УДК 519.816:58.002

На правах рукописи

Красинский Виталий Израилевич

ДИАГНОСТИКА ОБЪЕКТОВ, ХАРАКТЕРИЗУЮЩИХСЯ РАЗНОТИПНЫМИ ПРИЗНАКАМИ, ПО ОТНОШЕНИЮ К ПЕРЕСЕКАЮЩИМСЯ КЛАССАМ

Специальность 05.13.18 – математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск - 2002

Работа выполнена в Центральном сибирском ботаническом саду
Сибирского отделения Российской академии наук

Научный руководитель: доктор технических наук,
профессор В.З.Манусов

Научный консультант: заслуженный деятель науки РФ,
доктор биологических наук,
профессор И М.Красноборов

Официальные оппоненты: доктор технических наук,
профессор В.В Губарев,
Новосибирский государственный
технический университет

кандидат технических наук,
старший научный сотрудник А.Л.Осипов,
Новосибирский государственный университет

Ведущая организация: Институт систем информатики
им. А.П.Ершова СО РАН, г. Новосибирск

Защита диссертации состоится " 13 " июня 2002 года в 15 часов
на заседании диссертационного совета К 003.49 01
в Новосибирском институте органической химии
им. Н.Н.Ворожцова СО РАН
по адресу пр акад Лаврентьева, 9 г Новосибирск, 630090, Россия

С диссертацией можно ознакомиться в библиотеке Новосибирского института органической химии им. Н.Н.Ворожцова СО РАН

Автореферат разослан " 29 " апреля 2002 года.

Ученый секретарь диссертационного совета
кандидат химических наук М.И.Подгорная

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В последние годы значительные усилия исследователей в области искусственного интеллекта (ИИ) направлены на разработку методов решения задач классификации и распознавания объектов по плохо обусловленной исходной информации. Подобные задачи возникают при обработке зашумленных сигналов с датчиков технологических процессов, результатов социологических опросов, прогнозировании в геологии, диагностике в биологии.

Затрудняющим условием распознавания объектов зачастую является пересечение классов объектов по всем количественным и качественным признакам, вследствие чего неприменимы вероятностно-статистические методы анализа информации. В ботанической науке полиморфизм таксономических единиц по признакам растений является одной из основных проблемных ситуаций при диагностике особей.

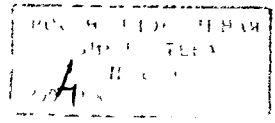
В настоящее время диагностика растений осуществляется по книжным определителям, составленным ведущими флористами "вручную", на основе своего опыта и литературных данных. Эти определители построены по типу двоичных вопросников относительно значений признаков и жестко привязывают пользователя к ветвям процесса диагноза, заложенным автором определителя. Не предусматривается ответ "не знаю" на вопросы о значениях признаков, что часто заводит в тупик диагностику растения. Авторы ручных определителей не в состоянии учесть все пересечения ветвей решений и комбинации значений признаков растений, поэтому процесс диагностики особей по таким определителям ненадежен.

Необходимость создания надежных определителей растений признается в мировом ботаническом сообществе.

Предметом исследования является математическое моделирование процесса диагностики объектов по совокупности их количественных и качественных признаков. Пересечение (многозначность) таксонов-классов объектов в разнотипном признаковом пространстве является заданной неопределенностью, что не позволяет применять методы статистического анализа, основанные на аксиоматической теории вероятностей.

Целями диссертационной работы являлись:

1. Разработка способов преобразования числовых и номинальных признаков многозначных объектов в значения функций принадлежности (ФП) к соответствующим нечетким множествам (НМ).
2. Выбор способов суперпозиции всех НМ для вычисления интегрального критерия ранжирования многозначных (списковых) объектов по совокупности разнотипных признаков.
3. Выбор критериев ранжирования разнотипных признаков многозначных объектов по степени априорной нечеткой диагностической способности признаков.



4. Создание алгоритма, минимизирующего число задаваемых вопросов о значениях разнотипных признаков диагностируемой особи (предмета).
5. Программирование, отладка и тестирование диалоговой программы-диагноста растений ©RECOFAM (Recognizer of Families).
6. Создание варианта общего алгоритма для предсказания ошибок в содержании документов базы данных (БД) гербарных этикеток.

Методом решения вышеназванных задач в диссертационной работе избрана теория нечетких множеств и связанная с ней теория возможностей, ввиду описанных особенностей исходных данных. Эти теории – один из разделов искусственного интеллекта. Общая теория систем, теория измерений, теория голосования также использовались в работе.

Достоверность результатов обосновывается тем, что при разработке алгоритма диагностики использовались методы теории НМ и теории возможностей, методы теории измерений, метод парных сравнений и методы обработки результатов голосования. Все эти методы хорошо себя зарекомендовали при решении практических задач в различных прикладных областях. Программа ©RECOFAM надежно работает с реальной исходной флористической информацией – быстро и безошибочно диагностированы все предъявленные специалистами-ботаниками растения (гербарные образцы). Верное на 76% предсказание ошибок в содержании документов одного из разделов БД гербарных этикеток также подтвердило надежность разработанного метода диагностики.

Научная новизна работы состоит в следующем:
1. Разработан и реализован в программе для ЭВМ оптимизирующий алгоритм диагностики объектов по совокупности разнотипных признаков, в условиях пересекающихся классов объектов.
2. Разработаны два алгоритмических (без участия человека-эксперта) способа вычисления ФП списковых объектов к НМ – для номинального признака на основе функций доверия Шефера[1], и для числового признака по фокальным элементам (ФЭ) слева и справа.
   В отличие от типичного применения постоянных ФП для всей задачи, в предложенных способах осуществляется динамическое перевычисление ФП на переменном числе объектов.
3. Разработан способ проверки правильности индексирования ключевыми словами документов базы данных.
4. Для оценки степени многозначности совокупности объектов по разным признакам предложен коэффициент эксцесса как средняя информационная избыточность описания объектов.

Теоретическая и практическая значимость проведенного исследования состоят в том, что продемонстрирован метод анализа нечеткой информации, применимый к широкому классу задач в различных прикладных областях. Например, многозначные числовые объекты могут отображать

технологические параметры в сомнительных ситуациях, сбойную работу электронных схем. Многозначные номинальные объекты могут моделировать качественные оценки промышленной продукции, состояния элементов экономики, символьные последовательности, возникающие при формализации естественных и абстрактных текстов (кодов). Ранжирование нечетких объектов по совокупности разнотипных признаков дает возможность классифицировать, в соответствии с семантикой нечеткости, реальные процессы, то есть извлекать знания из нечетких фактов. Эти знания могут использоваться как элементы экспертных систем, обучающих программ, прецеденты в нейро-ЭВМ.

Диалоговый характер алгоритма диагностики неизвестного объекта хорошо соответствует целям обучения. Имеются акты внедрения программы ©RECOFAM в Новосибирском и Омском государственных педагогических университетах на кафедрах ботаники. Меняя сценарии диалога и содержательное наполнение таблицы экспериментальных данных (ТЭД), можно использовать алгоритм в разных учебных курсах.

Представляемый алгоритм диагностики, описанный в [17], [19], может быть реализован в микро-ЭВМ как портативный вычислитель в качестве "интеллектуального советника" в неопределенных ситуациях.

Часть диссертационного исследования по созданию электронной БД гербария NS ЦСБС СО РАН и верификации ее содержимого входит как основной результат по проекту РФФИ 94-07-11932 за 1994-1996 годы.

Апробация работы. Основные результаты диссертации представлялись на Всероссийских и международных конференциях, семинарах:

Всероссийской конференции Петровской Академии наук и искусств "Наука и технологии XXI века", институт математики СО РАН, Новосибирск, октябрь 1998. Расширенном ученом совете ЦСБС СО РАН, Новосибирск, апрель 1999. Международной конференции KORUS'99, Новосибирск, НГТУ, июнь 1999. Семинаре на кафедре ВТ НГТУ, Новосибирск, декабрь 1999. Семинаре в Новосибирском филиале РосНИИ искусственного интеллекта, Новосибирск, март 2000. V рабочем совещании СО РАН и ИСИ СО РАН по электронным публикациям с участием иностранных ученых EL-PUB-2000, Новосибирск, июнь 2000. Два доклада на четвертом Сибирском конгрессе по прикладной и индустриальной математике (ИНПРИМ-2000), институт математики СО РАН, Новосибирск, июнь 2000. Три доклада на Российском совещании "Проблемы создания ботанических БД", ЦСБС СО РАН, Новосибирск, октябрь, 2000. Международной конференции "Информационные системы и технологии" ИСТ'2000, Новосибирск, НГТУ, ноябрь 2000. Семинаре на кафедре СЭСП НГТУ, Новосибирск, январь 2001. Семинаре в институте автоматики и электрометрии СО РАН, Новосибирск, февраль 2001. Всероссийской конференции, посвященной 90-летию А.А.Ляпунова, СО РАН, октябрь, 2001.

Публикации. Основные результаты исследования опубликованы в 19 работах. При этом в соавторских статьях диссертанту полностью принадлежат решения всех математических вопросов создания алгоритма диагностики объектов, моделирования исходной экспертной информации, создание программного обеспечения управления БД, программы-диагноста ©RECOFAM, проведение расчетов и обработка результатов; совместное с ботаниками участие в разработке структур БД, тестирование программ, верификация содержимого БД, обсуждение результатов.

На защиту выносятся следующие положения:

- Способы вычисления функций принадлежности списковых объектов к НМ по числовому и номинальному признакам.
- Способ нормализации нечетких множеств списковых объектов.
- Диалоговый алгоритм диагностики многомерных объектов, минимизирующий число вопросов о значениях разнотипных признаков.
- Программа для ЭВМ ©RECOFAM, реализующая алгоритм диагностики.
- Способ предсказания ошибок индексирования ключевыми словами документов баз данных
- Коэффициент эксцесса для оценки средней степени многозначности объектов по разнотипным признакам.

Структура и объем диссертации. Работа состоит из введения, пяти глав, заключения, списка литературы из 154 наименований, приложения с таблицами исходных данных о семействах двудольных растений Сибири, и двумя актами внедрения результатов работы.
Содержание диссертации изложено на 151 страницах печатного текста, включая два рисунка и 22 таблицы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность выбранной темы. Представлен обзор литературы, из которого следует, что отсутствует решение задачи в постановке, как в названии диссертации. Автором просмотрены новые поступления книг и журналов библиотеки ГПНТБ СО РАН и электронные рефераты журналов Biometrics, Biometrika, Mathematical Biosciences, Fuzzy Sets and Systems, Soft Computing за последние 5 лет.

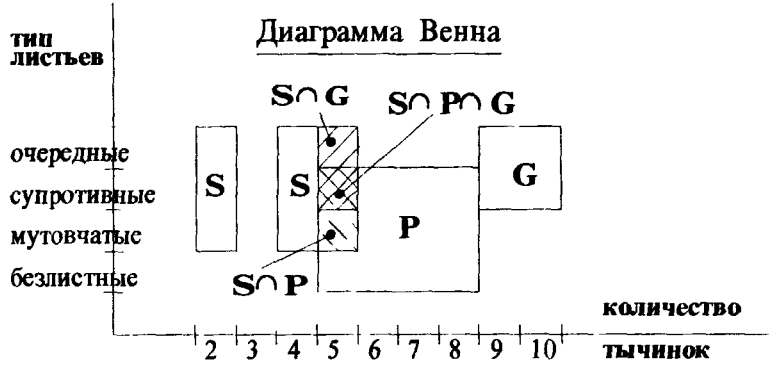
Сформулированы цели исследования. Описывается сложная исходная информация о 102 многозначных таксонах-семействах двудольных растений Сибири по 11 разнотипным признакам. Признаки следующие:
4 количественных – числа лепестков околоцветника, тычинок (андроцей), столбиков, пестиков; 7 номинальных – типы плодов, околоцветника, размножения, соцветия, листьев, завязи, гинецея.

Многозначность таксонов по каждому признаку соответствует пересечениям этих классов. Например, по признаку "тип соцветия" семейство растений Fabaceae характеризуется списком трех значений: головка, кисть,

цветки одиночные. По этому же признаку семейство Primulaceae имеет список из пяти значений: зонтик, кисть, метелка, цветки одиночные, цветки пазушные. Видно, что эти таксоны частично пересекаются (похожи) по признаку типа соцветия. В табл. 1 приведен фрагмент описания трех классов-семейств по двум разнотипным признакам. Ниже показана соответствующая диаграмма Венна.

Таблица 1

Классы	Тип листьев	Количество тычинок
Geraniaceae	очередные, супротивные	5, 9, 10
Primulaceae	безлистные, мутовчатые, супротивные	5, 6, 7, 8
Scrophulariaceae	мутовчатые, очередные, супротивные	2, 4, 5



Еще пример по числовому признаку: у растений семейства Brassicaceae число тычинок цветка бывает 2, 4, 6, – отсюда видна недостаточность интервального способа учета неопределенности для реальных биологических объектов представители названного таксона не могут иметь число тычинок цветка 3, 5.

Многозначность описания классов можно трактовать как информационную избыточность. Для оценки средней степени этой избыточности предлагается коэффициент эксцесса:

ex = $\sum x_{mni}$ / (M·N). Здесь i – индекс значения признака многозначного таксона; m, n – индексы признака и класса; M - число признаков; N - число классов; x_{mni} – значение признака класса-таксона в двоичной матрице инцидентности объект-свойство. Коэффициент ex равен единице в случае не пересекающихся по значениям признаков классов объектов.

Он позволяет приближенно оценить среднюю степень неопределенности описания классов как по каждому признаку в отдельности, так и в целом, по всем признакам. Будучи вычисленным по всей двоичной ТЭД (102*94), этот коэффициент получил значение ex=1.62, которое в точности равно пропорции золотого сечения, одного из фундаментальных законов природы и искусства! Соответствие строения и количества разных органов растений числам Фибоначчи известно ботаникам еще со средних веков.

Поскольку исходная ТЭД является обобщающей по большинству растений Сибири, факт ex=Ф=1.62 можно считать не случайным, а свидетельствующим об удачности выбора ботаниками системы признаков для описания семейств растений. Возможно, в "золотой" степени информационной избыточности ТЭД отражается универсальное антиэнтропийное свойство живых систем для сохранения своей устойчивости во внешнем мире.

Глава 1 посвящена краткому обзору теории нечетких множеств. Теория НМ сравнивается с другими способами учета неопределенности. Известные способы вычисления ФП объектов к НМ показывают, как математически описать в предметной области неопределенность ситуации и ее степень. Для решения поставленной задачи диагностики растений этих способов оказалось недостаточно ввиду многозначности флористических таксонов. Авторские способы вычисления ФП к НМ представлены в главе 2.

Нечеткие меры (доверия, возможности, необходимости) событий есть меры учета неопределенности в ситуациях, когда недостаточна модель полной группы несовместных событий, на которой основана теория вероятностей. Эти меры используются в теории возможностей, а именно, на основе способов Демпстера-Шефера[1,2], Дюбуа и Прада[3] возможен переход к значениям ФП объектов к НМ. Теория возможностей, совместно с авторскими способами вычисления ФП многозначных объектов к НМ, использовалась при разработке алгоритма нового метода диагностики многомерных объектов в неопределенных ситуациях.

Глава 2 содержит описание авторских способов вычисления ФП многозначных объектов к НМ. Семантика нечеткости по всем признакам определена как "ненадежность диагностики растения" чем многозначнее таксон-семейство, тем хуже диагностируются его представители по данному признаку. Такая семантика нечеткости подтверждена ботаниками и поэтому служит цели создания оптимизирующего диалогового алгоритма диагностики растений. Оптимизация состоит в минимизации числа задаваемых вопросов о значениях признаков диагностируемой особи.

Вычисление степени нечеткости многозначных объектов является преобразованием в сильную интервальную шкалу списковых признаков объектов, то есть объекты становятся ранжированными по каждому признаку (количественному и качественному), причем по единой мере, что создает условия для многомерного анализа (классификации).

Раздел 2.1. Применение теории возможностей для вычисления ФП объектов к НМ требует нормальности этих НМ: $\exists x\ \mu_A(x)=1$.

В большинстве случаев для обеспечения этого условия прибегают к нормализации решетчатой ФП максимальным размахом ее значений. Но в задачах многомерного анализа такой подход привел бы к разномасштабной "нормализации", т.е. к изменению соотношений нечеткости в исходных данных. Поэтому автором предложено осуществлять нормализацию всех НМ путем добавления к исходным объектам одного искусственного, имеющего принадлежность ко всем градациям всех признаков [7]. Значения фокальных

элементов для числовых признаков и мер возможностей термов для номинальных признаков изменяются при этом на несколько процентов, в отличие от 20-50 -процентного изменения их в случае нормализации размахом ФП. В итоге, критерий свертки всех НМ вычисляется точнее.

В разделах 2.2, 2.3 представлен новый способ, описанный в [7], вычисления ФП многозначных объектов к НМ по числовому признаку.

Известна мера необходимости пересечения множеств $N(A\cap B)=\min(N(A), N(B))$. Л.Заде предложил меру возможности объединения множеств $\Pi(A\cup B)=\max(\Pi(A), \Pi(B))$.

Меры необходимости и возможности взаимосвязаны:

$\Pi(A)\geq N(A)$; $N(A)=1-\Pi(\neg A)$; $\Pi(A)=1-N(\neg A)$.

При этом $\Pi(A)+\Pi(\neg A)\geq 1$; $N(A)+N(\neg A)\leq 1$.

Г.Шефером[1] доказано, что нижняя P_1 и верхняя P_2 неаддитивные вероятности Демпстера[2] значений признака многозначных объектов равны, соответственно, мерам необходимости и возможности этих значений, вычисляемым по относительным мощностям $m(A_i)=|A_i|/|X|$ вложенных фокальных элементов (подмножеств) универсума $A_1\subseteq A_2\subseteq\cdots\subseteq A_r\subseteq X$.

При этом должно соблюдаться нормирующее условие $\sum m(A_i)=1$.

Другой подход к учету неопределенности числовой информации, вместо вычисления диапазонов вероятностей Демпстера P_1 и P_2 для каждого значения признака, основан на доказательстве Д.Дюбуа и А.Прада[3] значение ФП объекта к НМ может вычисляться по относительным мощностям вложенных ФЭ: $\mu_A(x)=\sum m(A_i)\cdot T_i(x)$, и является необходимой (гарантированной) степенью нечеткости объекта. Видно, что в вычислении значения $\mu_A(x)$ участвуют все возможные значения признака по всем объектам. Здесь $T_i(x)$ есть функция Хевисайда соответствия объекта и квантованного значения признака. В итоге получается интегрированная оценка неопределенности каждого объекта, так как она связана со значениями признака всех многозначных объектов ТЭД.

Автором предложено построить для числового признака две последовательности ФЭ по формулировкам противоположных событий "быть меньше или равно K" и "быть больше K". Параметр K соответствует значениям признака. В случае однозначных объектов эти последовательности сводятся к стандартной гистограмме.

Итак, относительные мощности ФЭ слева обозначим $m_l(K)$, а справа – $m_r(K)$. По способу Дюбуа-Прада по относительным мощностям $m_l(K)$ и $m_r(K)$ двух последовательностей ФЭ получаем два НМ L и R.

---

[1] Shafer G. A Mathematical Theory of Evidence Princeton: Princeton Univ. Press, 1976. -297 p.

[2] Dempster A.P. Upper and lower probabilities induced by a multivalued mapping // Ann. of Math. Statistics, 1967. Vol. 38, pp. 325-339.

[3] Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике. Пер. с фр. – М.. Радио и связь, 1990. – 288 с.

Поскольку эти НМ построены по минимальным мерам необходимости, объединяем их по принципу обобщения Заде в итоговое НМ V: $\mu_V(x)=\mu_{L\cup R}(x)=\max\{\mu_L(x),\mu_R(x)\}$.

В алгоритме диагностики употребляются дополнительные НМ $\mu_F(x)=1-\mu_V(x)$, семантика которых соответствует априорной надежности диагноза неизвестного объекта по значениям числовых признаков.

Таких НМ столько, сколько числовых признаков.

Отметим, что для многозначных объектов, описанных порядковыми (балльными) признаками, тоже можно построить две фокальные последовательности слева и справа. После этого вычисление ФП объектов к НМ можно производить по изложенному выше способу, что значительно расширяет сферу его применения в задачах многомерного анализа.

В разделе 2.4 представлен способ вычисления ФП многозначных объектов к НМ по номинальному признаку. В отличие от числовых, по значениям номинальных признаков невозможно построить вложенные подмножества объектов. Эти объекты в ТЭД являются объектами-списками, а у диагностируемой особи значение признака будет одним элементом из этого списка, т.е. реализуется одно из нескольких возможных событий. Поэтому вычисление ФП объектов к НМ по номинальным признакам производится на основе мер возможностей значений признаков. Известен подход Г.Шефера[1], развитый в работах Д.Дюбуа, А.Прада и G.Resconi[4], в котором требование полноты группы несовместных событий заменяется распределением единичной "массы уверенности" на все возможные события. Мы используем этот подход для вычисления ФП многозначных таксонов-объектов к НМ.

А именно, для номинального признака сумма единиц в двоичной матрице инцидентности $\|x_{ni}\|$ "объект – значения признака" (n – индекс объекта, i – индекс значения признака в ТЭД) рассматривается как "масса размытости всех значений признака по всем объектам". Распределение этой массы по значениям C_i номинального признака соответствует распределению меры возможности конкретного значения признака (нечеткого события) в смысле Шефера: $q(C_i) = \sum_n x_{ni} / \sum_{n,i} x_{ni}$.

В соответствии с доказательством Дюбуа и Прада, вычисление ФП объектов универсума X к нормальному НМ по мерам возможности событий (значений признака) производится так: $\forall$ x, $\mu_A(x)=\sum q(C_i)\cdot T_i(x)$, где $T_i(x)$ есть характеристическая функция соответствия объекта и значения признака (элементы матрицы $\|x_{ni}\|$ по ТЭД). Последнее выражение для $\mu_A(x)$ означает переход от модальной логики $T_i(x)$ двоичных исходных данных к многозначной логике (нечеткому множеству, измеренному в сильной числовой интервальной шкале) посредством весовой функции $q(C_i)$, соответствующей постулату Шефера о единичной "массе уверенности", распределенной на универсуме: $\sum q(C_i)=1$.

В алгоритме диагностики используются дополнительные НМ $\mu_{\neg A}(x)$, семантика которых соответствует степени детерминированности классов-таксонов в пространстве терм-множеств номинальных признаков.

После получения единой числовой меры многозначных объектов по каждому из разнотипных признаков возможно применение различных методов классификации и распознавания. Например, попарное сравнение разнотипных признаков по степени априорной надежности диагноза неизвестного объекта описано в [14]. На максимальном уровне 0.8 "наиболее похожими" оказались три признака всех 102 семейств растений: тип соцветия, тип плода, количество тычинок цветка. От них "дальше всех" отстоит признак количества пестиков. Таже возможно использование известных программных пакетов СИГАМД, STATISTICA. В итоговом алгоритме последняя возможность не использовалась, т.к. основной целью являлось создание диалогового компьютерного диагноста растений.

Глава 3 содержит описание примененных способов свертки всех НМ (получение многомерного критерия) для ранжирования многозначных многомерных объектов и для сравнения разнотипных признаков по надежности диагностики неизвестных объектов.

Полученные по всем признакам решетчатые ФП объектов сводятся в числовую матрицу "надежности диагноза" размера (102*11): $U=\|\mu_{nm}(x)\|$.
Здесь n – индекс объекта-таксона, m – индекс признака.
Цель многомерного нечеткого вывода – нахождение минимального по некоторому критерию пути для диагноза особи может быть достигнута в результате анализа текущей матрицы U по строкам и столбцам. Свертка матрицы U по строкам-объектам означает итоговое ранжирование исходных разнотипных признаков по степени априорной надежности диагностики. Усреднение по столбцам дает итоговый критерий для ранжирования многозначных объектов по совокупности разнотипных признаков [9], [13].

Исходную ТЭД можно представить как результат голосования на выборах. При этом таксоны-семейства являются "избирателями", голосующими за любое число вариантов по каждому признаку. Вычисление матрицы нечеткостей объектов $\|\mu_{nm}(x)\|$ означает в этой модели приведение индивидуальных профилей избирателей к единой интервальной шкале весовых значений нечетких признаков. Полученная матрица U нетранзитивна по НМ-признакам, поэтому определение наилучшего признака для очередного диагностического вопроса является решением задачи группового выбора. Известна теорема К.Эрроу, согласно которой не существует в общем случае решения этой задачи при числе признаков-вариантов больше двух и при выполнении условий нескольких естественных аксиом. Поэтому для решения необходимы некоторые упрощающие условия, или эвристики[5], каковыми являются примененные три способа многокритериального выбора. В программе диагностики пользователь может выбрать любой из этих способов, а по умолчанию установлена осторожная стратегия.

Рискованная стратегия, или экстремальный выбор.
Этот принцип, применяемый в теории игр, состоит в том, что из нескольких альтернатив выбирается та, которая имеет абсолютный максимум критерия. Реализация этой стратегии осуществляется ранжированием 11 признаков-НМ

по убыванию максимальных значений ФП каждого НМ, то есть исходная матрица $||\mu_{nm}(x)||$ сворачивается по строкам операцией $S_m=\max_n\{\mu_{nm}(x)\}$, и остается вектор-строка $|S_m|$ из 11 значений. Максимальный элемент этого вектора соответствует тому признаку ТЭД, по которому надлежит задать вопрос на очередном шаге диагностики. Эта стратегия в небольшом числе случаев дает быструю диагностику особи (за 1-2 ответа на вопросы).

Усредненная, или осторожная стратегия выбора.

Эта стратегия является фактически процедурой Ж.-Ш.де Борда обработки результатов голосования[6]. Она требует от избирателей упорядочения всех вариантов по предпочтению. Числовая матрица **U**, элементы которой можно считать дробным "числом голосов", поданных объектами-избирателями за каждый из 11 вариантов, обеспечивает это сравнение признаков. Вычисляемые по каждому столбцу-признаку суммы $S_m=\sum_n \mu_{nm}(x)$ ранжируются по убыванию, что и определяет ранжирование исходных признаков-вариантов по степени их априорной прогнозной способности. В алгоритме диагностики особи на очередном шаге рекомендуется применять признак-победитель, соответствующий сумме S_1

Выбор по минимуму коэффициента нечеткости НМ.

Известен коэффициент нечеткости НМ А.Кофмана:

$\nu(F) = (2/N)\cdot(\sum \min(\mu_F(x_n),\mu_{\neg F}(x_n)))$ Используется следующая эвристика:

чем меньше коэффициент $\nu(F_m)$, тем лучше соответствующий признак m подходит для диагностики. В алгоритме применяется равноценное, по сравнению с выражением для $\nu(F)$, ядро $|\mu_F(x_n)-(1-\mu_F(x_n))|$, на основе которого можно получить расстояние Хемминга-Заде между НМ и его дополнением: $D_m = \sum_n |\mu(x_{nm})-(1-\mu(x_{nm}))|$. Оптимальным для очередного шага диагностики считается тот признак m, который имеет максимальное значение D_m.

Это соответствует минимуму нечеткости по Кофману, то есть объекты-таксоны менее всего пересекаются по градациям значений этого признака, по сравнению со всеми другими признаками. Если применить ядро информационной функции Шеннона для определения расстояния между НМ и его дополнением, то суммирование по всем объектам этой меры различия дает энтропию НМ. Выбор признака по максимуму нечеткой энтропии приводит к таким же практическим результатам при диагностике, как и по минимуму нечеткости Кофмана.

---

[4] Resconi G Uncertainty Theories by Modal Logic / Computational Intelligence Soft Computing and Fuzzy-Neuro Integration with Applications Ed by O Kaynak, L A Zadeh, B Turksen, I J Rudas Springer, Series F Computer and Systems Sciences, Vol 162 1999. – Pp. 60-79.

[5] Айзерман М.А., Алескеров Ф.Т. Задача Эрроу в теории группового выбора (анализ проблемы) // Автоматика и телемеханика, 1983, N 9. – С. 127-151.

[6] Вольский В.И., Лезина З.М. Голосование в малых группах: процедуры и методы сравнительного анализа – М.: Наука, 1991. 192 с.

[7] Zadeh L On fuzzy algorithms Memorandum No ERL-M325 Electronics Research Lab , College of Engng , Univ of California, Berkeley, February 22, 1972

Глава 4 посвящена описанию работы программы-диагноста растений ©RECOFAM. Алгоритм ее работы описан в [17], [19].

Общая постановка вопроса о сходимости нечетких алгоритмов не имеет смысла по существу[7], ввиду неопределенности исходной информации.

Но в задаче распознавания растений сходимость алгоритма предопределена. Дело в том, что признаковое описание пересекающихся классов-таксонов является экспертной информацией, предоставленной ботаниками. Учтены все возможные проявления (реализации) значений признаков у растений. Поэтому алгоритм диагностики особи является нечетким последовательным поиском прецедента-таксона.

Программа циклически выполняет три основных блока:

1) Для каждого из семейств-кандидатов на диагноз, оставшихся после очередного цикла программы (в начале работы это все 102 семейства), вычисляются значения ФП ко всем 11 НМ-признакам способами, описанными в главе 2.

2) Решается задача группового выбора очередного признака-вопроса, то есть ранжирование признаков по критерию априорной надежности диагноза особи относительно оставшихся таксонов. Как описано в главе 3, реализованы три стратегии выбора. Пользователь может применить любую из них, а "по умолчанию" установлена осторожная стратегия.

Программа выдает на экран список признаков, ранжированный по убыванию вычисленного критерия. Пользователь может не следовать рекомендации алгоритма и выбрать не первый, а любой признак из списка, например, в учебных целях, или для тестирования, или используя наиболее выраженные у особи признаки, но длина ключа при этом может возрасти (для диагностики принадлежности особи к конкретному классу-семейству потребуется ответить на большее количество вопросов).

Этот факт подтвержден многочисленными экспериментами с программой и относится ко всем стратегиям выбора признаков.

3) Программа выдает список значений того признака, который пользователь выбрал в п. 2) алгоритма. После этого пользователь, сверяясь с реальным строением органов растения (гербарным листом), должен указать одно значение из списка. Здесь возможен ответ "не знаю", если, например, соответствующий орган отсутствует или плохо сформирован. Программа при этом переходит к п. 1). Возможность ответа "не знаю" является характерной особенностью алгоритма – процесс диагностики становится надежным (устойчивым к сомнениям пользователя). Далее, в соответствии с ответом пользователя и исходной ТЭД, производится ограничение множества семейств-кандидатов на диагноз, затем осуществляется переход к п. 1), и так далее, до одного кандидата, который является ответом, либо до появления противоречий в ответах пользователя о значениях признаков особи (в ТЭД нет такого класса). Большинство растений правильно диагностируется по ответам всего на 2 – 4 вопроса, вместо полного вектора из 11 ответов (по числу признаков).

Все остающиеся после очередного цикла работы программы семейства-кандидаты ранжированы по убыванию итогового (многомерного)

коэффициента надежности диагноза, и этот список выводится на экран. Поэтому, если в конце работы (при исчерпании всех 11 вопросов о значениях признаков) останется несколько кандидатов, то первый в списке является самым предпочтительным. Такая ситуация может возникнуть, если пользователь часто отвечает "не знаю" на вопросы о значениях признаков. Визуализация в каждом цикле ранжированного списка семейств-кандидатов на диагноз может быть использована в учебных целях при изучении морфологии растений.

Эффективность представленного алгоритма минимизации числа вопросов, задаваемых при диагностике растения, ввиду отсутствия алгоритма-прототипа, демонстрируется примерами по таблицам 2–5.

В табл. 2 показана диагностика особи по рекомендациям алгоритма.

Таблица 2

Вопрос	Ответ	Осталось кандидатов (из 102)
1. соцветие	корзинка	2
2. околоцветник	простой венчиковидный	1 – **Asteraceae**

В табл. 3 показан путь диагностики того же растения, но вопреки рекомендациям алгоритма, то есть всегда выбирался последний признак из ранжированного списка, выдаваемого на шаге 2) описанного цикла работы программы. Из сравнения с табл. 2 видно, как неэффективно диагностировалось растение по примеру табл. 3 – пришлось отвечать на 9 вопросов о значениях признаков вместо двух.

Таблица 3

Вопрос	Ответ	Осталось кандидатов (из 102)
1. пестики	1	99
2. завязь	нижняя	21
3. гинецей	синкарпный	19
4. размножение	цветки обоеполые	16
5. листочки околоцветника	4	10
6. околоцветник	простой венчиковидн	3
7. андроцей	5 тычинок	3
8. столбики	1	3
9. листья	супротивные	1 – **Asteraceae**

В табл. 4 приведен пример диагностики растения при сомнении пользователя в значении типа плода – ответ "не знаю" не сорвал процесс диагностики вследствие избыточности исходной информации и оптимизационных свойств алгоритма.

Таблица 4

Вопрос	Ответ	Осталось кандидатов (из 102)
1. соцветие	метелка	26
2. плод	не знаю	26
3. андроцей	2 тычинки	5
4. столбики	3	1 – **Rosaceae**

В табл. 5 показана эффективность алгоритма при диагностике представителей трех трудных таксонов при выборе признаков-вопросов разными способами. Видно явное превосходство действий пользователя в соответствии с рекомендациями программы по сравнению с другими способами выбора признаков при диагностике одних и тех же растений.

Число вопросов до точного диагноза особи **Таблица 5**

Семейство	По алгоритму	По датчику случайных чисел	Вопреки алгоритму
Caryophyllaceae	2	7	7
Ranunculaceae	3	5	4
Rosaceae	2	5	9

Глава 5 содержит результаты применения способа вычисления ФП многозначных объектов к НМ по номинальному признаку для выявления степени уверенности в ошибках при индексировании документов БД ключевыми словами [6], [15], [16]. При создании электронного каталога гербария NS автором впервые предложено кодировать описание места сбора растения [1], [2]. Кодирование флористических и экологических параметров осуществлялось по правилам, разработанным ботаниками [3], [4], [5].
В результате была разработана технология работы БД каталога гербарных этикеток. В одном из разделов этой БД содержатся практически все документы, описывающие сборы по Туве рода Artemisia (Asteraceae) полынь, поэтому возможен содержательный анализ информации, использующий результаты кодирования текстов этикеток.

Была поставлена задача прогноза ошибок в документах БД как степени несоответствия вида растения и типа места обитания (экотопа) или типа фитоценоза этого растения. Приуроченность видов растений к экотопам и фитоценозам установлена ботанической наукой, но эта приуроченность многозначна, то есть каждый вид растений встречается во многих экотопах и фитоценозах. Таким образом, тип ТЭД в этой задаче такой же, как в случае многозначных семейств по номинальному признаку.

Таблица 6 составлена по результатам обобщения 441 документов БД, соответствующих сборам ведущих специалистов-ботаников.
По этой экспертной таблице, в соответствии с содержанием раздела 2.4 диссертации, можно вычислить ФП видов полыни к НМ как степень соответствия видов растений типам их места обитания или фитоценоза.

Подозрительными на ошибку считались 344 документа БД, отражающие сборы менее опытных ботаников. Если в этих документах характеристическое соответствие вида растения и типа места обитания или фитоценоза не совпадало с содержанием табл. 6, то такие документы, вместе с прогнозным значением степени несоответствия $BAD_i=1-\mu(x_i)$, фиксировались. В итоге было выявлено 21 "подозрительных" документов со степенью уверенности в ошибке BAD в диапазоне от 0.85 до 0.99. Тщательная проверка ботаниками соответствующих 21 гербарных образцов подтвердила наличие 16 ошибок

в определениях видов растений, либо типов их места обитания или фитоценозов. То есть доля фактических ошибок в документах оказалась 0.76 от предсказанных.

Приуроченность видов растений к экотопам и фитоценозам **Таблица 6**

Виды	**Экотоп, Фитоценоз**								
Artemisia (Asteraceae)	берег	болото	кустарники	лес	луг	осыпь	скала	сорные	степь
commutata	1	0	1	1	1	0	1	0	1
dolosa	1	0	1	0	0	1	0	0	1
dracunculus	1	0	1	1	1	0	0	1	1
frigida	1	0	1	1	1	1	1	0	1
glauca	0	0	1	1	1	0	0	1	1
laciniata	0	1	1	1	1	0	0	0	1
latifolia	0	0	1	1	1	0	0	1	1
leucophylla	1	0	1	1	1	0	1	0	0
macrantha	0	0	0	1	1	0	0	0	1
macrocephala	1	0	1	1	1	0	0	1	1
mongolica	1	0	0	1	1	0	0	1	1
obtusiloba	0	0	1	0	0	0	1	1	1
palustris	1	0	1	1	0	0	0	0	1
santolinifolia	1	0	1	1	1	0	1	1	1
scoparia	1	0	0	1	1	1	1	1	1
sieversiana	1	0	0	1	1	0	0	1	1
tanacetifolia	1	0	1	1	1	0	1	0	1
tomentella	0	0	0	1	0	0	1	1	1
vulgaris	1	0	1	1	1	0	1	1	1

Содержание табл. 6 можно интерпретировать как результат кодирования документов библиографической БД списками ключевых слов.
В роли ключевых слов выступают типы экотопа и фитоценоза, а виды растений можно считать индексами системы УДК для книг и статей. Таким образом, показана возможность применения теории НМ для прогноза степени ошибок кодирования ключевыми словами содержания документов различных БД.

ЗАКЛЮЧЕНИЕ

Основные результаты проведенного исследования состоят в следующем:
1 Предложены два способа вычисления функций принадлежности (ФП) многозначных (неопределенных, или списковых по значениям исходных признаков) объектов к нечетким множествам (НМ) – соответственно, по числовым и номинальным признакам.
Способ вычисления ФП объектов к НМ по числовым признакам на основе двух последовательностей фокальных элементов полностью применим к качественным признакам, измеренным в шкале порядка (баллов).
Вычисления ФП не предполагают участие человека-эксперта.

2. Предложен метод диагностики многомерных объектов в условиях пересечения классов по всем разнотипным признакам, когда неприменимы вероятностно-статистические модели анализа информации. Получены числовые критерии для ранжирования многомерных объектов по совокупности признаков и ранжирования разнотипных признаков по степени их априорной диагностической способности по отношению к неизвестным объектам. Работоспособность и эффективность метода продемонстрированы на примере обработки реальной ботанической информации.

3. Разработан алгоритм и диалоговая программа диагностики растений ©RECOFAM, позволяющая безошибочно определять принадлежность особи к ботаническому семейству при минимальном количестве вопросов, задаваемых пользователю относительно значений признаков диагностируемой особи. Допускается ответ "не знаю" на любой вопрос о значениях признаков. Программа внедрена в Новосибирском и Омском государственных педагогических университетах на кафедрах ботаники. Реализованный метод диагностики многомерных разнотипных объектов может применяться для решения сложных задач обработки зашумленной информации (списковые признаки объектов) и в других областях науки и техники.

4. На конкретном примере продемонстрирован результат вычисления степеней ошибок индексирования документов электронной базы данных одного из разделов гербария NS (БД была создана в процессе работы над темой исследования). Количество фактических ошибок составило 76% от предсказанных. Это показывает возможность использования предложенного метода диагностики многозначных объектов для выявления ошибок кодирования ключевыми словами документов, например, библиографических компьютерных БД.

5. Предложен коэффициент оценки усредненной избыточности описания многозначных объектов. Будучи примененным ко всей таблице исходных данных по семействам двудольных растений Сибири, коэффициент эксцесса получил значение пропорции золотого сечения Ф=1.62.

Факт проявления одного из основных законов природы и искусства можно трактовать как свидетельство удачного выбора флористами системы признаков для формализованного описания таксонов-семейств растений.

6. Предложен способ нормализации нечетких множеств путем добавления к исходным нечетким объектам одного искусственного, имеющего инцидентность со всеми значениями признаков. Этот прием позволяет практически сохранить масштаб функций принадлежности объектов к НМ, что важно для многомерного нечеткого анализа.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Красинский В.И. Практический опыт создания автоматизированных баз данных в гербарном деле.– Новосибирск, 1990. Рукопись, депонированная в ВИНИТИ, N 4290-В90. 10 с.
2. Красинский В.И., Артемов И.А. Ботаническая информационно-поисковая система "ФЛОРА" *опыт разработки и перспективы развития* / IV научный семинар с международным участием "Автоматизированные библиотечно-информационные системы" (тез. докл.). – Новосибирск, 1991. С. 93-95.
3. Красноборов И.М., Красинский В.И., Артемов И.А. Ботанические компьютерные базы данных в ЦСБС СО РАН / II совещание "Компьютерные базы данных в ботанических исследованиях" (тез. докл.). – С.-Пб., 1995. – С. 26-27.
4. Красноборов И.М., Красинский В.И., Артемов И.А., Николаев С.В. Создание ИПС для решения задач ботанической географии (1-ый этап, АРМ "ФЛОРИСТ-1"). / Информационные системы в науке 95 (тез. докл.). М.: ФАЗИС, 1995. – С. 61-62.
5. Красноборов И.М., Красинский В И., Артемов И.А. Ботанические компьютерные базы данных и анализ флористической информации / Труды IV междунар. симпозиума по результатам международной программы биосферного мониторинга "Эксперимент Убсу-Нур". М.: ИНТЕЛЛЕКТ, 1996. С. 81-87.
6. Красинский В.И., Красноборов И.М. Выявление знаний из содержимого гербарных этикеток (на примере сборов рода ARTEMISIA из Тувы в гербарии NS) / Сб. научн. трудов "Компьютерные базы данных в ботанических исследованиях". – С.-Пб.: БИН РАН, 1997. – С. 45-49.
7. Красинский В И. Применение теории возможностей для ранжирования многозначных ботанических объектов // Автометрия, 1999, N 3. С. 65-81.
8. Krasinsky V.I. Application of fuzzy sets for improvement of the diagnosis of multivariate multivalued botanical objects // Abstracts of the third Russian-Korean International Simposium on Science and Technology KORUS'99. June 22-25, 1999 at Novosibirsk State Technical University. Novosibirsk, Russia. Vol. 2, p. 509.
9. Krasinsky V.I. Comparison of objects by the degree of fuzzyness of nominal characters. // Proceedings of the International Conference "INTERACTIVE SYSTEMS: THE PROBLEMS OF HUMAN-COMPUTER INTERACTION". 22-24 Sept. 1999, Ulianovsk. – P. 68.
10. Красинский В.И. Диалоговый определитель семейств растений Сибири – учебный тренажер на основе теории возможностей // VII Всероссийский семинар "Нейроинформатика и ее приложения" 1-3 октября 1999 г. (тез. докл.). КГТУ: Красноярск, 1999. – С. 86.

11. Красинский В.И. Формализация знаний на основе метода парных сравнений (методические заметки)
/ Междунар. телеконференция "Биометрика-2000". ТГУ, Томск, 2000.
http://www.biometrica.tomsk.ru
12. Красинский В.И. Диагностика многомерных многозначных ботанических объектов на основе теории нечетких множеств – программа RECOFAM / Междунар. телеконференция "Биометрика-2000" ТГУ, Томск, 2000. http://www.biometrica.tomsk.ru
13. Красинский В.И. Классификация биологических объектов по совокупности списковых признаков переменной длины (на основе нечетких множеств) // Четвертый Сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-2000).
Тез. докл. Ч. III. – Новосибирск, ИМ СО РАН, 2000. – С. 65.
14. Красинский В.И. Сравнение разнотипных признаков по степени надежности распознавания многозначных объектов
// Четвертый Сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-2000). Тез. докл. Ч. III. Новосибирск, ИМ СО РАН, 2000. – С. 92.
15. Красинский В.И. Предсказание ошибок в документах базы данных на основе нечеткого дескриптора по ключевым словам
/ V рабочее совещание по электронным публикациям EL-PUB2000.
СО РАН, Новосибирск, 2000.
http://www-sbras.nsc.ru/ws/el-pub-2000/
16. Красинский В.И. Прогноз степени несоответствия определений вида растения и типа фитоценоза в этикетках электронного каталога гербария NS на примере сборов из Тувы рода Artemisia (Asteraceae)
/ Проблемы создания ботанических баз данных:
Тез. докл. совещания (Новосибирск, 24-26 октября 2000г.).
М.: ПАТЕНТ, 2000. – С. 34-35. http://www-sbras.nsc.ru/ws/botdb/
17. Красноборов И.М., Красинский В.И. Применение теории нечетких множеств для определения семейств растений
// Доклады Академии наук, 2000. Том 374, N 4 С. 565-567.
18. Красинский В.И. Нечеткая классификация объектов малой числовой выборки // Автометрия, 2001, N 5. – С. 117-125.
19. Красинский В.И. Распознавание растений по разнотипным признакам на пересекающихся классах-таксонах (на основе теории нечетких множеств)
/ Всероссийская конференция, посвященная 90-летию А.А.Ляпунова.
СО РАН, Новосибирск, 2001. http://www-sbras.nsc.ru/ws/Lyap2001/

Формат бумаги 60×84 1/16. Объём 1 печ. л.
Тираж 100 экз.

Отпечатано на ротапринте Новосибирского института органической химии СО РАН им. Н.Н. Ворожцова
630090, Новосибирск, 90, пр. акад. Лаврентьева, 9