

## Обнаружение закономерностей в массивах экспериментальных данных\*

Н. Г. ЗАГОРУЙКО<sup>1,2,3</sup>, И. А. БОРИСОВА<sup>1,2,3</sup>, О. А. КУТНЕНКО<sup>1,2,3</sup>, Д. А. ЛЕВАНОВ<sup>1,3</sup>

<sup>1</sup>*Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия*

<sup>2</sup>*Конструкторско-технологический институт вычислительной техники СО РАН,  
Новосибирск, Россия*

<sup>3</sup>*Новосибирский государственный университет, Россия*

e-mail: zag@math.nsc.ru

Приведены методы автоматического поиска закономерностей, скрытых в экспериментальных данных, основанные на построении их сжатого описания и очистки от шумов. При этом результаты наблюдений представляются в компактном виде, удобном для восприятия человеком-экспертом и решения классических задач анализа данных.

*Ключевые слова:* поиск закономерностей, анализ данных, выбор информативных признаков, цензурирование, конкурентное сходство.

### Введение

Методы автоматического поиска закономерностей, скрытых в экспериментальных данных, часто основаны на построении сжатого описания этих данных, а также на очистке их от шумов [1]. В результате из всего множества наблюдаемых признаков выделяется небольшое число наиболее информативных, а из множества наблюдаемых объектов — подмножество наиболее типичных представителей. Этот подход позволяет сокращать объёмы сохраняемой информации, представлять результаты наблюдений в виде, удобном для восприятия человека-эксперта, а также решать классические задачи анализа данных, такие как таксономия, распознавание и прогнозирование.

В общем виде в данной работе решаются две задачи. Одна состоит в цензурировании обучающей выборки, т. е. в выборе из исходного множества  $M$  объектов такого подмножества  $m < M$  объектов, которое в заданном пространстве  $N$  признаков отражает основные закономерности распределения лучше, чем все  $M$  объектов, и обеспечивает более высокую надёжность распознавания контрольной выборки. Вторая задача заключается в цензурировании множества признаков, т. е. в выборе из исходного множества  $N$  признаков такого подмножества  $n < N$  признаков, в пространстве которых  $M$  объектов обучающей выборки отражают основные закономерности распределения лучше, чем в исходном признаковом пространстве.

Эти две задачи могут решаться как по отдельности, независимо друг от друга, так и согласованно, в рамках одной задачи комбинированного типа. Полученная подтаблица размерности  $m$  на  $n$  может использоваться для классификации новых объектов с опорой на эталонные образцы в пространстве информативных признаков. В задаче заполнения

---

\*Работа выполнена при финансовой поддержке РФФИ (грант № 11-01-00156).

пробелов в таблицах объект—свойство и объект—свойство—время такие компетентные подтаблицы строятся отдельно для каждого прогнозируемого объекта и признака.

Для решения сформулированных выше задач предлагается использовать новую меру оценки сходства между объектами в виде функции конкурентного сходства (FRiS-функции) [2, 3], что позволяет оценивать компактность экспериментальных данных и при необходимости повышать её, исключая шумящие и дублирующие признаки и объекты. Нами разработаны алгоритмы FRiS-GRAD [4] для выбора признаков и FRiS-Censor для цензурирования объектов. Эти алгоритмы работают при любом распределении объектов и любом соотношении между числом объектов и признаков, позволяют глубже понять связи между объектами и признаками и показали свою высокую эффективность при решении важных прикладных задач из области медицины, криминалистики и пр.

## 1. Функция конкурентного сходства

При измерении таких характеристик объекта, как вес, длина, сопротивление и т. д., обычно используется сравнение изучаемого объекта с эталонным. Результат измерения определяется свойствами только этих двух объектов и не зависит от свойств других объектов. По этой причине результат измерения имеет характер абсолютной величины. Но объекты могут описываться и такими характеристиками, как “похож—не похож”, “близок—далёк”, “добрый—злой” и т. д. Эталонов для подобных характеристик не существует, два объекта с несовпадающими свойствами могут считаться “сходными” или “не сходными”, “близкими” или “далекими” в зависимости от свойств других объектов. Адекватная мера сходства должна определять величину сходства, зависящую от особенностей конкурентного окружения объекта  $z$ .

При распознавании принадлежности объекта  $z$  к одному из двух образов  $A$  или  $B$  важно знать не только расстояние  $r(z, A)$  от  $z$  до образа  $A$ , но и расстояние  $r(z, B)$  от  $z$  до конкурирующего образа  $B$ . Следовательно, сходство в распознавании образов является категорией не абсолютной, а относительной.

Для вычисления сходства между объектами предлагается использовать тернарную относительную меру, которая оценивает сходство объекта  $z$  с объектом  $a$  в конкуренции с объектом  $b$ , которая называется функцией конкурентного сходства или FRiS-функцией:

$$F(z, a|b) = \frac{r(z, b) - r(z, a)}{r(z, b) + r(z, a)}. \quad (1)$$

По мере передвижения объекта  $z$  от объекта  $a$  к объекту  $b$  можно говорить вначале о большом сходстве объекта  $z$  с объектом  $a$ , об умеренном их сходстве, затем о наступлении одинакового сходства, равного 0, с объектом как  $a$ , так и  $b$ . При дальнейшем продвижении  $z$  к  $b$  возникает умеренное, а затем и большое отличие  $z$  от  $a$ . Совпадение объекта  $z$  с объектом  $b$  означает максимальное отличие  $z$  от  $a$ , что соответствует величине конкурентного сходства  $z$  с  $a$ , равной  $-1$ .

Конкурентное сходство объектов с образами будем определять по тому же принципу, что и конкурентное сходство объектов с объектами:

$$F(z, A|B) = \frac{r(z, B) - r(z, A)}{r(z, B) + r(z, A)}.$$

В случае нормально распределённых образов с одинаковыми матрицами ковариации конкурентное сходство объекта с данными образами можно вычислять через сходство

с их средними. Если же образы имеют очень сложную структуру, то ориентироваться при вычислении FRiS-функции можно лишь на локальную окрестность (ближайших соседей) того объекта, для которого она вычисляется.

Одним из способов проявить особенности данных в задаче распознавания является переход к их сжатому описанию с использованием ограниченного множества типичных (или эталонных) представителей каждого образа, сохраняющего основные закономерности распределения образов. Такие эталонные объекты далее будем называть столпами. Чем сложнее структура образов и чем сильнее они пересекаются, тем больше столпов потребуется для описания таких образов. Если удастся перейти от образов  $A$  и  $B$  к множеству их столпов  $S_A$  и  $S_B$ , то вычислять конкурентное сходство объекта  $z$  с образом  $A$  в конкуренции с образом  $B$  можно, как  $F(z, S_A|S_B)$ , сравнивая  $z$  с двумя ближайшими столпами, принадлежащими разным образам.

## 2. Выбор эталонных объектов. Алгоритм FRiS-Stolp

Для построения сжатого описания данных в виде системы столпов используется алгоритм FRiS-Stolp [4]. Алгоритм работает при любом отношении количества объектов к количеству признаков и произвольном виде распределения образов. В качестве столпов выбираются объекты, которые обладают высокими значениями двух свойств: обороноспособностью по отношению к объектам своего образа и толерантностью по отношению к объектам других образов. Чем выше обороноспособность эталона, тем меньше будет ошибок первого рода (пропуск цели). Чем выше толерантность эталона, тем меньше будет ошибок второго рода (ложная тревога). Набор столпов считается достаточным для описания выборки, если конкурентное сходство  $F$  всех объектов обучающей выборки с ближайшими своими столпами в конкуренции с ближайшими объектами других образов превышает пороговое значение  $F^*$ , например,  $F^* = 0$ .

С помощью рис. 1 проиллюстрируем методику вычисления оценок толерантности и обороноспособности объекта на примере задачи распознавания двух образов  $A = \{a_1, \dots, a_{M_A}\}$  и  $B = \{b_1, \dots, b_{M_B}\}$ , представленных соответственно наборами из  $M_A$  и  $M_B$  объектов обучающей выборки.

Проверим, хорошо ли объект  $a_i$  защищает объекты  $a_j$ ,  $j = 1, \dots, M_A$ , образа  $A$ . Для объекта  $a_j$  определим расстояния  $r(a_j, a_i)$  и  $r(a_j, b_{j'})$ , где  $b_{j'} \in B$  является ближайшим

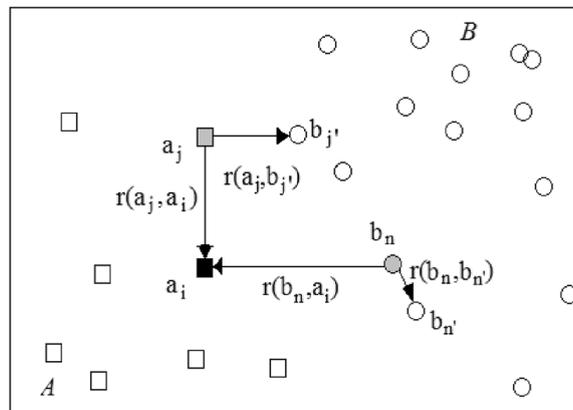


Рис. 1. Оценка обороноспособности и толерантности объекта  $a_i \in A$

соседом объекта  $a_j$ , т. е.  $j' = \arg \min r(a_j, b_m) | m = 1, \dots, M_B$ , и по (1) получим значение  $F(a_j, a_i | b_{j'})$  функции сходства объекта  $a_j$  с  $a_i$  в конкуренции с  $b_{j'}$  (см. рис. 1). Выделим объекты  $a_j \in A$ ,  $j = 1, \dots, M_A$ , сходство которых с  $a_i$  не меньше заданного порога  $F^*$ , т. е.  $F_j^+ = F(a_j, a_i | b_{j'}) - F^* \geq 0$ . Эти объекты надёжно защищены  $a_i$ . Получим оценку  $D$  обороноспособности объекта  $a_i$ :

$$D(a_i) = \sum_{j=1}^{M_A} F_j^+ | F_j^+ \geq 0.$$

Теперь оценим толерантность  $a_i$ , т. е. меру несходства с  $a_i$  объектов образа  $B$ . Для каждого  $b_n \in B$ ,  $n = 1, \dots, M_B$ , вычислим расстояния  $r(b_n, a_i)$  и  $r(b_n, b_{n'})$ , где  $b_{n'} \in B$  — ближайший сосед  $b_n$ . По (1) найдем величину сходства  $F(b_n, b_{n'} | a_i)$  объекта  $b_n$  с  $b_{n'}$  в конкуренции с  $a_i$  (см. рис. 1). Выделим те объекты образа  $B$ , у которых  $F_n^- = F(b_n, b_{n'} | a_i) - F^* < 0$ . Эти объекты больше похожи на  $a_i$ , чем на ближайшие объекты своего образа, что отрицательно влияет на оценку  $a_i$ . Получим оценку  $T$  “нетолерантности” объекта  $a_i$ :

$$T(a_i) = \sum_{j=1}^{M_B} F_j^- | F_j^- < 0.$$

Качество выполнения объектом  $a_i$  роли столпа образа  $A$  оценивается величиной

$$S(a_i) = D(a_i) - T(a_i).$$

Отметим некоторые особенности алгоритма FRiS-Stolp. Вне зависимости от вида распределения обучающей выборки столпами выбираются объекты, расположенные в центрах локальных сгустков и защищающие максимально возможное количество объектов с заданной надёжностью. При нормальных распределениях столпами в первую очередь будут выбраны объекты, ближайшие к точкам математического ожидания. Следовательно, при приближении закона распределения к нормальному решение задачи построения решающих функций стремится к статистически оптимальному. Если распределения полимодальны и образы линейно неразделимы, столпы будут стоять в центрах мод.

Сжатое описание образов через множество столпов можно использовать для распознавания новых объектов.

### 3. FRiS-компактность

Для произвольного объекта  $a \in A$  мера его конкурентного сходства со своим образом в конкуренции с образом  $B$  показывает, насколько этот объект похож на свой образ и не похож на образ  $B$ . Если эта величина для всех объектов образа  $A$  положительна, то можно считать данный образ компактным, так как подобная ситуация хорошо согласуется с интуитивным представлением о компактности как о сходстве объектов внутри образа и несходстве их с объектами конкурирующего образа.

Вычисляя среднее значение FRiS-функции по всем объектам образа  $A$ , можно оценить компактность данного образа. Если при этом вычислять FRiS-функцию с опорой на столпы, то такая оценка компактности будет автоматически адаптироваться к особенностям данных.

В случае двух образов  $A = a_1, \dots, a_{M_A}$  и  $B = b_1, \dots, b_{M_B}$  предлагается следующий вариант оценки компактности (далее — компактности).

1. С помощью алгоритма FRiS-Stolp строятся  $c$  столпов образов  $A$  и  $B$ :  $c = c_A + c_B$ , где  $c_A$  и  $c_B$  — число столпов образов  $A$  и  $B$  соответственно.

2. Для каждого элемента  $a \in A$  оценивается сходство со своим ближайшим столпом  $s_A(a_i)$  в конкуренции с ближайшим столпом  $s_B(a_i)$  образа  $B$ . Затем вычисляется компактность образа  $A$  в конкуренции с образом  $B$ :

$$C_{A|B} = \frac{1}{c_A M_A} \left( \sum_{i=1}^{M_A} F(a_i, s_A(a_i) | s_B(a_i)) - c_A \right). \quad (2)$$

3. Также вычисляется величина  $C_{B|A}$  компактности образа  $B$  в конкуренции с  $A$ .

4. Далее получим оценку компактности образов  $A$  и  $B$  путём усреднения величин  $C_{A|B}$  и  $C_{B|A}$ .

Отметим, что количество столпов  $c_A$  образа  $A$  зависит от структуры распределения объектов и величины порога  $F^*$ . С ростом  $F^*$  увеличиваются количество столпов и точность описания распределения, но растёт и сложность его описания, т. е. множитель  $c_A$  является штрафом за структурную сложность образа.

Если количество образов  $K$  больше двух, то при оценке компактности образа  $A_k$ ,  $k \in 1, \dots, K$ , объекты всех остальных образов объединяются в один виртуальный образ  $B_k$ . После получения оценок компактности  $C_{A_k|B_k}$ ,  $k = 1, \dots, K$ , всех образов общая оценка их компактности в данном признаковом пространстве может быть получена путём арифметического усреднения данных оценок:

$$C = \frac{1}{K} \sum_{k=1}^K C_{A_k|B_k}.$$

Если мы стремимся к тому, чтобы компактность самого некомпактного образа была максимально большой, то нужно использовать среднегеометрическую оценку:

$$C = \sqrt[k]{\prod_{k=1}^K C_{A_k|B_k}}. \quad (3)$$

Наши эксперименты с этими двумя критериями FRiS-компактности показали существенное преимущество второго из них [5].

#### 4. Повышение компактности данных путём цензурирования объектов. Алгоритм FRiS-Censor

Упрощение описания и повышение компактности можно получить за счёт “очистки” данных, устраняя “нетипичные” объекты, которые искажают представление о выборке и влияют на выбор эталонов. Значительные отличия свойств таких объектов от свойств остальных объектов образа могут объясняться их уникальностью, однако чаще причина отличий состоит во влиянии неучитываемых факторов, таких как сбой измерительных приборов, ошибки занесения данных в протокол и пр. Встречаются и объекты, которые

не являются “ошибочными”, но находятся на периферии распределения и оказываются глубоко в зоне пересечения с соседними образами. Они также могут неоправданно сильно усложнять решающие правила.

Для исключения (цензурирования) выбросов можно применять алгоритм FRiS-Censor [6], который включает в себя как составную часть алгоритм FRiS-Stolp и использует меру FRiS-компактности образов в качестве критерия, управляющего процессом повышения компактности данных.

Пусть даны два образа  $A$  и  $B$ , представленные наборами из  $M_A$  и  $M_B$  объектов,  $M = M_A + M_B$ . Оценка компактности образов  $A$  и  $B$  вычисляется по формулам (2), (3). Через  $M^*$  обозначим число объектов обучающей выборки, оставшихся после очередного этапа сокращения выборки. Величину  $(M^*/M)^\alpha$ ,  $\alpha \geq 0$ , будем использовать в качестве штрафа за исключение объектов из обучающей выборки. С учётом этого FRiS-компактность  $H_{A|B}$  образов на каждом шаге сокращения выборки будем оценивать следующим образом:

$$H_{A|B} = (M^*/M)^\alpha \sqrt{C_{A|B}C_{B|A}}.$$

Выбор оптимального значения параметра  $\alpha$  осуществляется методом машинного моделирования путём сравнения результатов работы алгоритма FRiS-Censor при разных значениях  $\alpha$ . Через  $d \in [0, 1]$  обозначим максимальную долю объектов обучающей выборки, которые можно исключить,  $m^*$  — максимальное количество объектов в удаляемом кластере.

Алгоритм FRiS-Censor тестировался на модельной задаче распознавания двух образов, каждый из которых представлял собой суперпозицию нескольких (от двух до четырёх) нормально распределённых кластеров в двумерном пространстве признаков.

Эксперименты показали, что повышение FRiS-компактности обучающей выборки путём цензурирования объектов более чем в 99% случаев приводит к повышению качества распознавания. Очищенная выборка описывается более простым решающим правилом, что повышает надёжность распознавания контрольной выборки. На рис. 2 показаны гистограммы распределения надёжности  $P$  распознавания контрольной выборки. По оси ординат отложено абсолютное число экспериментов  $N$  (из 1000), в которых была достигнута данная надёжность  $P$ . Кривая 1 соответствует надёжности без предварительного повышения FRiS-компактности, среднее значение равно 91.6%, а кривая 2 — надёжности с использованием процесса повышения FRiS-компактности. Здесь среднее

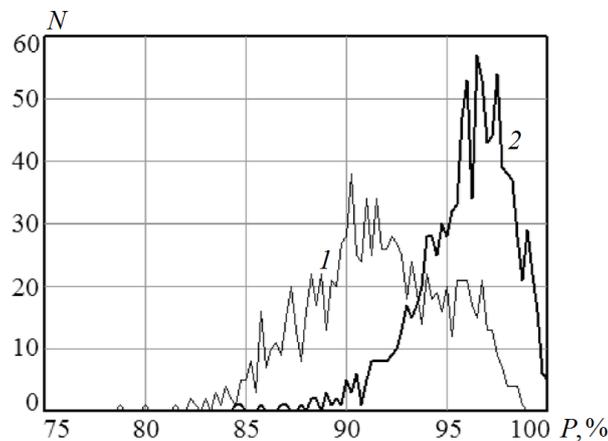


Рис. 2. Распределение надёжности  $P$  распознавания контрольной выборки

значение равно 95.9%. Среднее значение доли исключённых объектов  $d^*$ , при котором достигалось максимальное значение критерия  $H_{A|B}$ , было равно 12.7%.

## 5. Исключение неинформативных признаков.

### Алгоритм FRiS-GRAD

Упрощение описания и повышение компактности данных можно получить, устраняя малоинформативные признаки. В настоящее время преобладают задачи, у которых количество признаков  $N$  на порядки превышает количество объектов  $M$ . При этом информация, полезная для решения конкретной классификационной задачи, обычно представлена в нескольких признаках  $n \ll N$ . Выбор этих  $n$  признаков не только позволяет в дальнейшем существенно сократить затраты машинных ресурсов, но и повышает компактность образов и надёжность их распознавания. Признаки могут зависеть друг от друга, что не позволяет по оценкам информативности каждого признака в отдельности выбрать подмножество в виде списка из  $n$  наиболее информативных признаков. Точное решение можно получить, проверив все сочетания из  $N$  признаков по  $n$ , что в реальных задачах практически невозможно. По этой причине используются эвристические алгоритмы направленного перебора.

В разработанном нами алгоритме GRAD [3] используется следующий приём. Сначала методом полного перебора формируются информативные системы признаков (гранулы) малой размерности, а затем эти гранулы используются в качестве входных элементов — обобщённых признаков для алгоритма AdDel [3], который представляет собой комбинацию двух известных жадных алгоритмов Addition и Deletion. Эти алгоритмы дают оптимальное решение на каждом шаге, но не обеспечивают глобального оптимума.

В алгоритме AdDel в процессе прямого хода (алгоритм Addition) набирается некоторое количество информативных признаков (гранул признаков) и затем часть из них исключается в процессе обратного хода (алгоритм Deletion). Такое чередование алгоритмов Addition и Deletion продолжается до достижения заданного количества признаков (гранул признаков)  $n$ . В алгоритме FRiS-GRAD, использующем вместо отдельных признаков гранулы, в полученной системе некоторые признаки могут встретиться более одного раза.

Как показали эксперименты, по мере увеличения числа признаков качество распознавания вначале растёт, потом рост прекращается и начинается его снижение за счёт добавления неинформативных (шумящих) признаков. Перегиб кривой качества позволяет автоматически определить количество признаков в системе.

Информативность признака или системы признаков может оцениваться разными способами. От того, насколько универсальный и релевантный задаче критерий используется, зависит качество решения этой задачи. В алгоритме выбора признаков FRiS-GRAD [4] в качестве критерия информативности используется величина FRiS-компактности. Такой критерий применим к любому виду распределений и любому соотношению между  $M$  и  $N$ .

В процессе вычисления FRiS-компактности одновременно происходит выбор системы столпов. Поэтому алгоритм FRiS-GRAD может интерпретироваться как алгоритм сжатия данных за счёт сокращения числа объектов в выборке и числа описывающих выборку признаков. Это сокращённое описание выборки в виде множества столпов в пространстве информативных признаков далее может использоваться для решения

задачи распознавания. Объект  $z$  относится к образу, сходство со столпом которого в пространстве выбранных информативных признаков оказалось наибольшим, а величина сходства рассматривается как вероятность правильности принятого решения.

## 6. Экспериментальная проверка

В работе [7] представлены результаты сравнения десяти наиболее известных методов выбора признаков. Сравнение делалось по результатам решения девяти задач из области генетики. Для каждой выбранной системы признаков строились решающие правила четырёх наиболее известных типов. В результате каждая из девяти задач получала сорок различных решений. Мы выбрали лучшие из них (рекордные результаты) и сравнили их с результатами, полученными комбинацией алгоритма выбора признаков FRiS-GRAD с алгоритмом построения решающего правила FRiS-Stolp.

Результаты сравнительных экспериментов представлены в табл. 1. Здесь показаны имена задач, размерность признакового пространства  $N$ , количество объектов первого ( $M1$ ) и второго ( $M2$ ) образов и два столбца результатов. В последнем столбце показано место, занятое результатами решения всех девяти задач FRiS-методами.

Задачи решались в [7] с использованием десяти методов выбора признаков. Для каждой задачи по результату, полученному каждым методом, можно указать его рейтинг:

Т а б л и ц а 1. Результаты решения девяти задач

Задача	$N$	$M1/M2$	max of 40	FRiS-GRAD
ALL1	12625	95/33	100.0	100.0
ALL2	12625	24/101	78.2	88.0
ALL3	12625	65/35	59.1	87.2
ALL4	12625	26/67	82.1	87.4
Prostate	12625	50/53	90.2	95.1
Myeloma	12625	36/137	82.9	95.9
ALL/AML	7129	47/25	95.9	100.0
DLBCL	7129	58/19	94.3	95.0
Colon	2000	22/40	88.6	98.6
Среднее значение			85.7	94.1

Т а б л и ц а 2. Суммы рейтинговых мест, занятых методами выбора признаков

Methods of feature selection	Rating
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
maxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	9

Т а б л и ц а 3. Сумма рейтинговых мест, полученных решающими правилами

Решающее правило	Рейтинг
Between group analysis (BGA)	35
K-nearest neighbours (kNN)	32
Naive bayes classification (NBC)	25
Support vector machines (SVM)	19
FRiS-Stolp	9

лучший результат занимает первое место, худший — десятое. Если просуммировать места, занятые методом на всех задачах, то можно определить его общий рейтинг. Результаты таких подсчётов представлены в табл. 2, в последней строке которой показана сумма рейтинговых мест, занятых FRiS-методом.

Такой же анализ был проведён и по четырём использованным решающим правилам. Его результаты показаны в табл. 3, в которой так же, как и в табл. 2, чем меньше сумма рейтинговых мест, тем лучше.

Эти эксперименты показывают, что функция конкурентного сходства оказалась эффективным инструментом как для выбора из всего множества наблюдаемых признаков небольшого числа наиболее информативных, так и для выбора из множества наблюдаемых объектов — подмножества наиболее типичных представителей. Предложенные алгоритмы, основанные на FRiS-компактности, эффективны при решении сложных задач когнитивного анализа данных.

## Список литературы

- [1] АРКАДЬЕВ А.Г., БРАВЕРМАН Э.М. Обучение машины распознаванию образов. М.: Наука, 1964.
- [2] ВОРОНИН Ю.А. Введение мер сходства и связи для решения геолого-географических задач // Докл. АН СССР. 1971. Т. 199, № 5. С. 1011–1015.
- [3] ЗАГОРУЙКО Н.Г. Когнитивный анализ данных. Новосибирск: Акад. изд-во ГЕО, 2013.
- [4] BORISOVA I.A., DYUBANOV V.V., ZAGORUIKO N.G., KUTNENKO O.A. Use of the FRiS-function for taxonomy, attribute selection and decision rule construction // Collective monographic “Knowledge Processing and Data Analysis”. Springer, 2010. P. 257–271.
- [5] БОРИСОВА И.А., ДЮБАНОВА В.В., ЗАГОРУЙКО Н.Г., КУТНЕНКО О.А. Сходство и компактность // Математические методы распознавания образов: Тр. 14-й Всерос. конф. М., 2009. С. 89–92.
- [6] ЗАГОРУЙКО Н.Г., КУТНЕНКО О.А. Количественная мера компактности образов и метод её повышения // Интеллектуализация обработки информации: Сб. докл. 9-й Междунар. конф. М.: Торус Пресс, 2012. С. 29–32.
- [7] JEFFERY I., HIGGINS D., CULHANE A. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics, 2006, 7:359. ([http://www.biomedcentral.com/1471-2\[9\]5/7/359](http://www.biomedcentral.com/1471-2[9]5/7/359)).