и функциональным показателям)"; "Квалифицированная и специализированная помощь при лучевых поражениях"; "Организационно-управленческие решения при чрезвычайных ситуациях"; "Оценка научно-технических проектов".

Методы исследования. В диссертационной работе применялись методы дискретной математики, распознавания образов, искусственного интеллекта, прикладного и системного программирования.

Научная новизна содержится в следующих результатах:

1. Разработаны новые алгоритмы построения СДТ при представлении данных и знаний в пространстве троичных, номинальных и целочисленных признаков.
2. Разработаны новые алгоритмы принятия решений непосредственно в процессе построения СДТ на основе безусловных тестов в пространстве троичных, номинальных и целочисленных признаков.
3. Получены формулы для вычисления коэффициента разделения, используемого при построении дерева СДТ.
4. Разработаны новые алгоритмы построения безызбыточной двоичной матрицы импликаций, используемой для выявления закономерностей в данных и знаниях в пространстве троичных, номинальных и целочисленных признаков.

Апробация работы. Основные положения и результаты диссертации представлены в докладах: 3-й Всероссийской конференции с международным участием "Новые информационные технологии в исследовании дискретных структур" (Томск, 2000); научной сессии МИФИ–2002 (Москва, 2002); Международной научной конференции "Интеллектуализация обработки информации (ИОИ–2002)" (Симферополь, 2002); Международном семинаре "Компьютерная лингвистика и интеллектуальные технологии (Диалог–2002)" (Протвино, 2002); международной научно-технической конференции "Интеллектуальные САПР" (Таганрог, 2002); Международной конференции "Информационные системы и технологии (ИСТ-2003)" (Новосибирск, 2003); второй Международной конференции по проблемам управления (Москва, 2003); VI международном конгрессе по математическому моделированию (Нижний Новгород, 2004).

Публикации. Основные результаты диссертации были опубликованы в 11 работах, две из которых в журнале, а одна в периодическом сборнике.

Структура и объем работы. Диссертационная работа состоит из введения, пяти глав и заключения. Объем работы составляет 106 страниц. Список литературы содержит 91 наименование.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, дана общая характеристика работы.

В первой главе приведены модели представления знаний, основные подходы к принятию решений, а также методы и алгоритмы построения логических тестов.

Описываются следующие модели представления знаний: продукционная модель, фреймовая модель, семантическая сеть, логическая модель. Также приводится используемая в диссертационной работе матричная модель представления данных и знаний, предложенная Янковской А.Е.

В матричной модели знания представляются в виде: матрицы описаний (Q), элементы которой принимают двоичные, троичные, k-значные значения; целочисленных матриц различений (R) трех типов. Первый тип матрицы различений характеризуется вложенностью механизмов классификации, когда каждый последующий столбец задает более подробное разбиение объектов на классы эквивалентности, чем предыдущий. Второй тип может интерпретироваться как совокупность действий, которые необходимо выполнить для данного объекта (данной ситуации). Третий тип служит для представления независимых механизмов классификации, где столбцы отражают, например, мнения различных экспертов. Наличие матрицы различений 1-го или 3-го типа, обязательно. Строка матрицы различений задает обобщенный класс (образ). Если имеется единственный механизм классификации, матрица различений вырождается в столбец, а обобщенный класс превращается в класс, что соответствует традиционному представлению знаний в задачах распознавания образов.

Для изложения алгоритмов используется вектор-столбец образов (R'), строки которого сопоставлены объектам обучающей выборки (строкам матрицы Q), а значениями являются номера образов (обобщенных классов).

Данная модель позволяет представлять не только данные, но и знания экспертов, поскольку одной строкой матрицы описаний можно задавать в интервальной форме подмножество объектов, для которых характерны одни и те же решения, задаваемые строкой матрицы R.

При обнаружении пересечения описаний образов матрица описаний должна быть доопределена.

На рис. 1 приведен пример матриц Q и R, а также вектор-столбец образов R' для представления знаний с использованием троичных (1,0,–) признаков, где символ "–" означает, что значение признака безразлично (признак может принимать как нулевое, так и единичное значение).

Рассмотрены следующие методы и алгоритмы построения логических тестов: алгоритм КОРА (Бонгард М.М.), методы построения тупиковых тестов (Журавлев Ю.И.), алгоритм построения безусловных тестов на основе нахождения минимальных столбцовых покрытий безызбыточной матрицы импликаций (Янковская А.Е., Гедике А.И.), алгоритмы построения безусловных тестов непосредственно по матрицам описаний и различений (шагово-циклические алгоритмы) (Янковская А.Е.), алгоритм синтеза граф-схем для распознавания образов (Орлов В.А.), алгоритмы построения смешанных диагностических тестов на основе безусловных (Янковская А.Е., Кузоваткин А.Н.).

Во второй главе приводятся основные определения и понятия, используемые в диссертационной работе, вводится коэффициент разделения и приводятся неформальные описания алгоритмов построения смешанных диагностических тестов и принятия решения.

В диссертации используется понятие закономерностей, предложенное Янковской А.Е. К закономерностям относятся константные, устойчивые (константные внутри образа), неинформативные (не различающие ни одной пары объектов), альтернативные (в смысле включения в диагностический тест), зависимые (в смысле включения подмножеств различимых пар объектов), несущественные (не входящие ни в один безызбыточный тест), обязательные (входящие во все безызбыточные тесты), псевдообязательные (входящие в множество используемых при распознавании безызбыточных тестов и не являющиеся обязательными) признаки, а также все минимальные и все (либо часть – при большом признаковом пространстве) безызбыточные различающие подмножества признаков, являющиеся, по сути, соответственно минимальными и безызбыточными (тупиковыми) логическими тестами. К закономерностям также относятся и весовые коэффициенты признаков, характеризующие их индивидуальный вклад в различимость объектов из разных образов.

$Q =$

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	0	–	0	1	0	0	1	–	1	1	0
2	0	0	0	0	–	1	0	0	1	1	1	1	0
3	0	1	0	0	0	0	0	–	1	1	–	1	1
4	0	1	0	0	0	0	0	–	1	1	–	1	1
5	0	1	1	1	0	1	–	0	0	0	1	–	0
6	0	1	1	1	1	1	–	0	0	0	1	–	0
7	0	1	0	0	0	0	1	–	1	1	–	0	1
8	0	1	1	1	1	0	1	–	0	0	–	0	1
9	0	1	–	–	1	1	–	1	–	–	0	–	0
10	0	1	1	1	1	1	–	1	0	0	0	–	0
11	1	1	0	0	–	–	1	1	1	1	0	0	–
12	1	1	1	0	–	1	0	0	0	1	1	1	0
13	1	1	0	1	0	–	1	1	1	0	0	0	–
14	1	1	1	1	–	0	1	1	0	0	0	0	1

$R =$

	1	2
1	1	2
2	1	2
3	1	2
4	1	2
5	1	2
6	1	2
7	2	1
8	2	1
9	2	2
10	1	1
11	2	1
12	2	1
13	2	2
14	1	1

$R' =$

	1
1	1
2	1
3	1
4	1
5	1
6	1
7	2
8	2
9	3
10	4
11	2
12	2
13	3
14	4

Рис. 1. Матричное представление данных и знаний для троичных характеристических признаков

Реакция матрицы описаний на тест есть совокупность строк матрицы описаний, в которых значения признаков, входящих в тест, совпадает со значениями этих же признаков в описании исследуемого объекта.

Термин и понятие "смешанный диагностический тест" (СДТ), представляющий собой оптимальное сочетание безусловных и условных составляющих, введены в 1996 г. Янковской А.Е.

СДТ представляет собой дерево, ярусам которого (за исключением первого) сопоставлены необязательные признаки, входящие в безусловный тест, ребрам – значения этих признаков, а каждой вершине (за исключением корня и концевых вершин) сопоставлена пара (L, j), где L – реакция матрицы описаний, j – номер признака. Корню дерева сопоставляется пара (S, G), где S – совокупность всех строк матрицы описаний, G – совокупность обязательных признаков. Ребрам первого яруса дерева сопоставляется множество возможных значений обязательных признаков. Каждая реакция, соответствующая вершине дере-

ва, однозначно определяется значениями признаков на пути к этой вершине. Реакция, сопоставленная каждой концевой вершине, содержит множество строк матрицы описаний, описывающих объекты обучающей выборки, принадлежащие одному образу. Концевой вершине сопоставляется также номер образа, которому она соответствует.

Также приведем ряд определений и пример безызбыточной матрицы импликаций и выявленных закономерностей.

Для уменьшения степени неполноты смешанного диагностического теста вводится коэффициент разделения признаков.

Коэффициент разделения для всех входящих в соответствующий тест признаков вычисляется на каждом шаге построения дерева. Этот коэффициент отражает степень непересечения подмножеств образов, соответствующих различным значениям признака. Для его вычисления предложена следующая формула:

$$C_j^i = \frac{K_j}{\sum_{l \in T} K_{j,l}}, \qquad (1)$$

где C_j^i – коэффициент разделения i-го признака на j-ом ярусе;

K_j – количество образов, входящих в реакцию предыдущего яруса;

$K_{j,l}$ – количество образов, входящих в реакцию, соответствующую l-му значению j-го признака;

T – множество значений j-го признака.

Коэффициенты разделения не вычисляются для признаков, не входящих в безусловный тест, на основе которого строится СДТ, и для признаков, реакцией на значения которых является текущая реакция (реакция предыдущего яруса).

Алгоритм построения СДТ и принятия решений состоит из следующих шагов.

1. По матрице описаний и матрице различений осуществляются поиск различных закономерностей и построение безусловных (минимальных или безызбыточных) тестов.
2. Для каждого безусловного теста строится дерево СДТ.
3. Производится проход по построенным деревьям от корня к листьям в соответствии со значениями признаков исследуемого объекта. Тест отдает голос за образы, соответствующие листьям, лежащим на концах пути обхода дерева СДТ.
4. Применяется процедура голосования на множестве деревьев и принимается решение о принадлежности исследуемого объекта к тому образу, за который получено наибольшее число голосов.

Алгоритм принятия решений непосредственно в процессе построения СДТ состоит из следующих шагов.

1. По представленной в виде матрицы описаний и матриц различений (3-х типов) обучающей выборке осуществляются поиск различных закономерностей и построение безусловных (минимальных или безызбыточных) тестов.
2. Для каждого безусловного теста производится построение реакций, соответствующих только необходимой части дерева СДТ. Тест отдает голос за образы, соответствующие построенным концевым реакциям.
3. Применяется процедура голосования на множестве тестов и принимается решение о принадлежности исследуемого объекта к тому образу, за который получено наибольшее число голосов.

В третьей главе приводятся алгоритмы построения смешанных диагностических тестов и принятия решений для представления знаний в пространст-

ве троичных признаков.

В целях минимизации длины смешанного диагностического теста в диссертационной работе предлагается следующая модификация формулы (1) , позволяющая учесть равномерность разбиения реакции предыдущего яруса при вычислении коэффициента разделения C_i':

$$C_i' = \frac{T}{K_{i0}^{\lambda} + K_{i1}^{\lambda}}, \qquad (2)$$

где T – нормировочная константа, равная $\frac{K^{\lambda}}{2^{\lambda-1}}$;

λ – целое число, определяемое экспериментально и зависящее от базы знаний, задающее вклад построенных реакций при неравномерном разделении реакции предыдущего яруса;

K – количество образов в предыдущем ярусе;

K_{i0}, K_{i1} – количество образов в реакциях на нулевое (равное 0) и единичное (равное 1) значение *i*-го признака, соответственно.

В отличие от формулы 1, формула 2 позволяет сократить количество запросов за счет выбора признаков, разбивающих реакцию предыдущего яруса на более равномерное количество образов.

Формула 2 позволяет учитывать как степень непересечения подмножеств образов, так и равномерность разбиения реакции предыдущего яруса. При коэффициенте λ=1 получим формулу 1, учитывающую только непересечения; при росте λ увеличивается вклад равномерности разбиения реакции предыдущего яруса.

Обоснованием введения модифицированной формулы служит следующее утверждение

Утверждение 1. Длина дерева СДТ минимальна, если

$$\left|K_{i0} - K_{i1}\right| \le 1,\ K_{i0} + K_{i1} = K\ .$$

Для двоичного и троичного представления данных используется следующий алгоритм построения СДТ на базе построенных безусловных тестов.

1. Из матрицы описаний удаляются столбцы, соответствующие константным, неинформативным, несущественным и зависимым признакам.

2. Выделяется множество обязательных признаков, строки матрицы Q разбиваются на множество реакций по значениям обязательных признаков.

3. Используется следующая рекурсивная процедура по каждому из безусловных тестов.

Пусть L – реакция матрицы описаний, сопоставленная вершине j, находящейся на i-м ярусе. Если L состоит из множества строк, описывающих один образ, то вершина j – концевая. Иначе, вычисляются коэффициенты разделения признаков, производится выбор из безусловного теста следующего признака с максимальным коэффициентом разделения и разбиение L на реакции относительно значений выбранного признака на $(i+1)$-м ярусе.

Во второй части алгоритма осуществляется непосредственно принятие решения по построенным деревьям путем прохода от корня дерева к одной или нескольким концевым вершинам по следующему алгоритму.

1. Формируется: список голосов (G), отданных за каждый из образов (первоначально устанавливаются нулевые значения для элементов списка); список деревьев СДТ (T) (в начальный момент в него входят все построенные деревья); список указателей на вершины деревьев тестов (P) (каждый указатель

из P содержит ссылку на вершину соответствующего дерева теста и величину голоса, отдаваемого после принятия решения с использованием этого указателя); список указателей на концевые вершины деревьев (F) (в начальный момент список пуст).

2. Производится запрос значений всех обязательных признаков.

3. Производится выбор первого дерева из списка деревьев смешанных диагностических тестов. Создаются и заносятся в список P указатели на вершины первого яруса, каждая из которых соответствует совокупности введенных значений всех обязательных признаков. Величина голоса, соответствующая каждому указателю, вычисляется по следующей формуле: $g = (2^N)^{-1}$, где N –количество реакций на значения обязательных признаков.

4. Производится выбор первого указателя (p) из списка P.

5. Запрашиваются значения признака, соответствующего указанной в p вершине.

6. Если значение признака исследуемого объекта нельзя определить (неизвестно), то указатель перемещается на вершину (v) следующего яруса, соответствующую нулевому значению запрашиваемого признака; величина голоса уменьшается вдвое; создается и заносится в список P указатель на вершину дерева теста, соответствующую единичному значению; вес голоса созданного указателя устанавливается равным половине голоса указателя на входе. Иначе, p перемещается на соответствующую вершину следующего яруса.

7. Если вершина v концевая, то указатель на нее заносится в список F и удаляется p из списка P.

8. Если список P не пуст, то переход к пункту 7. Иначе, голос, отданный за все образы, соответствующие входящим в список F вершинам, увеличивается на величину $g = W/m$, где W – вес теста, m – количество образов, соответствующих вершинам, входящим в список F. Из списка T удаляется рассмотренное дерево и, если список T не пуст, то переход к пункту 3, иначе, конец.

На рис. 2 приводится пример принятия решений для матриц описаний и различений, представленных на рис. 1. В безусловный тест входят признаки {1,2,3,4,6,8}.

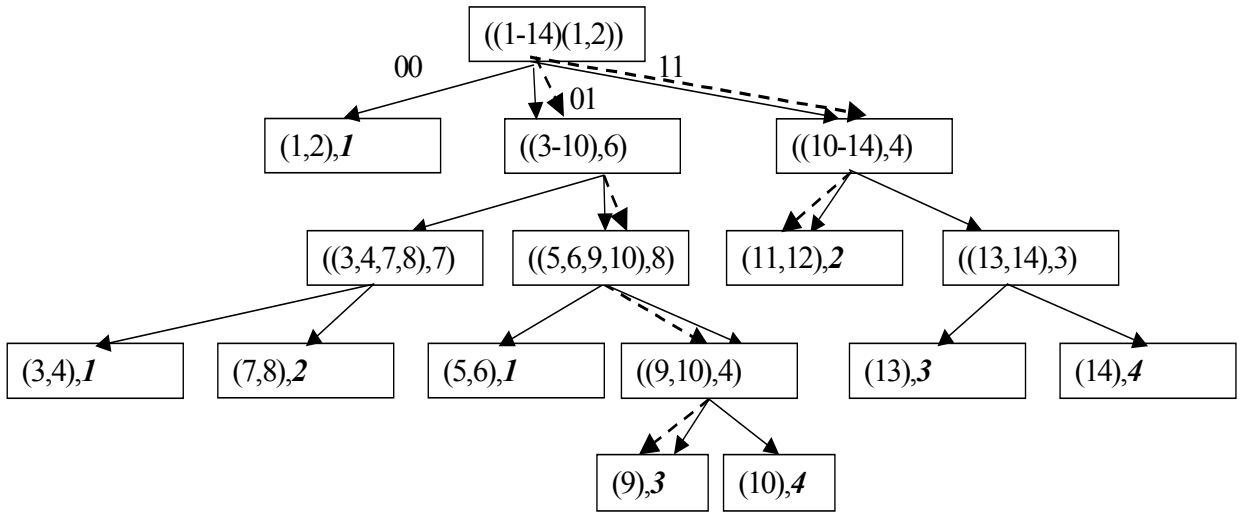

Рис. 2. Дерево смешанного диагностического теста

В разработанном в диссертационной работе алгоритме принятия решения непосредственно в процессе построения смешанного диагностического теста принятие решения осуществляется по следующему алгоритму.

1. Запрашиваются у пользователя значения всех обязательных признаков (процедура выполняется единожды при принятии решения по первому безусловному тесту).

2. Производится построение реакции матрицы описаний на множество обязательных признаков.

3. Упорядочивается множество столбцов матрицы описаний по неубыванию коэффициентов разделения.

4. Производится выбор признака с максимальным коэффициентом разделения и не нулевым весовым коэффициентом из безусловного диагностического теста (в случае, если существует несколько признаков с равными коэффициентами разделения, производится выбор признака с максимальным весовым коэффициентом).

5. Если значение выбранного признака известно (равно 0 или 1), то производится построение реакции с учетом значения этого признака. После этого производится удаление признака из множества признаков, участвующих в принятии решения, и замена исходной реакции полученной. Если значение признака не удается определить, то весовой коэффициент этого признака приравнивается 0 и производится пересчет коэффициентов разделения.

6. Пункты 4–5 выполняются до тех пор, пока во множестве признаков, участвующих в принятии решения, остаются признаки, весовые коэффициенты которых не равны 0, или пока не получим реакцию, состоящую из одного образа. При получении реакции, состоящей из одного образа, производится увеличение голоса, отданного за этот образ, на величину, равную весу теста, и переход к следующему безусловному тесту.

7. В случае если у исследуемого объекта часть признаков не была определена, они доопределяются следующим образом. Создается группа из 2^k (где k – количество неопределенных признаков) объектов, у которых все признаки определены. После этого производится принятие решения относительно каждого объекта из полученной группы, но величина условного голоса за каждый образ уменьшается в 2^k раз.

В примере, приведенном на рис. 2, путь прохода по дереву, необходимый для принятия решения, показан пунктирными стрелками.

Также в третьей главе производится сравнение предложенных алгоритмов с методами и алгоритмами построения логических тестов, описанными в первой главе диссертации.

В четвертой главе приводятся алгоритмы построения матрицы импликаций и алгоритмы построения смешанных диагностических тестов и принятия решений для матричного представления знаний в пространстве целочисленных признаков, заданных в константном и интервальном виде.

Для целочисленного матричного представления данных и знаний с использованием константного задания значений признаков алгоритм построения безусловных диагностических тестов состоит из следующих шагов:

1. Построение двоичной безызбыточной матрицы импликаций с одновременным вычислением весовых коэффициентов признаков на основе целочисленной матрицы описаний.

2. Поиск кратчайших столбцовых покрытий двоичной матрицы импликаций с одновременным выявлением закономерностей (алгоритм Янковской А.Е., Гедике А.И.)

Алгоритм построения безызбыточной матрицы импликаций состоит из следующих шагов.

1. Строки матрицы описаний группируются по принадлежности образам (строки, соответствующие одному образу, располагаются последовательно). В качестве метода группировки используется сортировка строк матрицы описаний по номерам образов. Формируется вещественная вектор-строка (B), элементами которой являются весовые коэффициенты признаков, используемых для построения матрицы импликаций. В начальный момент все компоненты этой строки равны 0.

2. Производится сравнение всех пар строк матрицы описаний, относящихся к разным образам. Для каждой пары строк производится построение двоичной строки матрицы импликаций и корректировка вещественной строки (B) по следующим правилам:

2.1. Если значения признака у сравниваемых строк совпадают, то соответствующему элементу формируемой двоичной строки присваивается значение "0", иначе, присваивается значение "1".

Добавление к каждому элементу строки весов признаков величины $\left|q_j - q_k\right|$:

$$\beta_i := \beta_i + \left|q_j - q_k\right|,$$

где q_j – значение *i*-го признака *j*-ой строки;

q_k – значение *i*-го признака *k*-ой строки;

β_i – вес *i*-го признака;

3. Производится проверка на избыточность (на поглощение) построенной строки с уже построенными строками матрицы.

3.1. Если в матрице импликаций есть строки, поглощающие построенную строку, то они удаляются, а построенная строка добавляется в матрицу импликаций.

3.2. Если построенная строка поглощает какую-либо строку матрицы импликаций, то она в матрицу импликаций не добавляется.

3.3. Если построенная строка не поглощается другими строками матрицы импликаций и не поглощает ни одну из уже построенных строк, то она добавляется в матрицу импликаций.

3.4. Если рассмотрены все пары строк матрицы импликаций, то переход к пункту 4, иначе, производится сравнение следующей пары строк.

4. Производится нормировка элементов вектор-строки (B) по формуле:

$$\beta_i = \frac{\beta_i}{\alpha_i \sum_{j=1}^{d} \sum_{k=j+1}^{d} \Phi_j \Phi_k},$$

где α_i – максимальный интервал признака;

i – номер признака;

j,k – номера строк $j,k \in \{1,...,d\}$, где *d* количество образов;

Φ_r – количество объектов, принадлежащих *r*-у образу, где $r \in \{j,k\}$.

Двоичную матрицу импликаций допустимо использовать для построения безусловных целочисленных тестов, так как вся необходимая информация о различии значений признаков и их важности содержится в весовых коэффициентах признаков.

Для целочисленного матричного представления данных и знаний с использованием интервального задания значений признаков алгоритм построения

безызбыточной матрицы импликаций отличается от алгоритма построения матрицы импликаций для целочисленного матричного представления данных и знаний с использованием константного задания значений признаков в следующих деталях.

1. Вычисляется вес каждой строки матрицы описаний по формуле:

$$\Phi_j = \prod_{i=0}^{n} \varepsilon_{ij} ,$$

где i – номер признака;

j – номер строки матрицы описаний $j \in \{1,...,n\}$ (n – количество строк матрицы описаний);

$\varepsilon_{i,j}$ – величина интервала значений признака объекта, представленного i-й строкой матрицы Q.

2. При вычислении весов признаков к каждой компоненте вектор-строки весов признаков добавляется следующая величина:

$$\beta_i := \beta_i + \frac{(\delta_{jk}^{1} + \delta_{jk}^{2``})\varepsilon_{ij}\varepsilon_{ik}}{2} * \Phi_j \Phi_k ,$$

где δ_{jk}^{1} – разница между минимальными значениями интервалов признака у сравниваемых строк;

δ_{jk}^{2} – разница между максимальными значениями интервалов признака у сравниваемых строк;

j, k – номера сравниваемых строк $j,k \in \{1,...,n\}$ (n – количество строк матрицы описаний).

3. Нормировка элементов вектор-строки весов признаков производится по следующей формуле:

$$\beta_i = \frac{\beta_i}{\alpha_i \sum_{j=1}^{n} \sum_{k=j+1}^{n} \Phi_j \Phi_k} ,$$

где α_i – максимальный интервал значений i-го признака;

i – номер признака;

j, k – номера строк $j,k \in \{1,...,n\}$ (n – количество строк матрицы описаний);

Φ_j – вес j-ой строки;

β_i – вес i-го признака.

Также приводятся алгоритмы построения СДТ и принятия решений для матричного представления знаний в пространстве целочисленных признаков, заданных как в константном, так и в интервальном виде.

В пятой главе приводятся сравнительные характеристики алгоритмов построения СДТ и принятия решений и результаты их экспериментальных испытаний в прикладных интеллектуальных системах.

Для иллюстрации работы алгоритмов построения СДТ и принятия решений был выбран алгоритм принятия решений непосредственно в процессе построения СДТ. Исследование одного из представленных в диссертации алгоритмов построения смешанных тестов и принятия решений на его основе позволяет оценить эффективность остальных алгоритмов. Для доказательства этого в диссертационной работе предлагается следующее утверждение.

Утверждение 2. При принятии решений, осуществляемом при непосредственном построении СДТ, и принятии решения по предварительно построен-

ным СДТ на основе безусловных диагностических тестов количество и последовательность запрашиваемых признаков совпадают.

Исследования проводились на следующих интеллектуальных системах, созданных на основе интеллектуальной системы ЭКСАПРАС, разработанной в лаборатории интеллектуальных систем ТГАСУ под руководством Янковской А.Е.

1. ГЕНЕТОН – для оценки состояния здоровья населения по генетическим и функциональным показателям, включая определение уровня эритроцитов с микроядрами, стабильности генома и уровня риска патологических изменений, а также выработку рекомендаций по диспансерному наблюдению и применению препаратов с антимутагенным эффектом.

Для испытаний была выбрана база знаний "Генетические изменения (оценка состояния здоровья ребенка по генетическим и функциональным показателям)".

2. ЭСМЕДИЧ – для принятия оптимальных решений в области медицины чрезвычайных ситуаций, включая постановку диагноза и определение комплекса лечебно-профилактических и сортировочно-эвакуационных мероприятий при оказании медицинской помощи пострадавшим с травмами груди, от лучевых поражений и отравлений сильно действующими ядовитыми веществами (СДЯВ), а также принятие организационно-управленческих решений в очаге поражения.

Для испытаний были выбраны две базы знаний:

а) "Квалифицированная и специализированная помощь при лучевых поражениях";

б) "Организационно-управленческие решения при чрезвычайных ситуациях".

3. Интеллектуальная система оценки научно-технических проектов на предмет целесообразности выделения грантов, а также очередности и объемов их финансирования.

В создании баз знаний для прикладных интеллектуальных систем, на основе которых проводились экспериментальные исследования, представленных в диссертации алгоритмов, принимали участие высококвалифицированные эксперты: Ильинских Н.Н., Черногорюк Г.Э., Бурматнов В.Ф. Базы знаний, применяемые в прикладных интеллектуальных системах, основанных на интеллектуальном инструментальном средстве ЭКСАПРАС, конвертированы для создания прикладных интеллектуальных систем на основе ИИС ИМЛОГ.

Выбранный для исследования алгоритм реализован в ИИС ИМЛОГ.

Для иллюстрации работы алгоритмов принятия решений на основе СДТ приводится пример и таблица, характеризующая порядок и количество запрошенных признаков при принятии решения в базе "Генетические изменения". Для приведенного в примере исследуемого объекта, отношение количества запрашиваемых признаков в процессе принятия решения с использованием безусловных тестов к количеству запрашиваемых признаков при принятии решения с использованием смешанных тестов для данного объекта составило (38:21). При принятии решений на основе построения смешанного диагностического теста без выделения безусловной составляющей аналогичное отношение равно (38:30).

При тестировании алгоритма принятия решений на основе смешанных диагностических тестов было сгенерировано 1000 описаний случайных исследуемых объектов и проведено принятие решений для каждого из них. Исследования проводились на базах знаний, использующих для принятия решений множество безусловных тестов, мощность которого больше 1. Для тестирова-

ния брались реальные базы знаний прикладных интеллектуальных систем, разработанных на основе системы ЭКСАПРАС. Результаты тестирования представлены в таблице 1, в которой приняты следующие обозначения:

N – количество тестов;

M – количество обязательных признаков;

$|T|$ – количество характеристических признаков, входящих в каждый минимальный безусловный тест;

$|\mathrm{U}(T)|$ – количество признаков, вошедших в объединение множеств признаков, входящих в тесты;

$A_{\text{с},1}$ – среднее количество признаков, необходимое для принятия решений с использованием смешанных диагностических тестов, для каждого теста;

$A_{\text{с},2}$ – среднее количество признаков, необходимое для принятия решений с использованием только условной компоненты смешанных диагностических тестов (без выделения безусловной составляющей), для каждого теста;

$A_{\text{о},1}$ – общее количество признаков, запрошенных для принятия решения по каждому объек3ту (берется средняя величина) с использованием смешанных диагностических тестов;

$A_{\text{о},2}$ – общее количество признаков, запрошенных для принятия решения по каждому объекту (берется средняя величина) с использованием только условной компоненты смешанных диагностических тестов (без выделения безусловной составляющей), для каждого теста.

Таблица 1

Результаты тестирования алгоритма принятия решений непосредственно в процессе построения смешанного диагностического теста на различных базах знаний

Название базы	N	M	$\|T\|$	$\|\mathrm{U}(T)\|$	$A_{\text{с},1}$	$A_{\text{с},2}$	$A_{\text{о},1}$	$A_{\text{о},2}$
Генетические изменения	146	3	10	38	5,29	5,01	21,39	30,1
Квалифицированная и специализированная помощь при лучевых поражениях	41	0	6	13	3,02	3,02	9,25	9,25
Организационно-управленческие решения при чрезвычайных ситуациях	586	9	20	38	12,91	11,45	19,04	24,06
Оценка научно-технических проектов	5	16	19	22	9,00	6,01	9	11,35

На основе проведенных экспериментальных исследований были выявлены следующие условия, позволяющие оптимизировать количество и порядок запрошенных признаков.

1. Больший выигрыш достигается, если количество признаков, входящих в объединение безусловных тестов, существенно больше (желательно в несколько раз) среднего количества признаков, входящих в безусловный тест (в этом случае при принятии решения отсекается большее количество признаков).

2. Количество запрашиваемых признаков может быть существенно

уменьшено в случае, если количество признаков, входящих в безусловную составляющую смешанного теста, составляет от трети до половины от количества признаков, входящих в тест. В случае, если количество признаков, входящих в безусловную компоненту, меньше трети, то реакция на их значения будет достаточно велика и количество запрошенных признаков не будет существенно отличаться от количества признаков, запрошенных при принятии решения с использованием только условной компоненты.

3. При увеличении количества признаков, входящих в безусловную составляющую, появляется вероятность запроса у пользователя "лишнего" (не участвующего в принятии решения) признака.

4. Скорость работы алгоритмов существенно зависит только от количества смешанных тестов, на основе которых принимается решение.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработан и программно реализован в системе ИМСЛОГ алгоритм принятия решения непосредственно в процессе построения смешанных диагностических тестов на основе безусловных для представления данных и знаний в пространстве троичных признаков.

2. Разработаны алгоритмы построения СДТ на базе безусловных тестов и принятия решений на их основе для представления данных и знаний в пространстве троичных, номинальных и целочисленных признаков.

3. Создан алгоритм принятия решения непосредственно в процессе построения СДТ на основе безусловных при представлении данных и знаний в пространстве троичных, номинальных и целочисленных признаков.

4. Выведены две формулы коэффициента разделения, позволяющие строить деревья смешанных диагностических тестов меньшей длины.

5. Созданы алгоритмы построения двоичной безызбыточной матрицы импликаций и вычисления весовых коэффициентов признаков при матричном представлении данных и знаний в пространстве целочисленных и номинальных признаков (для константного и интервального задания значений признаков).

6. Проведено тестирование алгоритма принятия решений непосредственно в процессе построения смешанного диагностического теста на 4 реальных базах данных и знаний.

7. Реализованные алгоритмы внедрены в Сибирском государственном медицинском университете и в Томском медико-генетическом центре "ГЕНЕТИК".

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Янковская А.Е., Гедике А.И., Аметов Р.В., Блейхер А.М., Гергет О.М., Муратова Е.А., Кузоваткин А.Н. Интеллектуальная система ИМСЛОГ // Новые информационные технологии в исследовании дискретных структур. Доклады 3-й Всероссийской конференции с международным участием. – Томск: Изд-во СО РАН, 2000. – С. 169–175.

2. Янковская А.Е. Кузоваткин А.Н. Принятие решений на базе построения целочисленных смешанных диагностических тестов // Научная сессия МИФИ–2002. Сборник научных трудов, Т. 3. – М: Изд-во МИФИ, 2002. – С. 53–54.

3. Янковская А.Е., Гедике А.И., Аметов Р.В., Кузоваткин А.Н. Принятие и обоснование решений в комплексе интеллектуальных систем // Интеллектуализация обработки информации (ИОИ–2002). Тезисы докладов международной

научной конференции. – Симферополь: Изд-во Крымского научного центра, 2002. – С. 97–98.

4. Янковская А.Е., Кузоваткин А.Н. Построение логических тестов в пространстве k-значных и номинальных признаков в системе распознавания образов // Интеллектуализация обработки информации (ИОИ–2002). Тезисы докладов международной научной конференции. – Симферополь: Изд-во Крымского научного центра, 2002. – С. 98–100.

5. Янковская А.Е., Гедике А.И., Аметов Р.В., Кузоваткин А.Н. Технология представления и обработки знаний на базе инструментального средства ИМСЛОГ–2002 // Компьютерная лингвистика и интеллектуальные технологии (Диалог–2002). Труды международного семинара. – Т. 2. – М.: Наука, 2002. – С. 555–567.

6. Янковская А.Е., Кузоваткин А.Н. Алгоритмы построения логических тестов в пространстве k-значных и номинальных признаков в системе распознавания образов // Искусственный интеллект. – Украина, Донецк: ІПШІ "Наука і освіта". – 2002. – № 2. – С 330–337.

7. Кузоваткин А.Н., Янковская А.Е. Интеллектуальная система принятия решений на основе смешанных диагностических тестов // Известия Таганрогского государственного радиотехнического университета. Интеллектуальные САПР. Материалы Междунар. научно-технической конф. – 2002. – № 3 (26). – С. 5–11.

8. Янковская А.Е., Кузоваткин А.Н. К вопросу минимизации длины смешанного диагностического теста для интеллектуальных распознающих систем // Вторая Международная конференция по проблемам управления. Тез. докладов. – Т. 1. – М.: Институт проблем управления, 2003. – С. 158.

9. Янковская А.Е., Кузоваткин А.Н. Принятие решений в интеллектуальном инструментальном средстве ИМСЛОГ–2002 на основе смешанных диагностических тестов // Информационные системы и технологии (ИСТ–2003). Материалы междунар. конференции. – Т. 3. – Новосибирск: Изд-во НГТУ, 2003. – С. 182–186.

10. Yankovskaya A.E., KuzovatkinA.N. Approbation of constructing algorithms of mixed diagnostic tests and of real knowledge bases decision-making based on matrix model of knowledge representation // Proceedings of 6th International Congress on Mathematical Modeling. – Nizhni Novgorod, 2004. – С. 96.

11. Янковская А.Е., Ильинских Н.Н., Черногорюк Г.Э., Кузоваткин А.Н. Результаты исследований алгоритма принятия решений непосредственно в процессе построения смешанных диагностических тестов, реализованного в интеллектуальных биомедицинских системах // Естествознание и гуманизм. Сб. научных трудов. – Т. 1, № 3. – Томск: Издание СибГМУ, 2004. – С. 95–99.