РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2008, том 44, № 3

УДК 519.95

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, ОСНОВАННЫЙ НА ФУНКЦИИ КОНКУРЕНТНОГО СХОДСТВА*

Н. Г. Загоруйко

Институт математики им. С. Л. Соболева СО РАН, г. Новосибирск
E-mail: zag@math.nsk.ru

Предлагается единый подход к построению методов интеллектуального анализа данных (ИАД), или Data Mining. Он основан на использовании функции конкурентного сходства (FRiS-функции), отражающей способы оценки сходства и различия человеком. Кратко описаны методы ИАД, основанные на этом подходе. Приводятся примеры решения модельных и реальных задач с помощью таких методов.

Введение. Для ориентации в окружающей среде человек наблюдает, запоминает результаты наблюдений, обнаруживает в них закономерности (знания) и действует, опираясь на полученные знания. Методы интеллектуального анализа данных (ИАД), или Data Mining, используются для поддержки всех этих этапов деятельности человека. Имеется большой опыт применения методов ИАД в геологии, генетике, медицине, экономике и т. д. Внешние признаки благополучия в области ИАД выглядят убедительно. Ежегодно проводятся десятки конференций, издаются десятки журналов, опубликованы сотни монографий, разработаны тысячи алгоритмов и программ, которыми пользуются десятки тысяч специалистов разных прикладных областей. Но даже поверхностного знакомства с методами ИАД достаточно, чтобы увидеть недостатки в развитии теоретических основ этого научного направления. Отсутствует единый подход к решению разных задач ИАД. Более того, для решения одних и тех же задач используются различные подходы, обоснованность которых не анализируется. Наибольшие трудности связаны с плохой обусловленностью многих задач, в которых количество признаков превышает число объектов наблюдения.

При выборе основы для построения теории ИАД нужно обратить внимание на то, что в нашем распоряжении имеется образец системы, замечательно решающей все задачи ИАД. Эта система – человек, который непрерывно и повседневно занимается тем, что обнаруживает знания и на их основе классифицирует, распознает, выбирает важные признаки, прогнозирует и т. д.

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 05-01-00241, № 08-01-00040).

Скорее всего, при решении этих разных и плохо обусловленных задач он пользуется некоторой универсальной психофизиологической функцией, отвечающей за ориентацию человека в окружающей среде.

В данной работе предлагается рассмотреть следующую гипотезу: основная функция, используемая человеком при решении задач сбора данных, классификации, распознавания, выбора признаков, прогнозирования и т. д., состоит в определении сходств и различий.

Покажем, что мера, воспроизводящая механизм оценки сходства объектов человеком, позволяет строить для решения задач ИАД разного типа единообразные алгоритмы, инвариантные к степени обусловленности и к виду распределений объектов в пространстве характеристик.

Функция конкурентного сходства. При решении задач распознавания образов часто используется решающее правило, основанное на сравнении степени «похожести» контрольного объекта на эталоны конкурирующих образов. В литературе описаны десятки различных мер сходства [1]. Как правило, в этих мерах сходство контрольного объекта Z с эталонами всех образов носит абсолютный характер и зависит только от расстояний до этих эталонов. Но легко убедиться, что человеческое восприятие похожести носит относительный характер. Чтобы ответить на вопросы типа близко–далеко, похож–не похож, нужно знать ответ на вопрос: по сравнению с чем?

В некоторых алгоритмах распознавания, например в правиле k ближайших соседей (kNN), решение о принадлежности объекта Z образу S_i принимается не в том случае, когда расстояние r_i до него «мало», а когда оно меньше расстояния r_j до конкурирующего образа S_j. Следовательно, чтобы оценить похожесть объекта Z на первый образ, нужно знать расстояние не только до него, но и до ближайшего конкурента и сравнивать эти расстояния в шкале порядка. Если же нас интересует мера конкурентного сходства, измеренная в более сильной шкале отношений, то можно использовать следующую величину:

$$F_{i/j} = (r_j - r_i)/(r_j + r_i), \tag{1}$$

которую в дальнейшем будем называть функцией конкурентного сходства или FRiS-функцией (Function of Rival Similarity). Эта функция имеет относительный характер и хорошо согласуется с механизмами восприятия сходства и различия человеком. Значение функции конкурентного сходства $F_{i/j}$ меняется в пределах от + 1 до –1. Если контрольный объект Z совпадает с эталоном первого образа, то $r_i = 0$ и $F_{i/j} = 1$, а $F_{j/i} = -1$. При расстояниях $r_i = r_j$ значения $F_{i/j} = F_{j/i} = 0$, что указывает на границу между образами. В точках границы объект в равной степени похож и не похож на эти конкурирующие образы.

Опыт работы с FRiS-функцией показал, что она может использоваться в качестве базового элемента для решения различных задач ИАД. При решении задачи распознавания в условиях, когда есть возможность оценивать дисперсии d_i и d_j распределений конкурирующих образов, нужно пользоваться нормированными расстояниями до их эталонов: $R_i = r_i / d_i$ и $R_j = r_j / d_j$. В результате нормированная функция конкурентного сходства имеет следующий вид:

$$F_{i/j} = (R_j - R_i)/(R_j + R_i). \tag{2}$$

Рассмотрим способы использования FRiS-функции при решении некоторых задач ИАД.

Построение решающих правил (алгоритм FRiS-Stolp). Для распознавания образов необходимо выбрать объекты-эталоны, с которыми будут сравниваться контрольные объекты. Выбор эталонов (столпов) для каждого образа можно осуществить с помощью алгоритма FRiS-Stolp.

Пусть решается задача распознавания «первый образ против всех остальных». Проверяется вариант, при котором первый случайно выбранный объект a_i является единственным столпом образа S_1, а все другие образы в качестве столпов имеют все свои объекты.

1. Для всех объектов $a_j \neq a_i$ первого образа S_1 находится расстояние r_{ji} до столпа a_i и расстояние r_{jt} до ближайшего объекта чужого образа S_t. По этим расстояниям вычисляются значения сходства $F_{i/t}$ объектов a_j со своим столпом. Находим те m_i объектов первого образа, значение функций сходства F_i которых выше заданного порога F^*, например $F^* = 0$.

2. Аналогичную процедуру повторяем, назначая в качестве столпа все M_1 объектов первого образа по очереди.

3. Находим объект a_i с максимальным значением m_i и объявляем его первым столпом A_{11} первого кластера C_{11} первого образа S_1.

4. Исключаем из первого образа m_i объектов, входящих в первый кластер. Для остальных объектов первого образа находим следующий столп повторением пп. 1–3. Процесс останавливается, если все объекты первого образа оказались включенными в свои кластеры.

5. Восстанавливаем все объекты образа S_1 и для всех остальных образов повторяем пп. 1–4.

На этом шаге заканчивается первый этап поиска столпов. Каждый столп A_i защищает подмножество объектов m_i своего кластера C_i. Найдем расстояния между столпом и всеми объектами его кластера. Среднее значение этих расстояний d_i будем считать внутренним расстоянием кластера C_i. Теперь расстояние от контрольного объекта Z до кластера C_i будет определяться нормированным расстоянием

$$R(Z, A_i) = r(Z, A_i)/d_i.$$

6. Столпы были выбраны в условиях, когда им противостояли все объекты конкурирующих образов. Теперь образы представлены только своими столпами. Для уточнения состава кластеров распознаем принадлежность всех объектов к кластерам в условиях, когда функция F определяется по нормированным расстояниям до ближайшего своего и ближайшего чужого столпов. Если в результате произошло перераспределение объектов между кластерами, то для нового состава каждого кластера нужно повторить процедуру определения нового внутреннего расстояния d_i.

7. В заключение необходимо найти среднее значение функций сходства F_s всех объектов со своими столпами. Величина F_s характеризует качество обученности системы и тесно связана с величиной ошибок, которые будут получаться при распознавании контрольных объектов.

Итогом работы алгоритма FRiS-Stolp является решающее правило в виде списка эталонов (столпов), которые описывают каждый образ, списка объектов, входящих в каждый кластер, значений внутренних расстояний для каждого кластера и среднего значения функций сходства F_s.

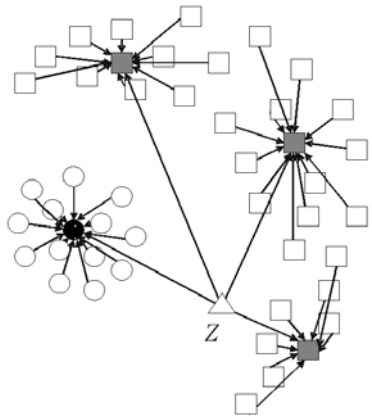

Рис. 1. Распознавание принадлежности объекта *Z* одному из двух образов (круги и квадраты), разделенных на кластеры

Алгоритмом FRiS-Stolp первыми выбираются столпы, расположенные в центрах локальных сгустков и защищающие максимально возможное количество объектов с заданной надежностью. По этой причине при нормальных распределениях в первую очередь будут выбраны столпы, расположенные в точках математического ожидания. Если распределения полимодальны и образы линейно не разделимы, столпы стоят в центрах мод. С ростом сложности распределения число столпов *k* будет увеличиваться.

Процесс распознавания с опорой на столпы очень прост и состоит в оценке нормированных функций конкурентного сходства контрольного объекта *Z* со всеми столпами и выборе образа, чей столп получил максимальное значение *F* (рис. 1).

Еще одним важным преимуществом такого решающего правила является возможность использования значения *F* в качестве оценки надежности принятого решения при распознавании конкретного объекта в условиях, когда закон распределения образов не известен. Результаты распознавания объектов контрольной выборки, получивших разные значения функции сходства *F*, представлены на рис. 2.

Как и ожидалось, при значениях *F*, близких к нулю, вероятность ошибочного распознавания составила около 50 %. С увеличением значения функции сходства *F* вероятность ошибки *P* быстро уменьшается. Используя значение функции сходства контрольного объекта с ближайшим столпом, можно указать вероятность того, что результат распознавания окажется правильным.

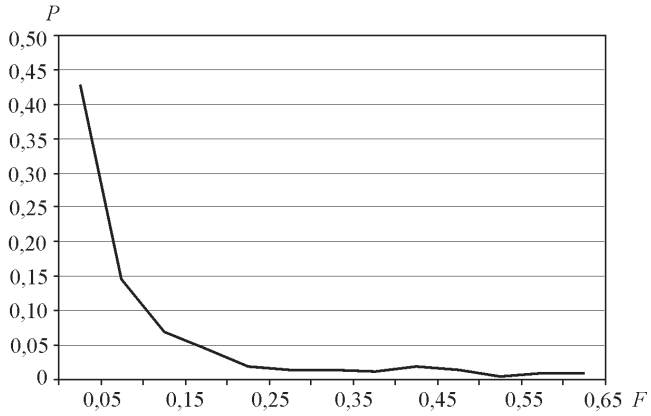

Рис. 2. Вероятность ошибки распознавания *P* в зависимости от величины функции сходства *F*

Выбор информативных признаков (алгоритм FRiS-GRAD). При решении задачи выбора подмножества наиболее информативных элементов из их большого исходного множества может быть применен любой алгоритм направленного перебора, например алгоритм GRAD [2, 3]. Он позволяет автоматически указать как состав, так и наилучшее количество характеристик. Здесь мы обращаем внимание на новый критерий информативности алгоритма, основанный на использовании FRiS-функции.

Критерий информативности. Для оценки информативности признаков или их сочетаний обычно используется критерий в виде количества ошибок U распознавания обучающей выборки в режиме скользящего экзамена. Главным недостатком этого критерия является то, что он не учитывает надежность распознавания объектов, которые распознаны правильно, и грубость ошибки тех объектов, которые распознаны неправильно. Было показано [4], что учесть эти особенности можно, если использовать в качестве критерия информативности нормированную функцию сходства F_s объектов с эталонами своих образов. Дополнительно нужно учитывать количество кластеров k, на которые распадаются классы. Чем меньше число k кластеров отличается от числа образов K, тем информативнее пространство признаков. В итоге получается критерий информативности подпространства в виде величины $Q = hF_s$, где $h = K/k$. В наилучшем случае, когда достаточно взять по одному столпу на образ, h будет равно единице. В наихудшем случае, когда число столпов равно числу объектов M, $h = K/M$.

При постепенном увеличении размерности выбираемого алгоритмом GRAD подпространства признаков качество классификации вначале растет, достигает некоторого максимума и затем начинает уменьшаться. Точка с максимальным значением Q соответствует наилучшему подмножеству признаков.

Преимущества описанного критерия информативности Q перед критерием ошибок на обучающей выборке U можно проиллюстрировать результатами их экспериментального сравнения. Исходные данные состояли из 200 объектов двух образов (по 100 объектов каждого образа) в 100-мерном пространстве. Признаки генерировались так, чтобы они обладали разной информативностью. В итоге около 30 признаков оказывались в той или иной степени информативными, а остальные – генерировались датчиком случайных чисел и были заведомо неинформативными. Дополнительно эта исходная таблица искажалась шумами разной интенсивности, и при каждом уровне шума (от 0,05 до 0,30) алгоритмом GRAD выбирались наиболее информативные подсистемы размерности n (от 1 до 22). При этом для обучения случайно выбиралось по 35 объектов каждого образа. На контроль предъявлялись остальные 130 объектов. Сравнение критериев Q и U представлено на рис. 3.

Результаты контроля показывают, что критерий Q обладает более высокими прогностическими свойствами и помехоустойчивостью по сравнению с критерием U, который создает необоснованные иллюзии.

Оценка пригодности признаков (проблема А. Н. Колмогорова). В 1933 г. А. Н. Колмогоров опубликовал работу [5], в которой обратил внимание на трудности, связанные с решением проблемы выбора подмножества информативных предикторов при построении регрессионных уравнений для случая, когда количество потенциальных предикторов сравнимо или превышает количество наблюдаемых объектов. Дело в том, что встречаются задачи, в которых значительная часть характеристик играет роль случайного шума. Чем больше таких характеристик, тем выше вероятность обнаружения «псевдо-

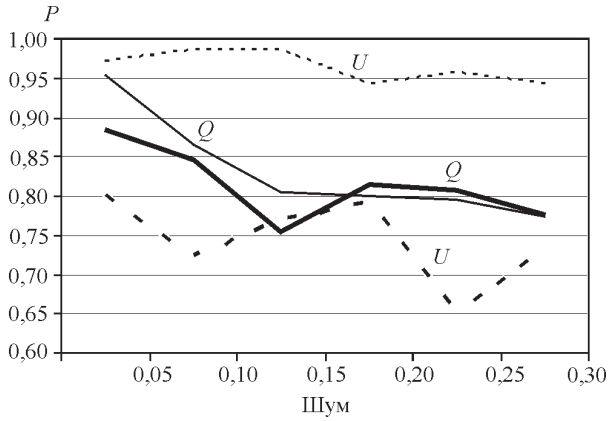

Рис. 3. Результаты обучения и распознавания по критериям U и Q при разных уровнях шумов (тонкие линии – обучение, жирные – контроль)

информативного» набора из шумовых предикторов. Вопрос А. Н. Колмогорова о том, как отличить «пригодную» систему признаков от непригодной, не теряет своей актуальности.

Для ответа на этот вопрос мы предлагаем следующую процедуру (алгоритм FRiS-Test). В анализируемой таблице данных выбираем n наиболее информативных признаков и оцениваем среднее значение функции сходства F_s объектов обучающей выборки со своими эталонами. Затем превращаем эту таблицу в случайную путем случайной перестановки значений каждого признака. Такая пертурбация разрушает имеющиеся в таблице зависимости между описывающими и целевым признаками. По этой таблице выбирается наилучшая подсистема из n признаков и оценивается среднее значение функции сходства F_s^* объектов с эталонами. Повторяем процедуру формирования случайных таблиц и выбора в них наилучшего подпространства признаков k раз. Среди полученных оценок F_s^* находим максимальное значение $F_{s\,\max}^*$. Эта верхняя граница коридора случайных результатов сравнивается со значением F_s, полученной на исходной таблице. Если $F_s > F_{s\,\max}$, то подсистема признаков, найденная по исходной таблице, может считаться неслучайной. Если же величина F_s для исходной таблицы попадает в пределы коридора значений F_s^* для случайных таблиц, то можно считать, что выбранные признаки «псевдоинформативны». Они не пригодны для дальнейшего использования.

Построение классификаций (алгоритм FRiS-Class) [6]. Автоматическая классификация объектов в виде иерархии классов или списка классов одного иерархического уровня делается с помощью алгоритма FRiS-Tax. Его работа состоит из двух этапов. На первом этапе алгоритмом FRiS-Cluster выбираются объекты, находящиеся в центрах локальных сгустков объектов. Такие объекты становятся эталонами (столпами) кластеров. На втором этапе с помощью алгоритма FRiS-Class происходит процедура укрупнения кластеров в классы (таксоны) путем объединения некоторых соседних кластеров в один класс. Это позволяет создавать классы произвольной формы, не обязательно линейно разделимые.

Кластеризация. Условия использования функции сходства в задаче построения классификации отличаются тем, что принадлежность объектов вы-

борки к тому или иному классу неизвестна. Все объекты как бы принадлежат одному образу. В связи с этим на первом этапе вводится виртуальный образ-конкурент, ближайший столп которого удален от каждого объекта выборки на фиксированное расстояние, равное R_2^*. В результате будем использовать модификацию функции сходства, которая по расстоянию R_1 от любого объекта a_i до объекта, играющего роль центра кластера, будет определяться как

$$F_i = (R_2^* - R_1)/(R_2^* + R_1). \quad (3)$$

Пользователь задает предельное число кластеров K, среди которых он хотел бы выбрать наилучший вариант кластеризации. Алгоритм ищет решения задачи кластеризации последовательно для всех значений $k = 1, 2, \ldots, K$, выполняя следующие операции:

1. Вначале все M объектов поочередно рассматриваются в качестве эталона единственного кластера. Для каждого из них в конкуренции с виртуальным столпом вычисляется значение функции сходства с ним всех остальных объектов. Объекты, для которых $F > F^*$, считаются принадлежащими данному эталону. Столпом A_1 первого кластера C_1 назначается объект, набравший наибольшее количество защищаемых им объектов.

2. Затем для случая $k = 2$ определяется, какой объект будет наилучшим вторым столпом. Для этого на роль второго столпа по очереди назначаются все $(M - 1)$ объектов, не совпадающих с первым столпом. Наличие двух реальных столпов и одного виртуального позволяет отнести каждый объект к первому или второму реальному кластеру. Вторым столпом A_2 выбирается такой объект, при котором оба столпа (A_1 и A_2) набирают в составы своих кластеров наибольшее количество объектов.

3. После выбора эталонов двух кластеров происходит уточнение состава этих кластеров. Объекты переходят в состав того кластера, расстояние до столпа которого меньше.

4. Дальнейшее расширение списка столпов делается аналогичным способом. В итоге будут получены наилучшие варианты классификации для всех значений $k = 1, 2, \ldots, K$. Качество каждого варианта оценивается средним значением F_s функций сходства всех объектов со своими столпами. На этом первый этап кластеризации заканчивается.

Построение классификации. Далеко не всегда класс сложной формы будет состоять из одного кластера. С учетом этого в алгоритме предусмотрен следующий механизм объединения нескольких кластеров в один класс:

5. Каждую пару кластеров C_i и C_j проверяют на наличие спорных объектов, которые находятся около разделяющей их границы. Объект a считается относящимся к зоне конкуренции столпов A_i и A_j, если они являются двумя ближайшими к нему и абсолютная величина его сходства со своим столпом меньше некоторого порога $|F_a| < F^*$.

6. Расстоянием D_{ij} между кластерами C_i и C_j считается минимальное расстояние между двумя объектами разных кластеров, попавшими в зону конкуренции. Для объектов a из C_i и b из C_j, которые находятся в зоне конкуренции и на которых достигается минимум F, определяются расстояния D_a и D_b от каждого из них до ближайшего своего соседнего объекта.

7. Кластеры C_i и C_j объединяются, если значения трех величин (D_{ij}, D_a и D_b) мало отличаются друг от друга.

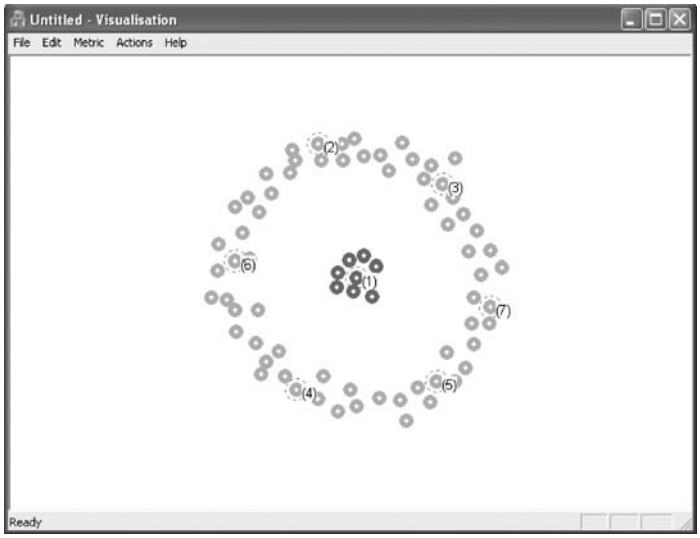

Рис. 4. Пример результата работы алгоритма FRiS-Tax. Номера столпов соответствуют порядку их появления при увеличении числа кластеров k

Важным преимуществом данного алгоритма является возможность автоматического выбора локально-оптимального числа кластеров. Для этого можно использовать значение качества кластеризации F_s при разных количествах кластеров. Лучшим вариантам кластеризации соответствуют локальные максимумы величины F_s. Работу алгоритма FRiS-Tax в двумерном случае иллюстрирует рис. 4. При $k = 7$ получено два класса, что хорошо согласуется с экспертным решением.

Эффективность предложенного алгоритма при работе с пространствами большей размерности в сравнении с существующими алгоритмами таксономии проверялась на прикладных задачах. В одной из них обучающая выборка состояла из рентгеновских спектров 160 образцов, которые по химическому составу делились на пять групп. Каждый спектр представлял собой 1024-мерный вектор. Проводилось разбиение обучающей выборки в пространстве спектральных характеристик на классы (их число варьировалось от 2 до 18) несколькими известными алгоритмами таксономии. Аналогичная задача решалась с помощью алгоритмов FRiS-Cluster и FRiS-Tax. Эффективность алгоритмов оценивалась через величину однородности полученных таксонов с точки зрения химического состава объектов, попавших в них.

Всего в тестировании участвовали пять алгоритмов, оперирующие понятием центра таксона, которые после окончания работы предоставляют пользователю помимо разбиения еще и набор эталонных образцов – столпов. Это следующие алгоритмы:

– самый популярный за рубежом алгоритм K-means [7, 8];

– алгоритм Forel, «раскатывающий» множество исследуемых объектов на таксоны сферической формы [9];

– алгоритм Scat, который из сферических таксонов, созданных алгоритмом Forel, конструирует таксоны более сложной формы [9];

– алгоритм FRiS-Cluster, обеспечивающий построение линейно разделимых кластеров;

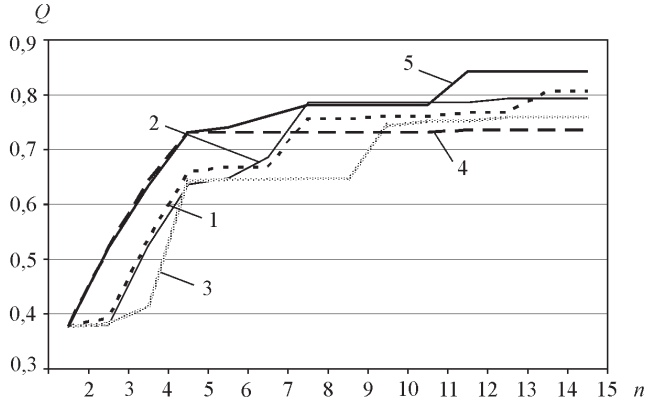

Рис. 5. Сравнение качества пяти алгоритмов таксономии: кривая 1 – FRiS-Cluster, 2 – K-means, 3 – Forel, 4 – Scat, 5 – FRiS-Tax

– алгоритм FRiS-Tax, объединяющий несколько кластеров в один таксон произвольной формы.

Оценки качества таксономии (определяемого мерой однородности таксонов), полученной данными алгоритмами для разного числа таксонов, приведены на рис. 5. Нетрудно видеть, что результаты алгоритма FRiS-Cluster и в пространстве большой размерности оказались не хуже, а результаты FRiS-Tax – лучше имеющихся аналогов. Кроме того, использование FRiS-функции позволило определить вариант таксономии, который следует предпочесть, число таксонов, на котором следует остановиться. Локально-максимальное значение качества кластеризации ($F_s = 0{,}776$) было достигнуто при восьми кластерах, которые затем объединились в пять классов. Анализ их содержания показал максимальную однородность по химическому составу.

Заключение. Предложенная в данной работе функция конкурентного сходства может использоваться в качестве универсального ядра для алгоритмов, решающих все основные задачи ИАД. Алгоритмы, основанные на FRiS-функции, применимы для решения задач с любой степенью обусловленности и при любом характере распределения анализируемых объектов в пространстве признаков. FRiS-функция, имитирующая способы оценки сходства человеком, позволяет получать легко интерпретируемые результаты. Использование этой функции в качестве критерия информативности признаков повышает точность оценки вероятности правильного распознавания контрольной выборки, а также позволяет решать такие новые задачи, как оценка пригодности признакового пространства и автоматическое определение числа кластеров. Качество решений известных задач ИАД с помощью FRiS-функций не уступает качеству, получаемому существующими методами.

Автор выражает благодарность И. А. Борисовой, О. А. Кутненко и В. В. Дюбанову за активное участие в обсуждении и развитии представленной в работе проблемы и проведение многочисленных вычислительных экспериментов.

СПИСОК ЛИТЕРАТУРЫ

1. **Воронин Ю. А.** Начала теории сходства. Новосибирск: Изд-во ВЦ СО АН СССР, 1989.

2. **Zagoruiko N. G., Kutnenko O. A.** Recognition methods based on the AdDel algorithm // Pattern Recogn. and Image Analysis. 2004. **14**, N 2. P. 198.

3. **Zagoruiko N. G., Kutnenko O. A., Ptitsyn A. A.** Algorithm GRAD for selection of informative genetic features // Proc. of the Intern. Moscow Conf. on Computational Molecular Biology. Moscow, 2005. P. 8.

4. **Загоруйко Н. Г., Кутненко О. А.** Алгоритм GRAD для выбора признаков // Тр. Междунар. конф. «Применение многомерного статистического анализа в экономике и оценке качества». М.: Изд-во МЭСИ, 2006. С. 81.

5. **Колмогоров А. Н.** К вопросу о пригодности найденных статистическим путем формул прогноза // Завод. лаборатория. 1933. № 21. С. 164.

6. **Борисова И. А.** Алгоритм таксономии FRiS-Tax // Науч. вестн. НГТУ. Новосибирск: НГТУ. 2007. № 3(28). С. 3.

7. **Шлезингер М. И.** О самопроизвольном разделении образов // Читающие автоматы и распознавание образов. Киев: Наук. думка, 1965. С. 46.

8. **MacQueen J.** Some methods for classification and analysis of multivariate observation // Proc. of the 5th Berkley Symp. on Mathematical Statistic and Probability. Berkley: University of California Press, 1967. Vol. 1. P. 281.

9. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний. Новосибирск: Изд-во ИМ СО РАН, 1999.

Поступила в редакцию 1 июня 2007 г.