Цензурирование ошибочно классифицированных объектов выборки

Борисова И. А. [1,2], Кутненко О. А. [1,2]

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

17-я Всероссийская конференция
«Математические методы распознавания образов – 2015»

19–25 сентября, 2015, Россия, г. Светлогорск

Введение

Предмет исследования —

цензурирование выборок, содержащих значительное число неверно классифицированных объектов.

Цель исследования —

построение алгоритма, последовательно удаляющего объекты, максимально ухудшающих качество описания выборки, и автоматически обнаруживающего момент окончания цензурирования данных.

Мотивация исследования —

анализ возможности использования FRiS-функции для решения задачи восстановления структуры данных, изначально содержащих значительное число неверно классифицированных объектов.

.

Области приложений:

Анализ данных и распознавание образов.

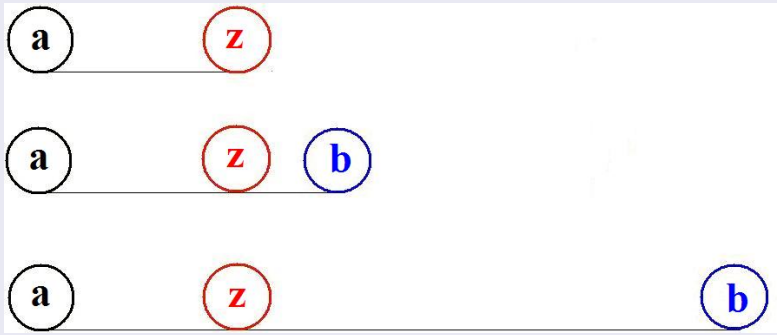
Введение

Различные алгоритмы распознавания при наличии шумовых объектов и выбросов в обучающей выборке обрабатывают их по-разному. В алгоритмах построения решающих деревьев для уменьшения влияния таких объектов предусмотрена процедура редукции (pruning)- удаление поддеревьев, имеющих низкую статистическую надежность из-за того, что для их построения использовались объекты-выбросы. (Teng, 2001; Quinlan, 1986) В других алгоритмах предусмотрена предобработка данных, в процессе которой шумовые объекты с помощью некоторого критерия выявляются и отфильтровываются. (Frinay and Verleysen, 2014; Segata and Blanzieri, 2010; Massie et al.,2007; Son and Kim, 2006; Delany et al., 2012), В некоторых случаях даже предпринимается попытка корректировки отдельных признаков объекта-выброса с целью преобразовать его в типичный объект, хотя это рискованный процесс, который может заменить один шум другим. (Yang et al., 2004; Teng, 2001;Brodley and Friedl, 1999). Большинство процедур выявления шумовых объектов основано на использовании правила ближайшего соседа. (Wilson and Martinez, 2000; Jankowski and Grochowski, 2004; Brighton and Mellish, 2002).

1. Teng, Choh Man. 2001. "A Comparison of Noise Handling Techniques."Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference: 269-273.
2. Quinlan, J R. 1986. "Induction of Decision Trees."Machine Learning: 81-106.
3. Frenay, Benoit, and Michel Verleysen. 2014. "Classification in the Presence of Label Noise: a Survey."IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 25 (5): 845-869.
4. Segata, N, and E Blanzieri. 2010. "Noise Reduction for Instance-Based Learning with a Local Maximal Margin Approach."Journal of Intelligent Information Systems 35 (October).
5. Massie, Stewart, Susan Craw, and Nirmalie Wiratunga. 2007."When Similar Problems Don't Have Similar Solutions."In: Proceedings of the 7th International Conference on Case-Based Reasoning (ICCBR 07), Springer-Verlag, Berlin, Heidelberg: 92-106.
6. Son, Seung-hyun, and Jae-yearn Kim. 2006. "Data Reduction for Instance-Based Learning Using Entropy-Based Partitioning."iN Proceedings of the International Conference on Computational Science and Its Applications: 590-599.
7. Delany, Sarah Jane, Nicola Segata, and Brian Mac Namee. 2012. "Profiling Instances in Noise Reduction."Knowledge-Based Systems 31 (July): 28-40.
8. Yang, Ying, Xindong Wu, and Xingquan Zhu. 2004. "Dealing with Predictive-but-Unpredictable Attributes in Noisy Data Sources."In: Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy.
9. Brodley, Carla E, and Mark A Friedl. 1999. "Identifying Mislabeled Training Data."Journal of Artificial Intelligence Research 11: 131-167.
10. Wilson, D.R., and T.R. Martinez. 2000. "Reduction Techniques for Instance-Based Learning Algorithms."Machine Learning 38 (3): 257-286.
11. Jankowski, Norbert, and Marek Grochowski. 2004. "Comparison of Instances Seletion Algorithms I. Algorithms Survey."Artificial Intelligence and Soft Computing: 1-6.
12. Brighton, Henry, and Chris Mellish. 2002. "Advances in Instance Selection for Instance-Based Learning Algorithms *."Data Mining and Knowledge Discovery 6: 153-172.

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.

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

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

$$F(z, A|B) = \frac{r(z, B) - r(z, A)}{r(z, B) - r(z, A)} \quad (1)$$

FRiS-компактность и качество описания выборки

Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A. A construction of a compressed description of data using a function of rival similarity // Journal of Applied and Industrial Mathematics. 2013. V. 7, № 2, P. 275–286.

Компактность образов — усредненное значение FRiS-функции по всем объектам образов A и B:

$$F_{AB} = \frac{\sum_{a \in A} F(a, A|B) + \sum_{b \in B} F(b, B|A)}{|A| + |B|} \qquad (2)$$

FRiS-компактность и качество описания выборки

Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A. A construction of a compressed description of data using a function of rival similarity // Journal of Applied and Industrial Mathematics. 2013. V. 7, № 2, P. 275–286.

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. 2008. V. 18, № 1. P. 1–6.

FRiS-компактность образа A по множеству столпов S_A и S_B образов A и B:

$$C_{A|B} = \frac{\sum_{a \in A} F(a, S_A|S_B) - |S_A|}{|S_A||A|}, \tag{3}$$

где $S_A \bigcup S_B$ - достаточный для описания выборки набор столпов. Аналогично вычисляется величина $C_{B|A}$ FRiS-компактности образа B в конкуренции с A.
Компактность образов A и B:

$$C_{AB} = \sqrt{C_{A|B} C_{B|A}} \tag{4}$$

FRiS-компактность и качество описания выборки

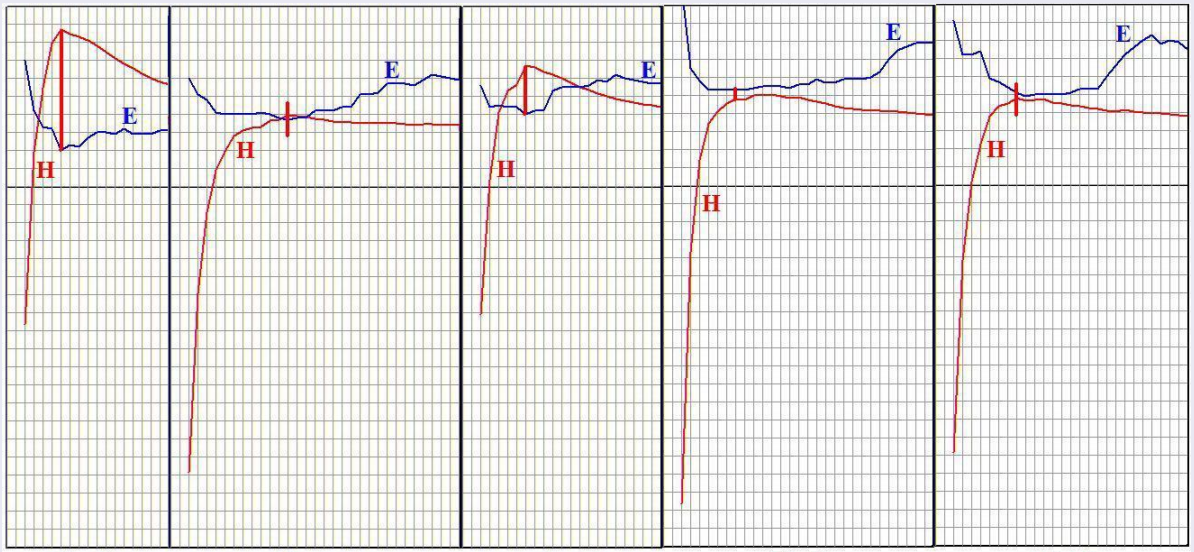
Zagoruiko N. G., Кутненко О. А., Зырянов А. О., Леванов Д. А. Обучение распознаванию без переобучения //Машинное обучение и анализ данных. 2014. Т. 1, № 7. С. 891–901.

Качество описания выборки набором столпов S_A и S_B образов A и B:

$$H(S_A, S_B) = \frac{\sum_{a \in A} F(a, S_A|B) + \sum_{b \in B} F(b, S_B|A)}{|S_A \cup S_B||A \cup B|} \qquad (5)$$

При изменении набора столпов меняется качество описания H обучающей выборки и ошибка распознавания E независимой тестовой выборки.

Рис. 2. Графики качества описания обучающей выборки (H) и графики ошибки распознавания (E) в зависимости от числа выбранных эталонов.



Цензурирование выборки без построения столпов

Борисова И. А., Кутненко О. А. Цензурирование ошибочно классифицированных объектов выборки //Машинное обучение и анализ данных. 2015. Т. 1, № 11. С. 1632-1641.

Дано: выборка содержит неверно классифицированные объекты (или шумовые объекты).

Требуется: исключить шумовые объекты.

Сложность: малый объем выборки; большая доля шумов.

Для цензурирования ошибочно классифицированных объектов разработан алгоритм, который ориентируется только на локальные характеристики объектов выборки.

Цензурирование выборки без построения столпов

Количественные характеристики локальной компактности объектов в той или иной части выборки оцениваются с помощью FRiS-функции.

Локальная компактность выборки:

$$F_{AB} = \frac{\sum_{a \in A} F(a,A|B) + \sum_{b \in B} F(b,B|A)}{|A|+|B|},$$

где в качестве расстояния от объекта до образа используется среднее расстояние до k ближайших объектов этого образа.

Локальная разделимость обучающей выборки:

$$G_{AB} = \frac{M^*}{M} F_{AB}, \tag{6}$$

M^*/M — штраф, регулирующий количество исключенных объектов.

Цензурирование выборки без построения столпов

Алгоритм цензурирования:

1. Вычисляется разделимость для всей выборки.

2. Отыскивается объект, удаление которого из выборки максимально повышает ее разделимость. Этот объект признается выбросом и исключается из выборки.

3. Процедура повторяется до момента, когда исключение любого объекта из обучающей выборки только ухудшает ее разделимость, т.е. процесс цензурирования продолжается до достижения точки перегиба функции разделимости.

Чувствительность (Sensitivi) - способность метода давать правильный результат:

$$Se = \frac{TP}{TP + FN} \times 100\%,$$

TP - количество истино положительных результатов,
FN - количество ложноотрицательных результатов.

Чувствительность алгоритма характеризуется его вероятностью определять шумовые объекты.

Специфичность (Specificity) - способность метода не давать ложноположительных результатов:

$$Sp = \frac{TN}{TN + FP} \times 100\%,$$

TN - количество истино отрицательных результатов,
FP - количество ложноположительных результатов.

Специфичность алгоритма характеризуется отсутствием ошибочных результатов для нешумовых объектов.

Тестирование алгоритма. Результаты экспериментов

Алгоритм тестировался на серии из 10 модельных задач, отличающихся сложностью и структурой, каждая из которых решалась 100 раз на разных обучающих выборках, т. е. общее количество экспериментов при различных численных реализациях данных было равно 1000. Объем обучающих выборок был 100 объектов. Уровень шума (α) составлял 15%. Решающее правило - правило k ближайших соседей, $k = 3$.

Тестирование алгоритма. Результаты экспериментов

	Чувствительность (Se)	Специфичность (Sp)
$G_0 = F_{AB}$	98.28	50.72
$G_1 = \frac{M^*}{M} F_{AB}$	97.48	96.31
$G_2 = (\frac{M^*}{M})^2 F_{AB}$	96.16	98.94

Рис. 3. Связь чувствительности (Se) и специфичности (Sp) алгоритма: 0 - критерий G_0, 1 - критерий G_1, 2 - критерий G_2.

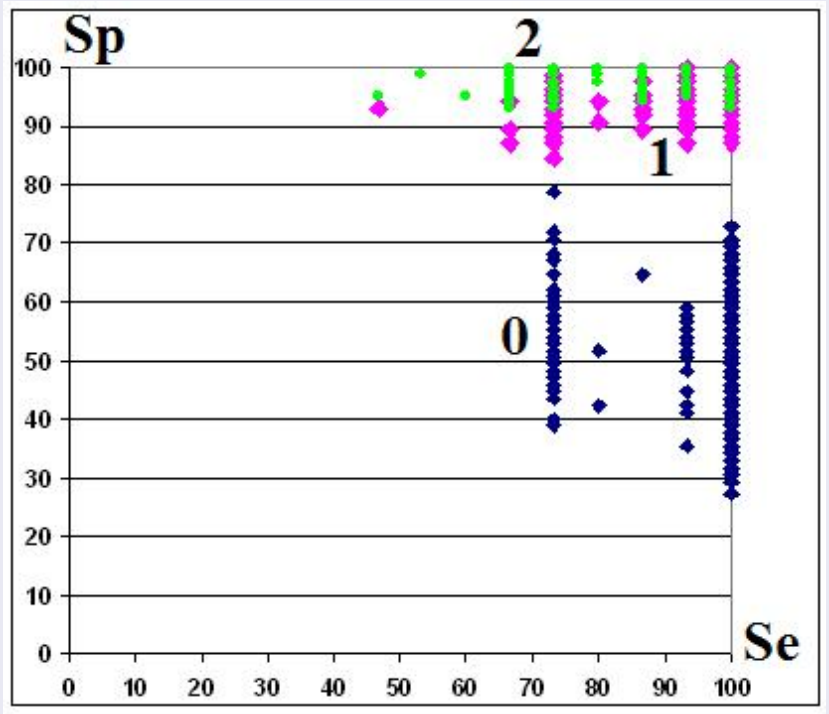
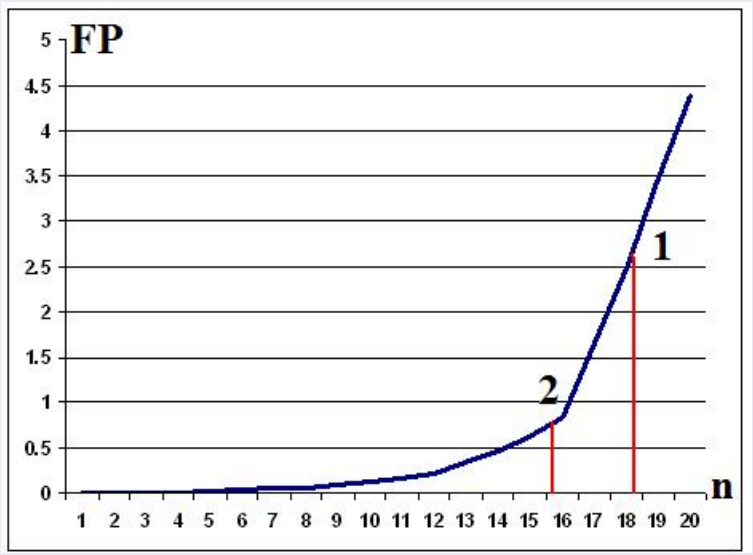


Рис. 4. Зависимость доли ложноположительных результатов (FP) от точки останова алгоритма: 1 - критерий G_1, 2 - критерий G_2.

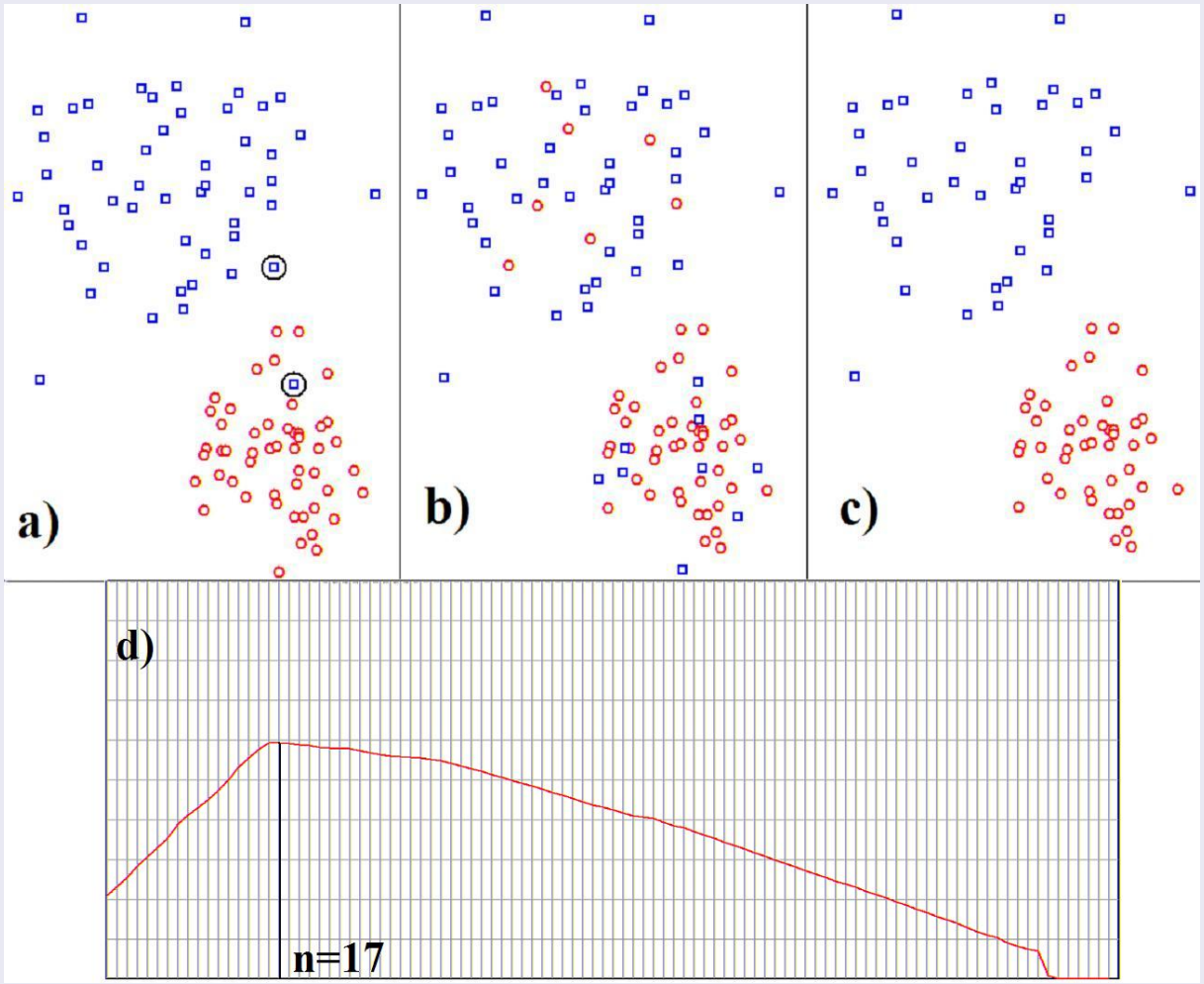


$$G_0 : n = 56.55;$$

$$G_1 : n = 17.69;$$

$$G_2 : n = 15.17.$$

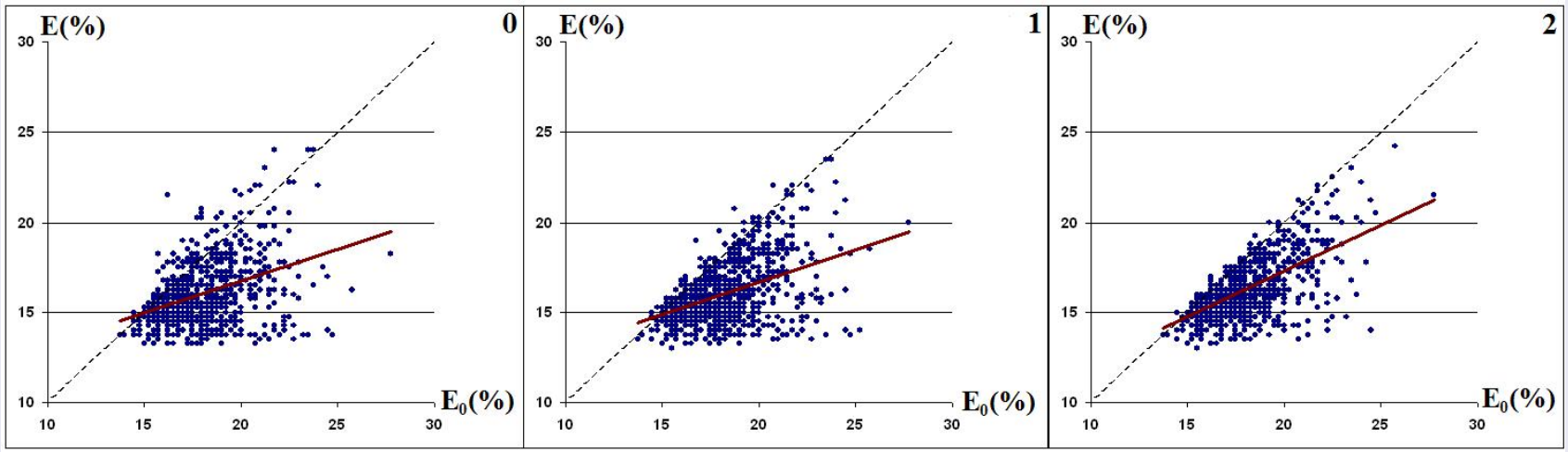
Рис. 5. a) исходная выборка, b) зашумленная выборка ($\alpha = 15\%$), c) очищенная выборка, d) график локальной разделимости описания обучающей выборки (G_{AB}).



Тестирование алгоритма. Результаты экспериментов

Сравнение результатов распознавания тестовой выборки до и после цензурирования. Пунктирная диагональ задает порог, при котором $E_0 = E$.

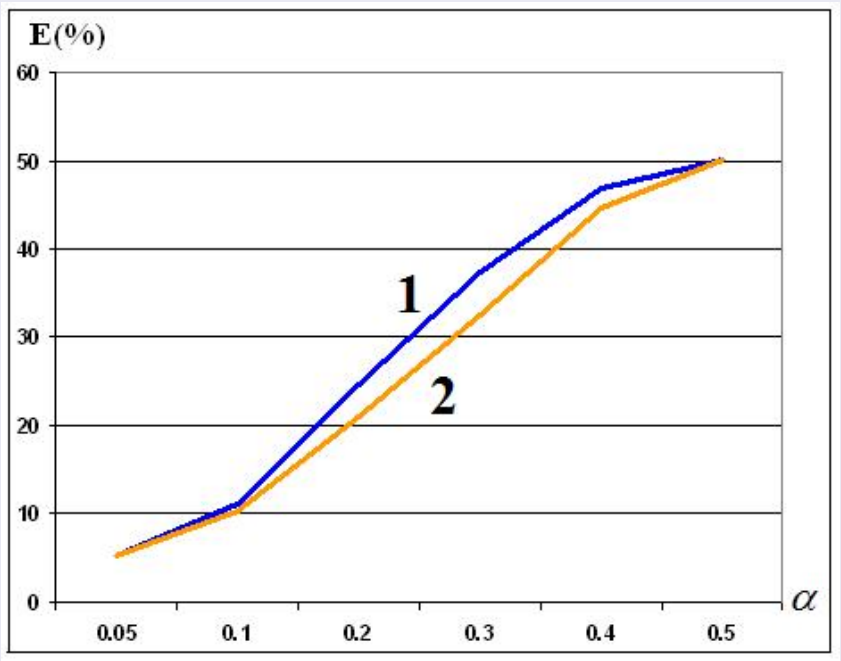
Рис. 6. Связь ошибки распознавания до (E_0) и после цензурирования (E): 0 - критерий G_0, 1 - критерий G_1, 2 - критерий G_2.



Тестирование алгоритма. Результаты экспериментов

В качестве оценки эффективности алгоритма использовалось разница в величинах ожидаемой ошибки распознавания до и после цензурирования обучающей выборки. Ожидаемая ошибка распознавания оценивалась методом Cross Validation.
α — доля объектов, для которых изменялась целевая характеристика.
Серия — 100 выборок. Объем обучающей выборки — 100 объектов.

Рис. 7. Зависимость ошибки распознавания от α — доли неверно классифицированных объектов выборки до (1) и после (2) цензурирования.



Заключение

Исследовалась возможность цензурирования ошибочно классифицированных объектов обучающей выборки для случая, когда доля таких объектов достаточно велика, а объем выборки ограничен. В этом случае цензурирование осуществляется путем снижения сложности выборки. Сложность при этом оценивается величиной локальной разделимости классов, которая вычисляется с помощью функции конкурентного сходства. Проведенные эксперименты на широком спектре модельных задач подтверждают работоспособность предложенного алгоритма цензурирования.

Спасибо за внимание!