# Обучение распознаванию без переобучения

Загоруйко Н. Г.[1,2], Кутненко О. А. [1,2], Зырянов А. О.[2],
Леванов Д. А.[1]

[1] *Институт математики им. С.Л. Соболева СО РАН, Новосибирск;*
[2] *Новосибирский государственный университет, Новосибирск*

10-я Международная конференция
«Интеллетуализация обработки информации»

4–11 октября, 2014, Греция, о. Крит

# Введение

**Предмет исследования —**

обучение алгоритмов распознавания.

**Цель исследования —**

построение алгоритма, автоматически выбирающего подмножество наиболее информативных объектов и признаков, и обнаруживающего момент начала переобучения.

**Мотивация исследования —**

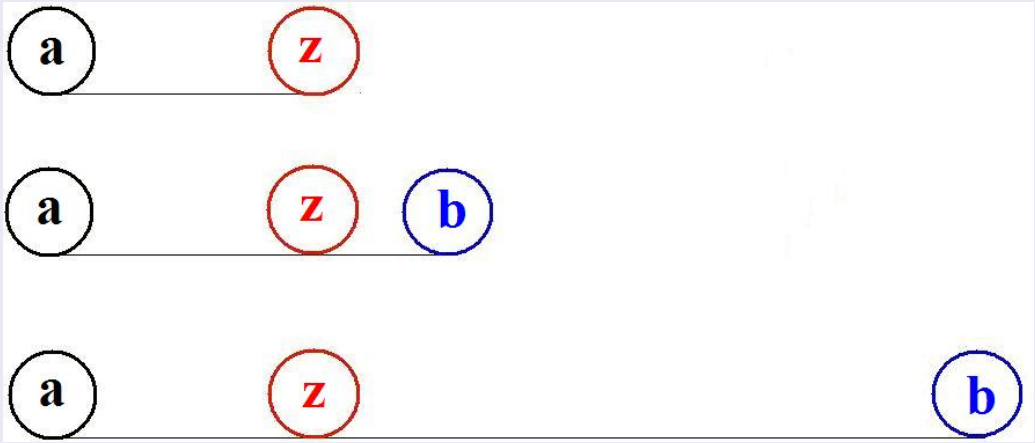
отсутствие алгоритмов, решающих проблему переобучения алгоритмов распознавания.

**Области приложений:**

Анализ данных и распознавание образов.

# Функция конкурентного сходства (FRiS-функция)
(Function of Rival Similarity)

Рис. 1. Иллюстрация относительности сходства объектов $a$ и $z$.




**Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Methods of recognition based on the function of rival similarity // Pattern Recognition and Image Analisys. V. 18, № 1. 1–6.

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

$F(z,a|b) \in [-1,1]$,
если $r(z,a) = r(z,b)$, то $F(z,a|b) = 0$,
$F(z,a|b) = -F(z,b|a)$.

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

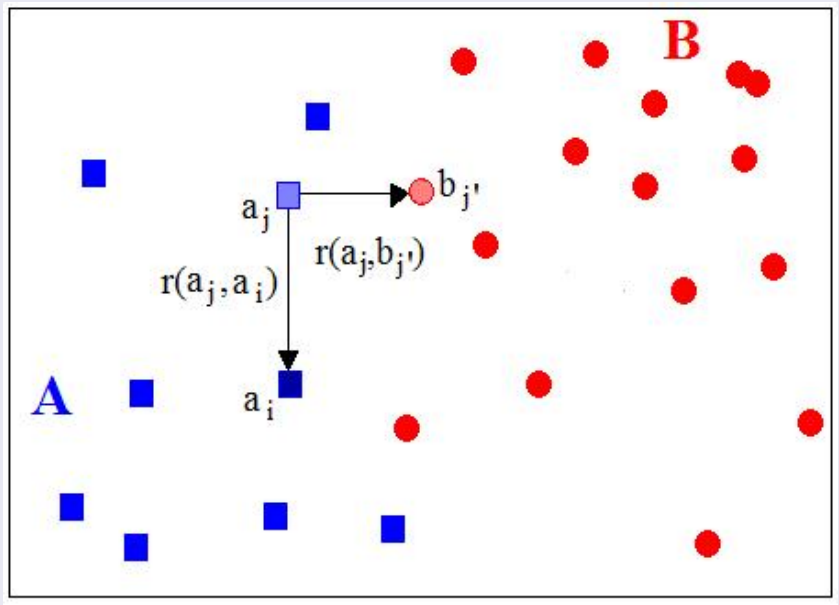
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

**Загоруйко Н. Г., Кутненко О. А.** Количественная мера компактности образов и метод ее повышения // 9-ая международная конференция «Интеллектуализация обработки информации», Республика Черногория, Будва: Торус Пресс, 2012. С. 29–32.

В качестве столпов выбираются объекты, обладающие высоким значением обороноспособности по отношению к объектам своего образа.

$A = \{a_1, \ldots, a_{M_A}\}$ и $B = \{b_1, \ldots, b_{M_B}\}$

Рис. 2. Оценка веса объекта $a_i \in A$.

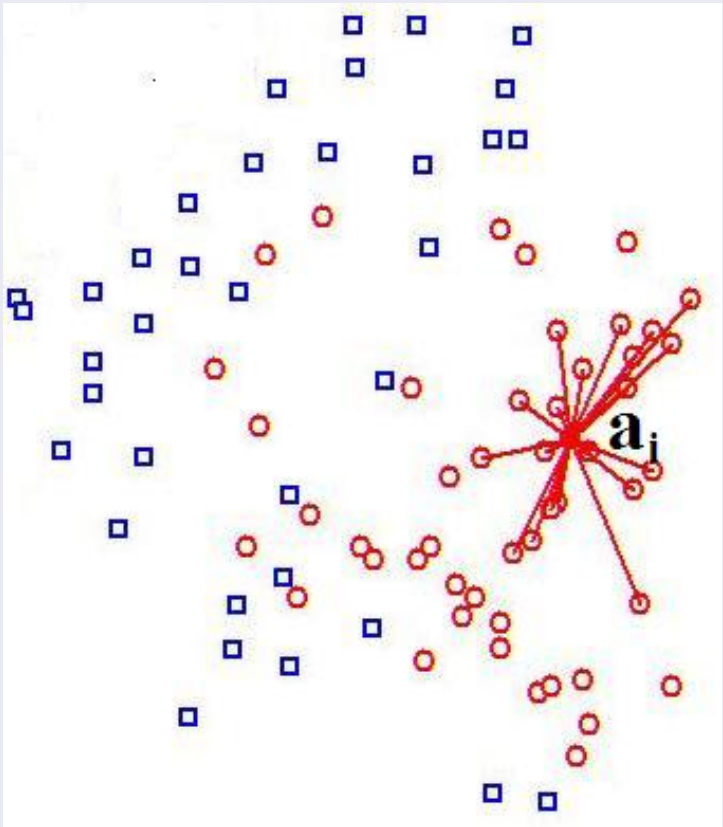



$F(a_j, a_i|b_{j'})$ - функция сходства объекта $a_j$ с $a_i \in A$ в конкуренции с $b_{j'} \in B$, $j' = \arg\min_{m=1,\ldots,M_B} r(a_j, b_m)$.

# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Объекты $a_j \in A$, $j = 1, \ldots, M_A$, сходство которых с $a_i$ не меньше заданного порога $F^*$, т. е. $F_j = F(a_j, a_i | b_{j'}) - F^* \geq 0$, образуют кластер.

Рис. 3. Пример кластера, образованного столпом $a_i \in A$.



# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Вес кластера – сумма сходств объектов, входящих в данный кластер, со своим столпом $a_i$ в конкуренции с ближайшим объектом другого класса:

$$V(a_i) = \sum_{j=1}^{M_A} F_j|_{F_j \geq 0} \tag{2}$$

– является оценкой качества исполнения объектом $a_i$ роли столпа класса $A$.

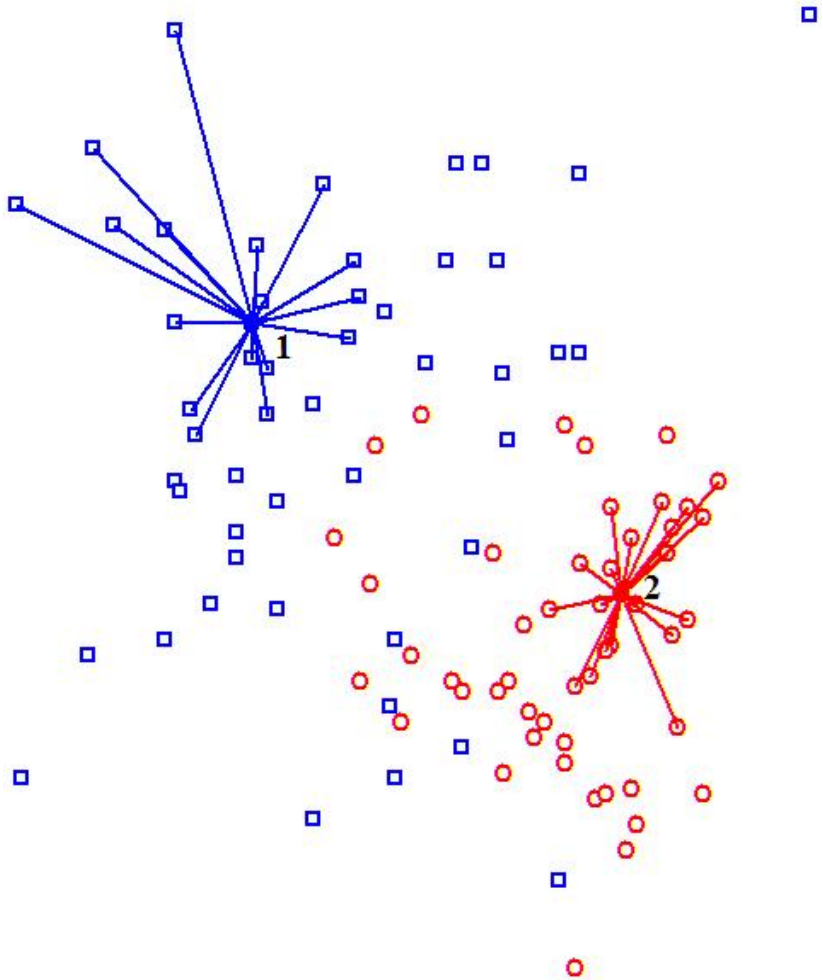
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Если в кластеры вошли не все объекты, то среди оставшихся («незащищенных») выбирается объект на роль третьего столпа. Третьим столпом назначается объект любого класса, сходство с которым незащищенных объектов этого класса в конкуренции с ближайшими объектами другого класса максимально.

Процесс уточнения описания выборки путем увеличения количества столпов продолжается до включения в кластеры всех объектов.

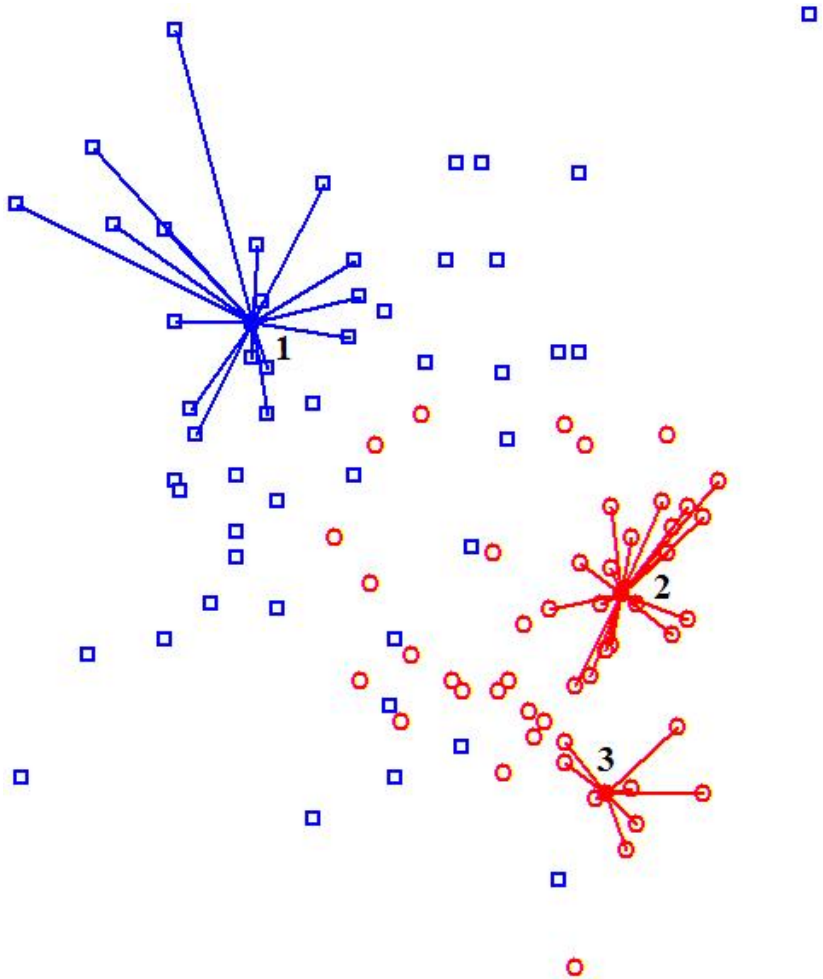
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 4. Иллюстрация работы алгоритма. Выбор 1 и 2 столпов.



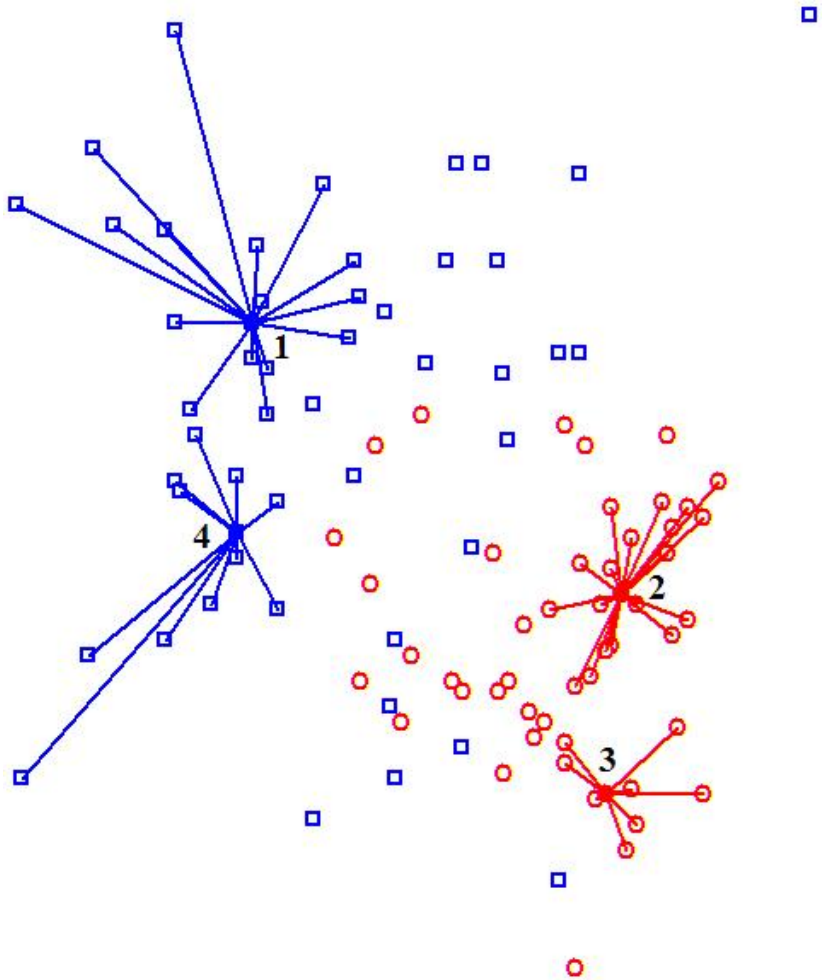
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 5. Иллюстрация работы алгоритма. Выбор 3 столпа.



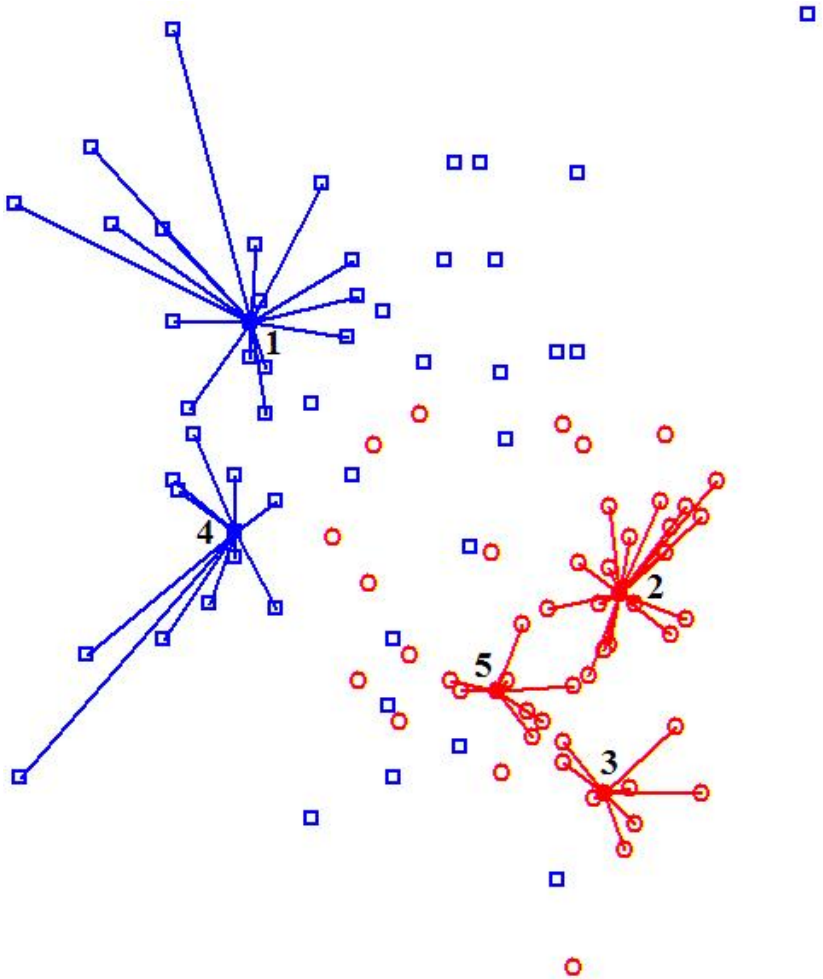
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 6. Иллюстрация работы алгоритма. Выбор 4 столпа.



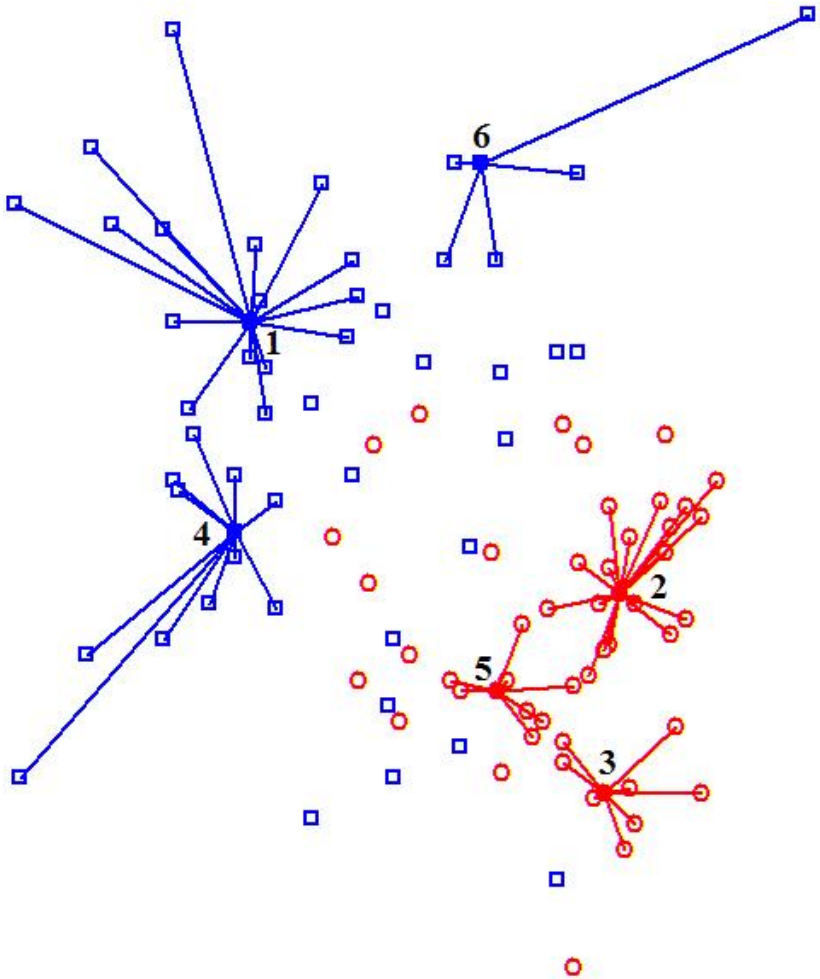
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 7. Иллюстрация работы алгоритма. Выбор 5 столпа.



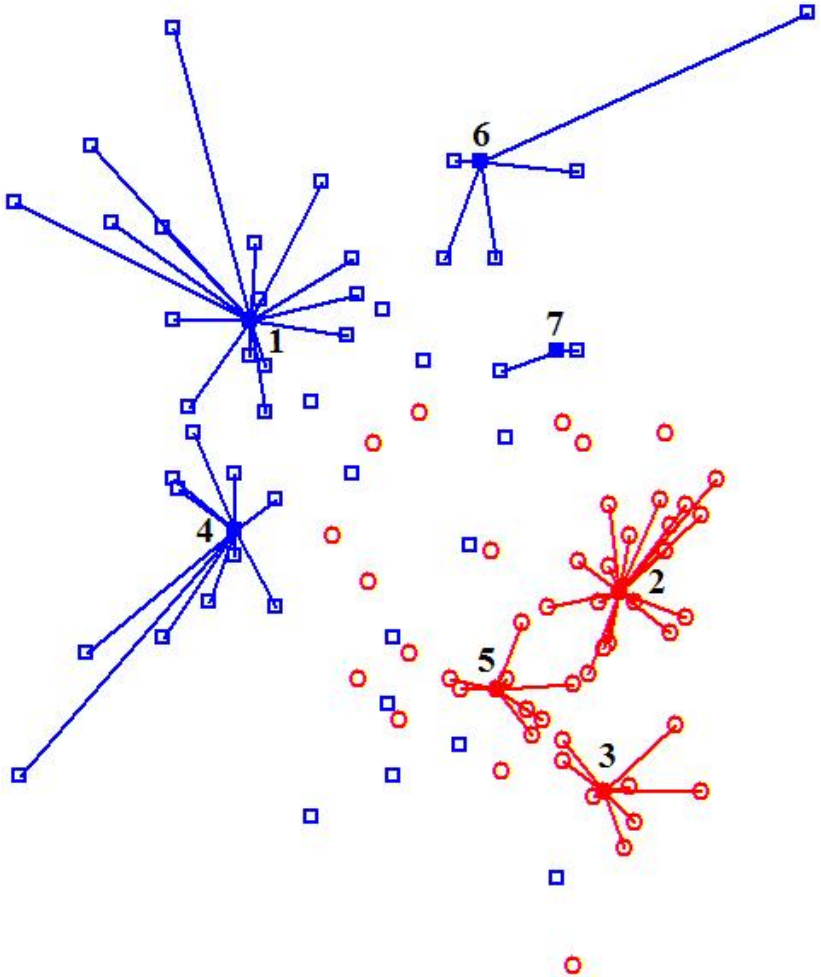
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 8. Иллюстрация работы алгоритма. Выбор 6 столпа.



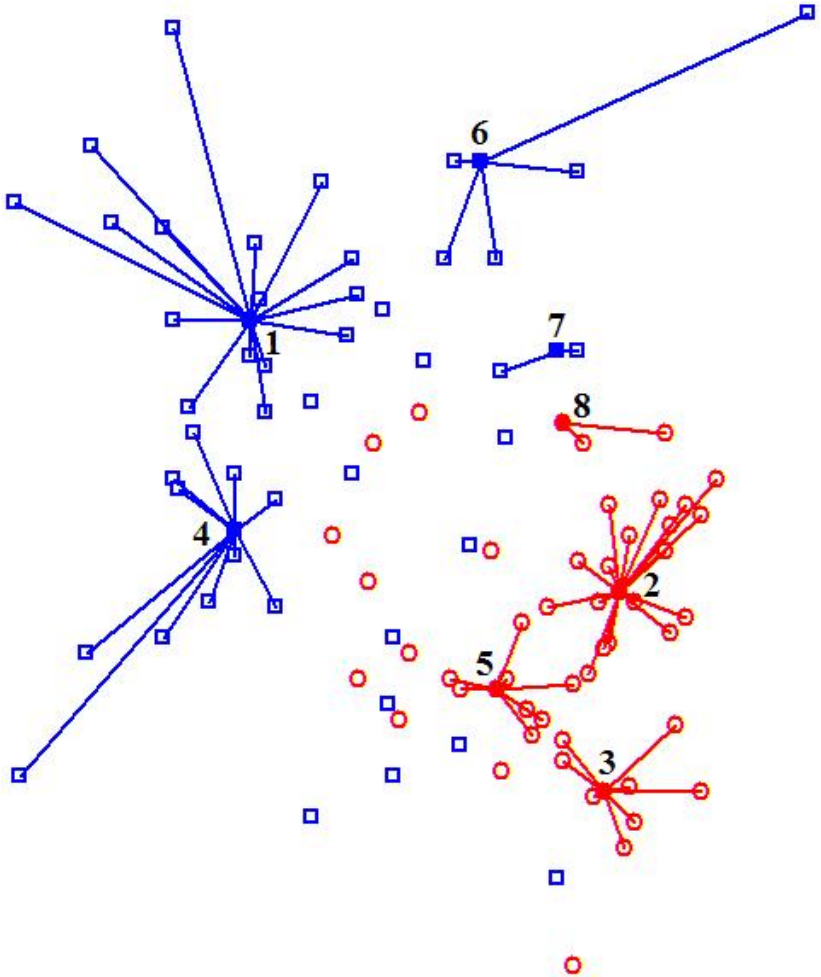
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 9. Иллюстрация работы алгоритма. Выбор 7 столпа.



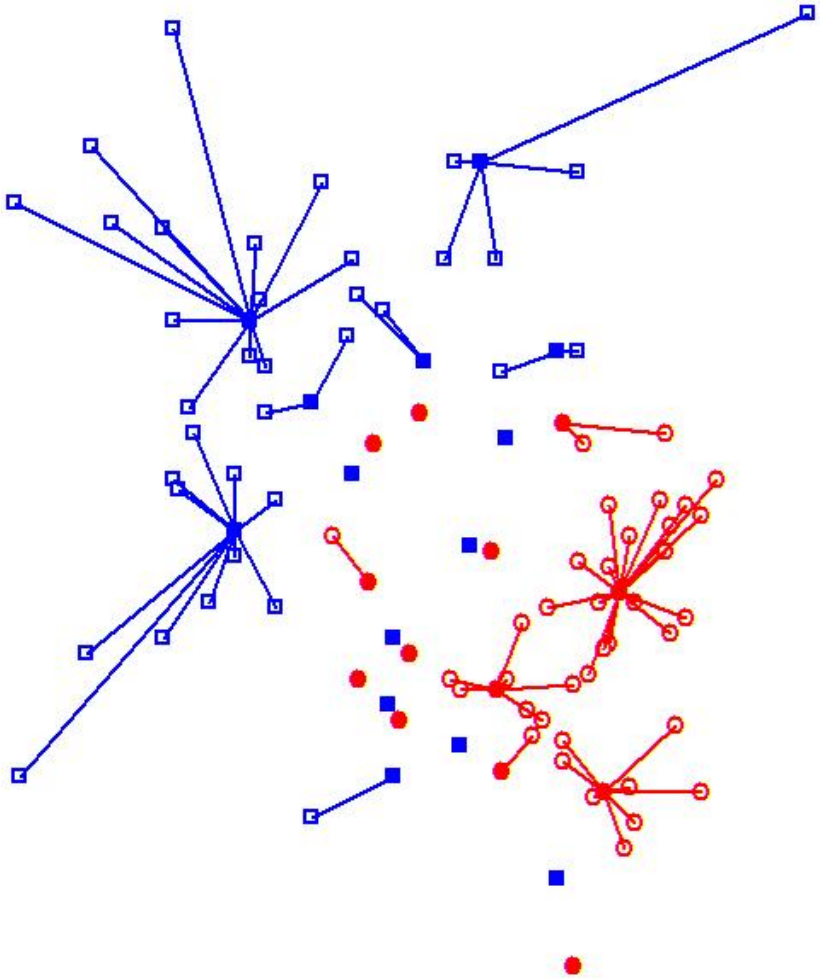
# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 10. Иллюстрация работы алгоритма. Выбор 8 столпа.



# Выбор эталонных объектов (алгоритм FRiS-Stolp)

Рис. 11. Иллюстрация работы алгоритма. Построение всех столпов.



# Качество описания выборки
(оценка разделимости классов)

**Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.** Сходство и компактность // Труды 14-й Всероссийской конференции «Математические методы распознавания образов». 2009. С. 89–92.

Качество описания обучающей выборки (или оценка разделимости классов) зависит от набора выбранных эталонов.

$A = \bigcup_{k=1}^{K} A_k = \{a_1, \ldots, a_M\}$ - обучающая выборка, состоящая из $M$ объектов, разделенных на $K$ классов.
Пусть $s_l(k) \in A$ — $l$-ый столп в описании выборки $L$ столпами, являющийся эталоном $k-$го класса, $k \in \{1, \ldots, K\}$;
Качество описания обучающей выборки $L$ столпами:

$$H(L) = \frac{K}{L \times M} \sum_{i=1}^{M} F(a_i(k), s_l(k) | a_{i'}(\overline{k})). \qquad (3)$$

$a_i(k) \in A-$ объект $k-$го класса;
$s_l(k)$ - ближайший к $a_i(k)$ эталон $k-$го класса;
$a_{i'}(\overline{k}) \in A \backslash A_k$, $i' = \arg\min_{m=1,\ldots,M, a_m \in A \backslash A_k} r(a_i, a_m)$.

# Распознавание контрольного объекта
## (принятие решения с учетом взвешенных расстояний)

Сжатое описание образов через множество столпов используется для распознавания новых объектов.

**Распознавание контрольного объекта** $z$**:** определяются взвешенные расстояния $r_l = r(z, s_l)/V(s_l)$ от $z$ до всех столпов $s_l$, $l = 1, \ldots, L$, описывающих обучающую выборку. Выбираются два минимальных значения $r_{l_1}$ и $r_{l_2}$ таких, что столпы $s_{l_1}$ и $s_{l_2}$ принадлежат разным классам. Объект $z$ считается принадлежащим классу, взвешенное расстояние до столпа которого оказалось меньше. По величине сходства $F(z, s_{l_1}|s_{l_2})$ можно судить о достоверности принятого решения.

# Обнаружение начала переобучения

При изменении количества столпов $L$ меняется качество описания $H$ обучающей выборки и надежность распознавания $P$ тестовой выборки.

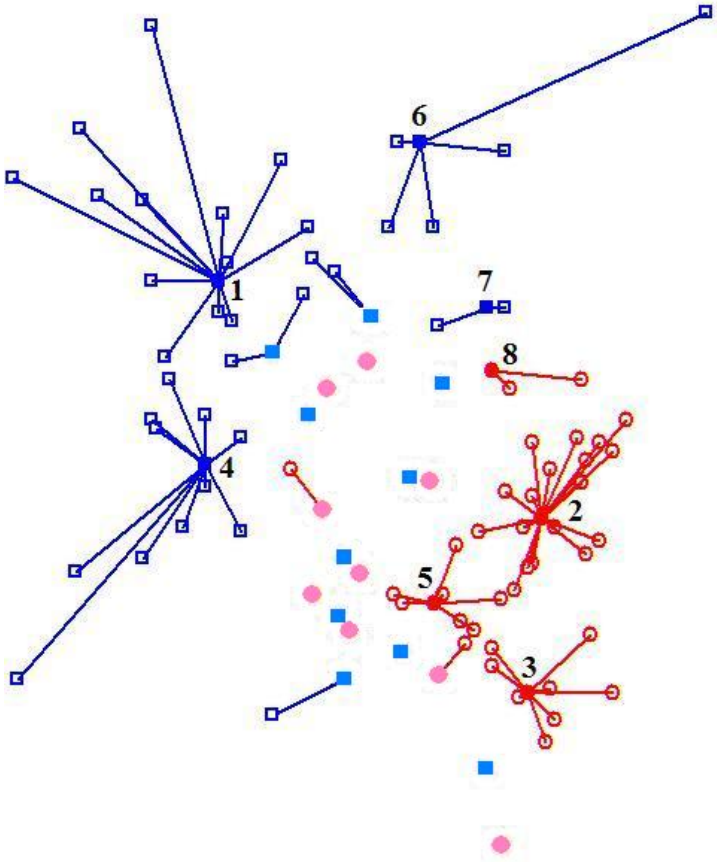
Выдвигается и проверяется гипотеза о том, что между функциями $H = f(L)$ и $P = f(L)$ имеется закономерная связь, используя которую можно найти такое количество столпов $l^*$, что дальнейшее увеличение числа столпов ведет к переобучению.

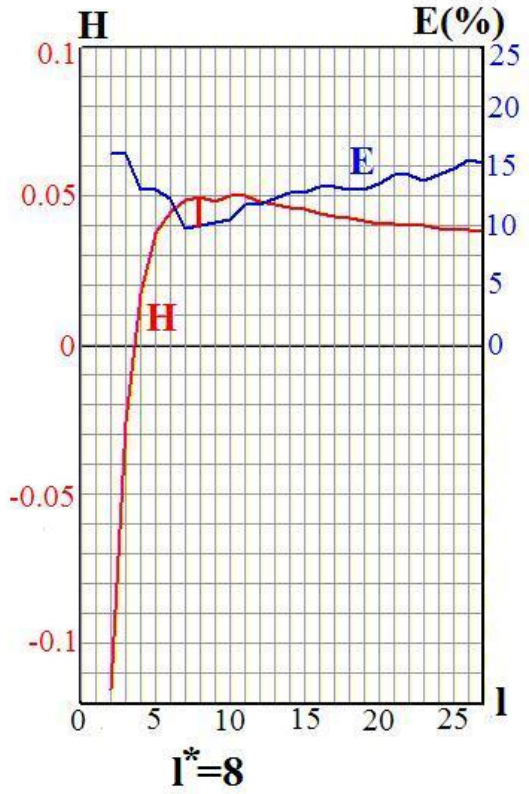
# Обнаружение начала переобучения
## (результаты экспериментов)

Тестирование проводилось на модельной задаче распознавания двух образов, каждый из которых представлял собой суперпозицию нескольких (от 2-х до 4-х) нормально распределенных кластеров в двумерном пространстве признаков. Рассматривалось 10 распределений, которые отличались друг от друга количеством образующих нормальных компонентов, их дисперсиями, координатами математических ожиданий и количеством объектов в компонентах. Каждый образ был представлен 250 объектами. При каждом распределении выборка 100 раз случайным способом делилась на две части: обучающую (по 50 объектов первого и второго образов), и контрольную (по 200 объектов каждого образа). Количество экспериментов при различных численных реализациях исходных данных было равно 1000.

# Обнаружение начала переобучения
(результаты экспериментов)

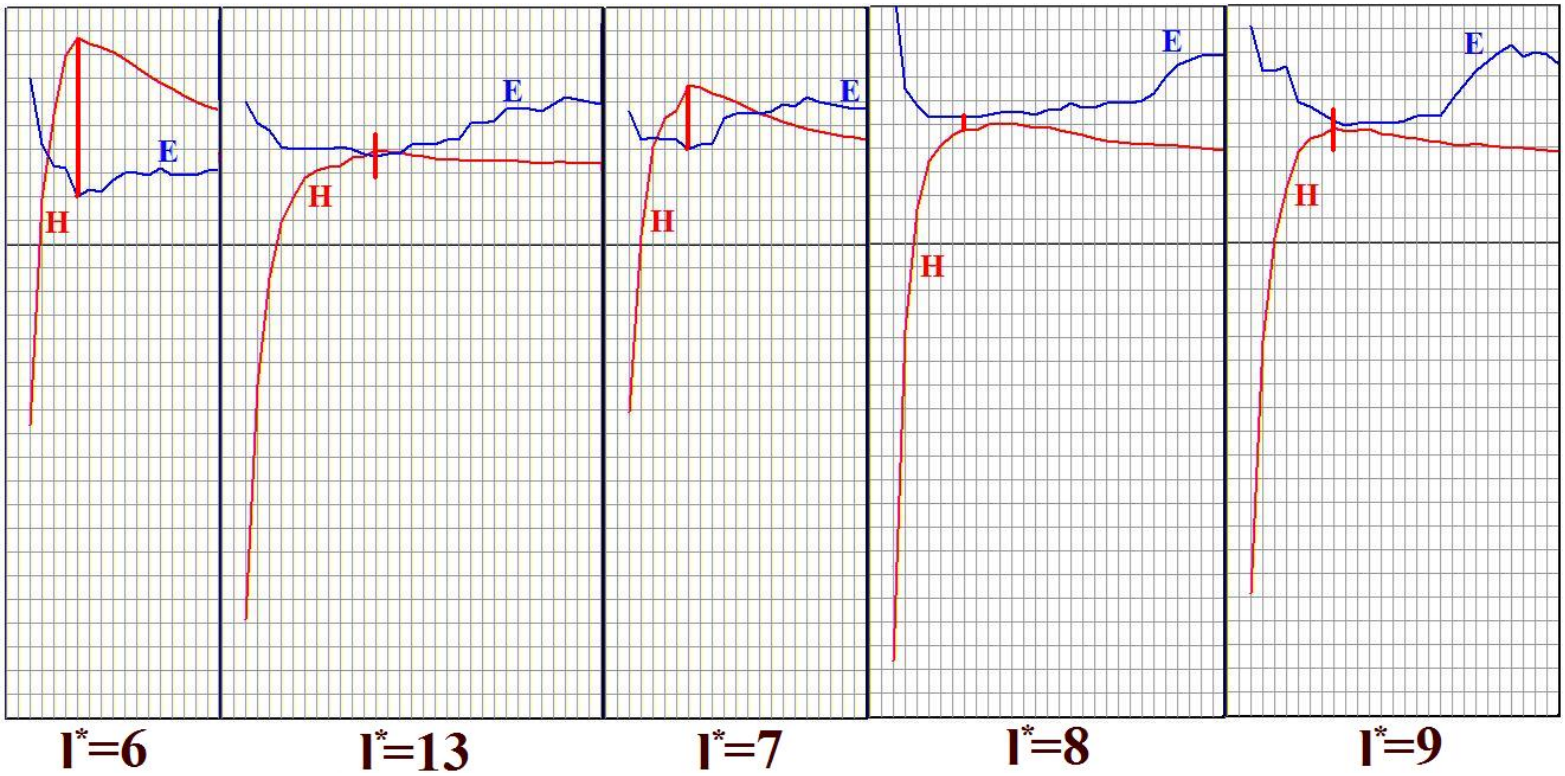
Рис. 12. График качества описания обучающей выборки - кривая $H$ и график ошибки распознавания - кривая $E$ в зависимости от $l$ - числа выбранных эталонов для данного множества.



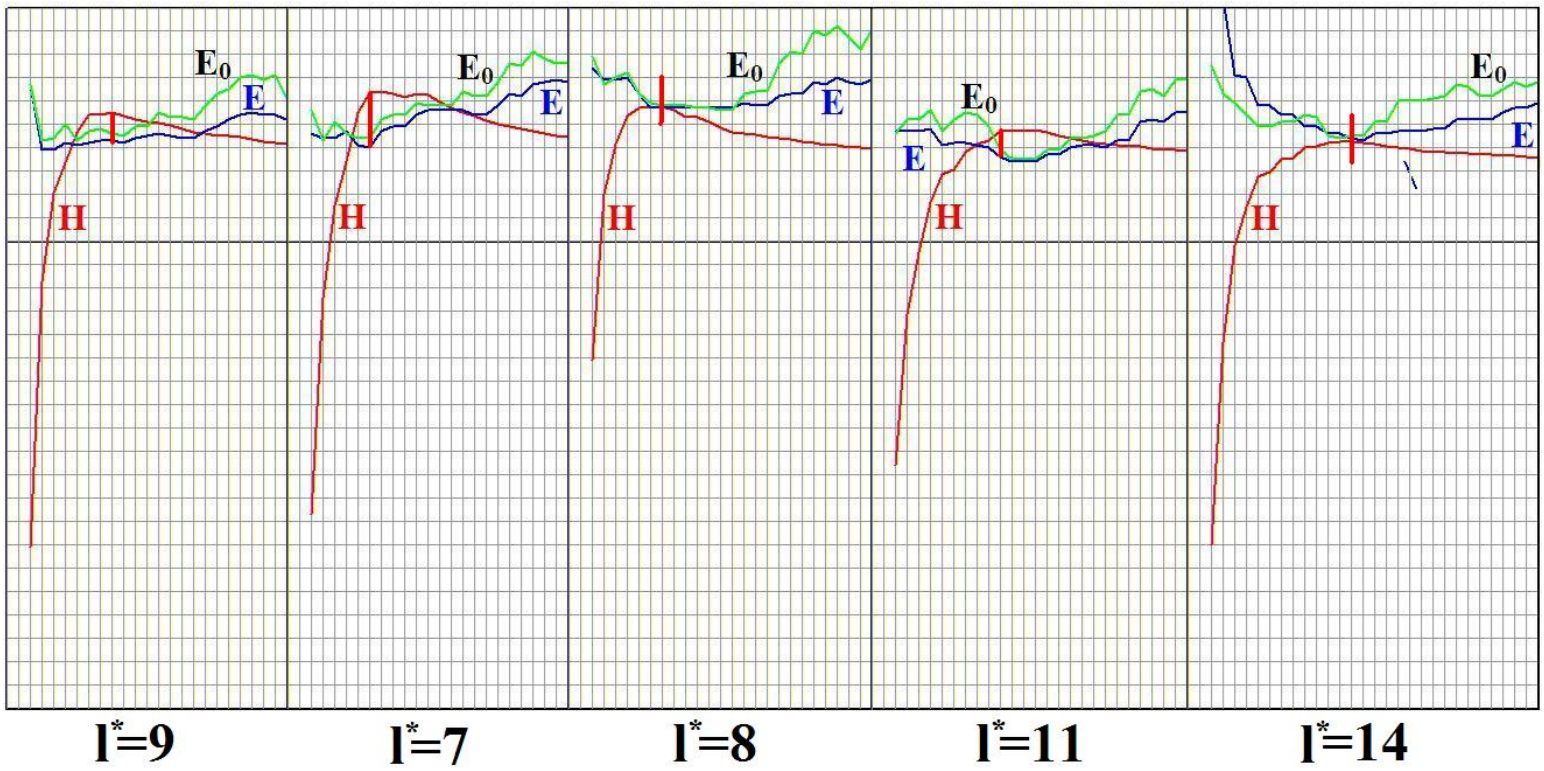
# Обнаружение начала переобучения
(результаты экспериментов)

Рис. 13. Графики качества описания обучающей выборки ($H$) и графики ошибки распознавания ($E$) в зависимости от числа выбранных эталонов.



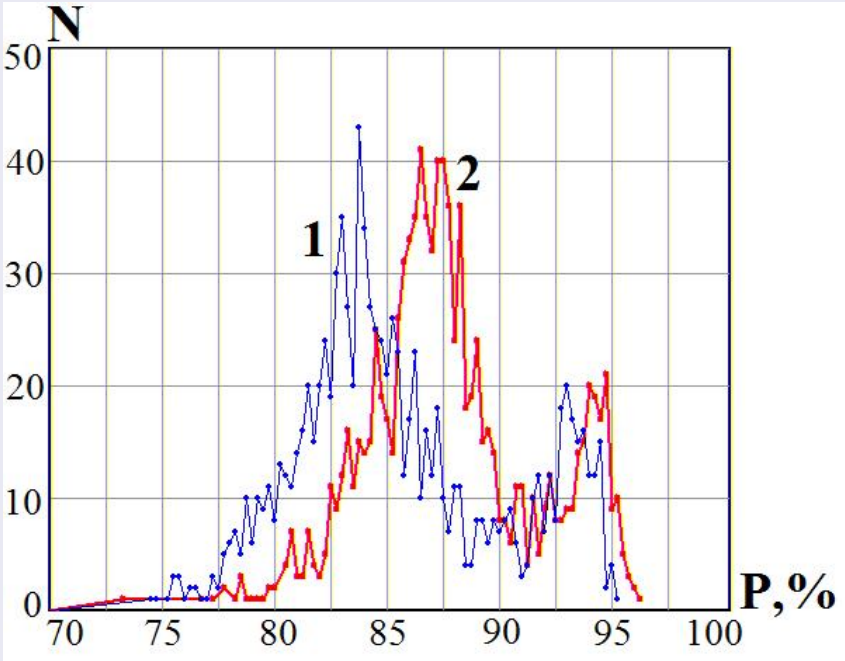
# Обнаружение начала переобучения
(результаты экспериментов)

Рис. 14. Графики качества описания обучающей выборки ($H$) и графики ошибки распознавания ($E$ - с учетом веса кластера, $E_0$ - без учета веса кластера) в зависимости от числа выбранных эталонов.



# Обнаружение начала переобучения
(результаты экспериментов)

Рис. 15. Распределение надежности $P(\%)$ распознавания контрольной выборки без цензурирования (1) и с цензурированием (2).




$\overline{P(l)} = 85.58\%$, $\overline{P(2)} = 87.86\%$; $\overline{l^*} = 10.81$, $\overline{L} = 28.91$.

Сформулирована и подтверждена гипотеза о том, что точка перегиба кривой, описывающей разделимость классов, может служить сигналом о начале переобучения.

# Сокращение пространства признаков
## (алгоритм FRiS-GRAD)

**Загоруйко Н. Г., Кутненко О. А.** Алгоритм GRAD для выбора признаков // Труды VIII Межд. конференции «Применение многомерного статистического анализа в экономике и оценке качества», Москва: МЭСИ, 2006. С. 81–89.
**Загоруйко Н. Г.** Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 268 с.
**Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Attribute selection through decision rule construction (algorithm FRiS-GRAD) //Proc. of 9th Intern. Conf. Pattern Recognition and Image Analysis: New Information Technologies. Nizhniy Novgorod, 2008. V. 2. P. 335–338.

В алгоритме FRiS-GRAD методом полного перебора формируются информативные системы признаков (гранулы) малой размерности, а затем эти гранулы используются в качестве входных элементов для алгоритма AdDel, который представляет собой комбинацию двух известных жадных алгоритмов Addition и Deletion. Как показали эксперименты, перегиб кривой качества распознавания позволяет автоматически определить количество признаков в системе.

В алгоритме выбора признаков FRiS-C-GRAD информативность каждой гранулы и системы признаков проверяется по критерию качества $H$ описания выборки при разных количествах столпов. В итоге автоматически формируется нуклеус обучающей выборки, обеспечивающий построение решающих правил, избегающих переобучения.

Эффективность процедуры сокращения пространства признаков и объектов иллюстрируется на примере решения задачи «Colon» распознавания двух классов объектов (пациентов) по генетическим признакам.
$N = 2000,\ M_1 = 40,\ M_2 = 22.$

Таблица: Результаты эксперимента

| | Без цензурирования объектов | С цензурированием объектов |
|---|---|---|
| $\overline{l}$ | 10 | 3 |
| $\overline{Err}(\%)$ | 28.3 | 23.3 |
| $\overline{N}^*$ | 30 | 43 |
| $\overline{L}(N^*)$ | 6.8 | 4 |

# Заключение

Использование FRiS-функции было полезным при построении решающих правил, автоматической классификации (таксономии) и выборе информативных признаков, при получении количественной оценки компактности. В данной работе показана полезность применения этой меры сходства и для решения задачи защиты от переобучения. Описан алгоритм, который выбирает подмножество наиболее информативных объектов и признаков, и останавливает процесс обучения в точке, в которой начинается переобучение.

**Борисова И. А.** Алгоритм таксономии FRiS-Tax // Научный вестник НГТУ. 2007. № 3. С. 3–12.

**Borisova I. A., Dyubanov V. V., Kutnenko O. A., Zagoruiko N. G.** Use FRiS-Function for Taxonomy, Attribute Selection and Decision Rule Construction // Knowledge Processing and Data Analysis. Springer-Verlag Berlin Heidelberg. 2011. P. 256–270.

**Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А.** Количественная мера компактности и сходства в конкурентном пространстве // Сибирский журнал индустриальной математики. 2010. Т. XIII. № 1(41). С. 59–71.

# Спасибо за внимание!