# Цензурирование ошибочно классифицированных объектов выборки*

***И. А. Борисова***[1,2,3]***, О. А. Кутненко***[1,2,3]
biamia@mail.ru
[1]Институт математики им. С. Л. Соболева СО РАН, Новосибирск; [2]Новосибирский государственный университет, Новосибирск; [3]Конструкторско-технологический институт вычислительной техники СО РАН, Новосибирск

Рассматривается задача цензурирования выборок, изначально содержащих значительное число неверно классифицированных объектов. Предложен алгоритм цензурирования, ориентированый только на локальные характеристики объектов выборки. Для оценки вероятности принадлежности объекта к одному из двух образов используется тернарная относительная мера — функция конкурентного сходства (function of rival simularity — FRiS-функция). В фиксированном признаковом пространстве цензурирование состоит в последовательном удалении объектов, максимально ухудшающих качество описания выборки (или оценку разделимости классов). Результаты тестирования алгоритма на широком спектре модельных задач позволили сделать вывод, что объекты, удаленные до точки перегиба функции, описывающей разделимость классов, как правило, являются выбросами, искажающими структуру данных.

**Ключевые слова**: *анализ данных; функция конкурентного сходства; компактность образов; разделимость классов; распознавание образов; цензурирование объектов*

# Outliers detection in datasets with misclassified objects*

***I. A. Borisova***[1,2,3]***, and O. A. Kutnenko***[1,2,3]
[1]Sobolev Institute of Mathematics, SB RAS, Novosibirsk; [2]Novosibirsk State University, Novosibirsk; [3]Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk

**Background:** The problem of outliers detection is one of the important problems in Data Mining. Here, outliers are considered as initially misclassified objects of the dataset. Such objects in small datasets can seriously interrupt the process of classification. This paper describes an algorithm of censoring such data, focusing only on the local characteristics of objects in the dataset.

**Methods:** Censoring procedure in a fixed feature space consists of sequential removals of objects, which deteriorate the quality of dataset description (a value of classes' separability) in the strongest way. This value depends on the number of objects in the dataset and similarity of objects with their class in competition with the rival class. To evaluate the similarity of the object $z$ with class $A$ in competition with class $B$, the ternary relative measure called the function of rival similarity (FRiS-function) is used.

**Results:** The proposed algorithm was tested on a wide range of model problems. Accuracy of $k$ nearest neighbors classification before and after outliers elimination from the datasets was in use to estimate efficiency of the censoring algorithm. In the most tasks, it is appeared to be improvement in classification accuracy after censoring. Analysis of objects which were recognized as outliers showed up to 96% sensitivity and 99% specificity.

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00039.

**Concluding Remarks:** According to the obtained results, it is possible to conclude that the objects, which were deleted before the inflection point of the classes separability function, usually distort the structure of the data. Therefore, their exclusion from the analyzed dataset increases the reliability of recognition.
**Keywords**: *Data Mining; function of rival similarity; compactness; class separability; outliers detection; classification*

## 1 Введение

Одним из следствий бурного развития технологий в последние десятилетия явилось лавинообразное накопление информации, получаемой, обрабатываемой и сохраняемой с их помощью. Этот же факт приводит к тому, что в собранных таким образом данных фигурирует большое количество нерелевантной информации, шумов, ошибок, от которых эти данные необходимо очищать, прежде чем приступать к решению тех или иных задач Data Mining.

Задача цензурирования шумовых объектов, равно как и задача фильтрации нерелевантных признаков, давно и довольно успешно решаются в области анализа данных. Хороших результатов при решении задачи фильтрации нерелевантных признаков или ее аналога — задачи выбора информативных признаков (feature selection) — удалось достичь благодаря использованию функции конкурентного сходства (FRiS-функции) [1]. Это объясняется тем, что FRiS-функция оказалась достаточно простым и надежным способом оценивать вероятность ошибки распознавания каждого отдельного объекта. Переход от бинарного индикатора «есть ошибка – нет ошибки» к количественной величине «вероятность ошибки» позволил более точно оценивать компактность выборок в том или ином пространстве признаков и, как следствие, более качественно выбирать наиболее информативное подпространство признаков и отфильтровывать шумы.

Данное свойство FRiS-функции — оценивать вероятность ошибочного распознавания каждого конкретного объекта — также может оказаться полезным для фильтрации шумовых объектов при решении задачи цензурирования выбросов (outliers detection). Существуют различные интерпретации этой задачи. В одних источниках выбросами называют объекты, порожденные механизмами, отличными от механизмов порождения остальной выборки [2], в других — объекты, резко повышающие сложность модели, в третьих — ошибки измерения и ввода данных, которые оказываются далеко от всех типичных объектов выборки [3]. Соответственно, и подходы к решению задачи поиска выбросов различаются. Наиболее распространенным является статистический подход, при котором выбросы отыскиваются с помощью статистических тестов в предположении об известном законе распределения анализируемой выборки [4]. Но, как правило, распределения выборок в реальных прикладных задачах не вписываются в рамки стандартных моделей. В связи с этим все большую популярность набирает непараметрический подход, опирающийся на метрические характеристики выборки, при котором выбросы отыскиваются среди объектов или кластеров, удаленных от основной массы объектов [5]. Все эти подходы объединены установкой, что, во-первых, количество выбросов обычно незначительно в сравнении с объемом выборки, а во-вторых, выбросы приводят к переобучению алгоритмов распознавания и тем самым увеличивают вероятность ошибки.

В [6] была рассмотрена задача цензурирования периферийных объектов выборки, неоправданно повышающих сложность структуры данных в условиях, когда эта сложность оценивается величиной FRiS-компактности. Для этих целей сначала формировался

набор типичных объектов — столпов, отражающих структуру выборки, и затем с его помощью отфильтровывались выбросы.

В данной статье под задачей цензурирования понимается задача исключения из обучающей выборки ошибочно классифицированных объектов, которые оказались в ней на этапе сбора и сохранения данных. Если таких объектов достаточно много, а объем выборки невелик, то восстановление ее структуры оказывается серьезно затруднено. Для решения подобных задач предлагается осуществлять цензурирование на основе анализа локального окружения объектов. Данный подход опирается на гипотезу локальной компактности [7]. Оценивать количественные характеристики локальной компактности объектов в той или иной части выборки также предполагается с помощью функции конкурентного сходства.

## 2 FRiS-компактность и качество описания выборки

Гипотеза компактности является одной из важных концепций в анализе данных, а ее выполнение для объектов обучающей выборки — необходимым требованием для успешного решения задачи распознавания. Для получения количественной оценки компактности образов в фиксированном признаковом пространстве предлагается использовать FRiS-функцию, с помощью которой формализуется представление о компактности как о «высоком» сходстве объектов одного образа друг с другом и «низком» сходстве с объектами других образов.

Идея конкурентного сходства возникла в связи с желанием учитывать конкурентную ситуацию — контекст при оценке похожести объекта на другой объект или принадлежности объекта к образу. Данная концепция хорошо согласуется с человеческими особенностями оценки похожести объектов. Два объекта с несовпадающими свойствами могут считаться «сходными» или «не сходными», «близкими» или «далекими» в зависимости от свойств других объектов. Хорошо известная бытовая фраза «все познается в сравнении» на самом деле отражает фундаментальный закон познания. Адекватная мера сходства должна определять величину сходства, зависящую от особенностей конкурентного окружения объекта $z$. В распознавании образов сходство также является категорией не абсолютной, а относительной. При распознавании принадлежности объекта $z$ к одному из двух образов $A$ или $B$ важно знать не только расстояние $r(z, A)$ до образа $A$, но и расстояние $r(z, B)$ до конкурирующего образа $B$.

Для вычисления конкурентного сходства объекта $z$ с объектом $a$ в конкуренции с объектом $b$ предлагается использовать следующую величину:

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

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, как с объектом $a$, так и с $b$. 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 сходству $z$ с $a$, равному $-1$.

Сходство $F$ между объектами не зависит от положения начала координат, поворота координатных осей и одновременного умножения их значений на одну и ту же величину. Но независимые изменения масштабов разных координат меняют вклад, вносимый отдельными характеристиками в оценку и расстояния, и сходства.

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

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

при этом расстояние от объекта $z$ до образов $A$ и $B$ может вычисляться по-разному. В качестве него может использоваться и расстояние $r(z, a)$ до ближайшего объекта $a$ образа $A$, и среднее расстояние до всех объектов образа, и среднее расстояние до $k$ ближайших объектов образа, и расстояние до центра тяжести образа, и т. д. В дальнейшем в качестве расстояния от объекта до образа по умолчанию будет использоваться расстояние до ближайшего объекта этого образа.

Для произвольного объекта $z \in A$ мера конкурентного сходства этого объекта со своим образом в конкуренции с образом $B$ показывает, насколько этот объект похож на представителей своего образа и не похож на представителей образа $B$, поэтому при решении задачи распознавания FRiS-функция может интерпретироваться как оценка вероятности принадлежности объекта $z$ к образу $A$. Усредняя значения FRiS-функции из (1) по всем объектам образов $A$ и $B$, можно оценить важную характеристику решаемой задачи распознавания — компактность образов, аналогами которой у других авторов [8] выступают такие понятия, как отделимость классов, сложность выборки, качество выборки и т. д.:

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

Варьируя способ вычисления расстояния от объекта до образа, с помощью (2) можно моделировать различные варианты компактности.

В [9] был описан один из таких вариантов получения количественной оценки компактности, который затем с успехом использовался при решении задачи выбора информативного набора признаков. Его особенность заключается в том, что вместо всех объектов выборки для вычисления компактности по формуле (2) используются только типичные представители образов — столпы.

Построение множества столпов, сохраняющего основные закономерности задачи, необходимые для хорошего распознавания как объектов исходной выборки, так и новых объектов, является одним из способов проявить особенности данных, перейти к их сжатому описанию, оценить сложность выборки. Чем сложнее структура образов, чем сильнее они пересекаются, тем больше столпов потребуется для описания таких данных. Для построения сжатого описания данных в виде системы столпов используется алгоритм FRiS-Stolp [10], который работает при любом соотношении количества объектов к количеству признаков и при произвольном виде распределения образов. Набор столпов считается достаточным для описания выборки, если сходство $F$ всех объектов обучающей выборки с ближайшими своими столпами в конкуренции с ближайшими объектами других образов превышает пороговое значение $F^*$, например $F^* = 0$.

Чтобы вычислить величину FRiS-компактности образа $A$ по множеству столпов $S_A$ и $S_B$ образов $A$ и $B$, соответственно, используется следующая формула:

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

Здесь $S_A$ и $S_B$ — достаточный для описания выборки набор столпов.

Аналогично вычисляется величина $C_{B|A}$ FRiS-компактности образа $B$ в конкуренции с $A$. Итоговая величина компактности образов $A$ и $B$ вычисляется как геометрическое усреднение величин $C_{A|B}$ и $C_{B|A}$:

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

Отметим, что количество столпов образа зависит от структуры распределения объектов и величины порога $F^*$: с ростом $F^*$ увеличивается как число столпов, так и точность описания распределения, но растет и сложность его описания, т. е. множитель $1/S_A$ в (3) является штрафом за структурную сложность образа.

Безошибочное распознавание всех объектов обучающей выборки является неким гарантом того, что построенная система столпов сохраняет основные свойства выборки. Однако, как правило, требование безошибочного распознавания при решении задачи классификации приводит к чрезмерному усложнению решающих правил и, как следствие, к переобучению.

Для решения этой проблемы вводится понятие качества описания выборки набором столпов. Эта величина, с одной стороны, показывает, насколько рассматриваемый набор столпов отражает основные закономерные связи между описывающими характеристиками и целевым признаком, которые можно наблюдать на всей выборке, а с другой — регулирует количество столпов.

Чтобы оценить качество описания выборки набором столпов $S_A$ и $S_B$ образов $A$ и $B$, используется следующая формула:

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

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

В [11] было экспериментально показано, что если наращивать число столпов, выбирая на роль каждого следующего столпа объект, обеспечивающий максимальный рост величины $H$ в формуле (5), то между $H$ и $E$ имеется закономерная связь, используя которую можно найти количество столпов, не приводящее к переобучению. Примеры кривых $H$ и $E$ для различных модельных задач приводятся на рис. 1.

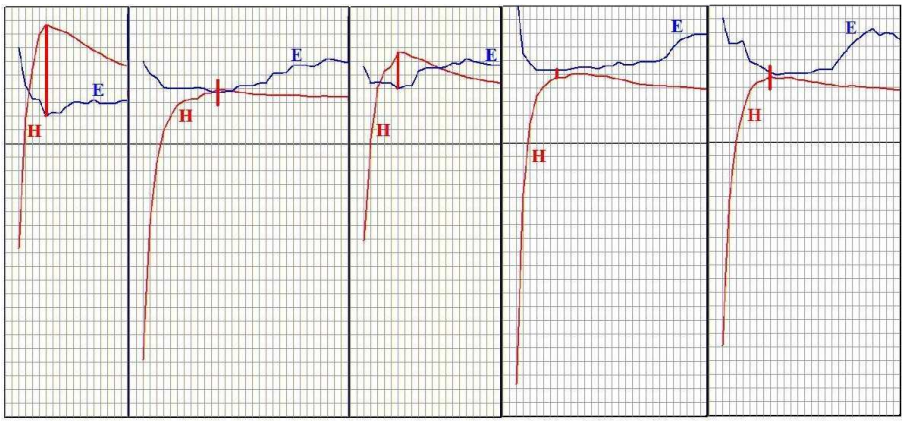


**Рис. 1** Графики качества описания обучающей выборки ($H$) и графики ошибки распознавания ($E$) в зависимости от числа выбранных эталонов

## 3 Цензурирование выборки с опорой на столпы

При построении столпов наряду с объектами, хорошо отражающими структуру образов, принимают участие и шумящие объекты, и даже мелкие кластеры таких объектов, влияние которых было бы целесообразно исключить. Для их цензурирования можно применять описанный в [6] алгоритм, использующий в качестве критерия, управляющего процессом повышения компактности обучающей выборки, меру FRiS-компактности образов, вычисляемую по формуле (4), и включающий как составную часть алгоритм FRiS-Stolp.

Для регулирования доли цензурируемых объектов в рассмотрение вводится величина штрафа за исключение объектов из обучающей выборки $(M^*/M)^\gamma$, где $M$ — размерность выборки; $M^*$ — число объектов обучающей выборки, оставшихся после очередного этапа сокращения выборки; $\gamma \geqslant 0$ — параметр, регулирующий вклада штрафа в общую оценку разделимости. С учетом этого $Q_{AB}$ — качество описания выборки после цензурирования образов на каждом шаге сокращения выборки — оценивается следующим образом:

$$Q_{AB} = \left(\frac{M^*}{M}\right)^\gamma C_{AB}\,. \tag{6}$$

Сначала алгоритмом FRiS-Stolp строится достаточный набор столпов, стоящих в центрах своих кластеров, и по формулам (3), (4), (6) вычисляется качество описания выборки $Q_{AB}$. Затем выбирается кластер, исключение которого обеспечивает максимальное значение $Q_{AB}$ на оставшихся объектах. При этом могут исключаться только кластеры, содержащие не больше $m^*$ объектов. Процесс построения столпов и исключения кластеров повторяется, пока либо доля удаленных объектов не превысит заданный порог, либо не будет кластеров, содержащих не больше $m^*$ объектов.

По списку найденных оценок качества выборки выбирается вариант, соответствующий максимуму величины $Q_{AB}$. Набор столпов, который был зафиксирован при этом, служит основой решающего правила, используемого для распознавания контрольной выборки.

Алгоритм тестировался на модельных задачах распознавания двух образов. Эксперименты показали, что повышение компактности обучающей выборки более чем в 99% случаев приводит к повышению качества распознавания. Очищенная выборка описывается более простым решающим правилом, что повышает надежность распознавания контрольных объектов. Трудоемкость алгоритма зависит от исходной компактности образов — чем она выше, тем меньше времени требуется для выбора наилучшего варианта цензурирования.

## 4 Цензурирование выборки без построения столпов

В описанной выше задаче процедура цензурирования использовалась для упрощения структуры данных путем исключения из выборки непредставительных, периферийных объектов либо выбросов, попадающих в выборку с очень малой вероятностью. В этих условиях процедура построения столпов для заданной выборки оказывается устойчивой, а полученный набор столпов корректно описывает структурные особенности выборки.

Однако если рассматривать специфический случай задачи цензурирования — задачу классификации в условиях, когда обучающая выборка имеет малый объем и изначально содержит большое количество неверно классифицированных объектов, ситуация меняется. Алгоритм FRiS-Stolp, как и большинство алгоритмов классификации, строящих сложные решающие правила, не сможет корректно работать. Для цензурирования ошибочно классифицированных объектов в этом случае будет рассматриваться не требующий предварительного построения столпов алгоритм, который ориентируется только на локальные

характеристики объектов выборки. Отметим, что под цензурированием в данной задаче понимается исключение из обучающей выборки неверно классифицированных объектов, снижающих ожидаемую надежность распознавания новых объектов.

Для оценки меры разделимости (простоты) выборки $G_{AB}$ в данной задаче будем использовать локальную компактность выборки, вычисленную по формуле (2) при условии, что в качестве расстояния от объекта до образа используется среднее расстояние до $k$ ближайших объектов этого образа. Кроме того, так как с увеличением количества исключенных объектов повышается вероятность переобучения алгоритма, введем штраф $M^*/M$, регулирующий количество исключенных объектов. В результате локальная разделимость обучающей выборки будет вычисляться по следующей формуле:

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

Сам алгоритм цензурирования при этом будет следующим:

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. Другими словами, процесс цензурирования продолжается до достижения точки перегиба функции разделимости.

## 5 Тестирование алгоритма. Результаты экспериментов

Для изучения возможности цензурирования ошибочно классифицированных объектов на основе анализа изменения локальной разделимости выборки был сгенерирован ряд модельных задач распознавания образов с распределениями разной степени сложности. Целевая характеристика для заданной доли объектов, обозначаемой в дальнейшем $\alpha$, изменялась, тем самым в анализируемые выборки вводилась шумовая компонента, состоящая из неверно классифицированных объектов.

В качестве оценки эффективности предложенного алгоритма использовалась разница в величинах ожидаемой ошибки распознавания до и после цензурирования обучающей выборки. В качестве решающего правила во всех экспериментах использовалось правило $k$ ближайших соседей, $k = 3$.

Для оценки того, как изменяется ошибка распознавания без цензурирования и с цензурированием в зависимости от доли ошибочно классифицированных объектов в анализируемой выборке, был проведен следующий эксперимент. Генерировалась серия из 100 выборок, с одними и теми же распределениями и задаваемой параметром $\alpha$ долей шумов. По каждой выборке методом Cross Validation оценивалась ожидаемая ошибка распознавания. Объем обучающей выборки составлял 100 объектов.

Результаты эксперимента приводятся на рис. 2. Здесь штриховой линией изображена зависимость ошибки распознавания от доли шумов в выборке без цензурирования, а сплошной линией — та же зависимость, но для отцензурированных выборок.

Для более глубокого изучения свойств предложенного алгоритма проводилось сравнение результатов распознавания тестовой выборки до и после цензурирования на серии из 10 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. Уровень шума при этом

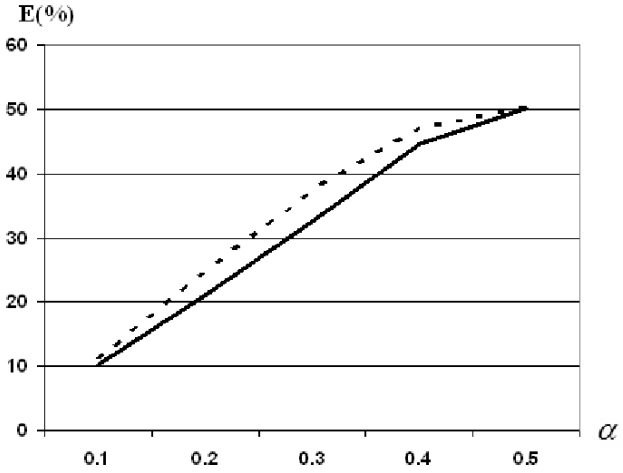


**Рис. 2** Зависимость ошибки распознавания от $\alpha$ — доли неверно классифицированных объектов выборки до и после цензурирования

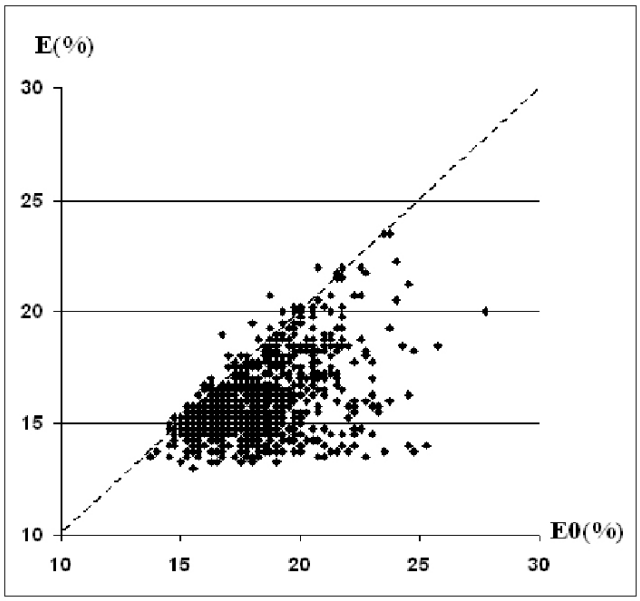


**Рис. 3** Связь ошибки распознавания до ($E0$) и после цензурирования ($E$)

составлял 15%, объем обучающих выборок был 100 объектов. Результаты этого эксперимента приведены на рис. 3. Каждой точке соответствуют результаты распознавания по одной выборке, координата точки по оси абсцисс — это ошибка без цензурирования, по оси ординат — ошибка с цензурированием. Штриховая диагональ задает порог, при котором ошибка без цензурирования равна ошибке с цензурированием. Для экспериментов, результаты которых оказались ниже этого порога, надежность распознавания после цензурирования улучшилась.

Параллельно в этой серии задач отслеживалось количество шумовых объектов, которые реально удается отфильтровать в процессе цензуририрования. Оказалось, что в среднем чувствительность по отношению к шумам составила 96%, специфичность — 99%. Эти результаты позволяют сделать вывод о применимости предложенного подхода к решению задачи фильтрации неверно классифицированных объектов.

## 6 Заключение

В данной работе 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.

## Литература

[1] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* A quantitative measure of compactness and similarity in a competitive space // J. Appl. Ind. Math., 2011. Vol. 5. No. 1. P. 144–154.

[2] *Hawkins D.* Identification of outliers. — London, U.K.: Chapman and Hall, 1980.

[3] *Aggarwal C. C.* Outlier analysis. — Springer, 2013.

[4] *Barnett V., Lewis T.* Outliers in statistical data. — New York, NY, USA: John Wiley, 1994.

[5] *Knorr E., Ng R.* Algorithms for mining distance-based outliers in large datasets // 24th Conference (International) on Very Large Data Bases (VLDB) Proceedings, 1998. P. 392—403.

[6] *Загоруйко Н. Г., Кутненко О. А.* Цензурирование обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, 2013. № 1(22). С. 66–73.

[7] *Аркадьев А. Г., Браверман Э. М.* Обучение машины распознаванию образов. — М.: Наука, 1964.

[8] *Субботин С. А.* Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов // Математичні машини і системи, 2010. № 1. С. 25–39.

[9] *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 // J. Appl. Ind. Math., 2013. Vol. 7. No. 2. P. 275–286.

[10] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of recognition based on the function of rival similarity // Pattern Recognition Image Anal., 2008. Vol. 18. No. 1. P. 1–6.

[11] *Загоруйко Н. Г., 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.

## References

[1] Zagoruiko, N. G., I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko. 2011. A quantitative measure of compactness and similarity in a competitive space. *J. Appl. Ind. Math.* 5(1):144–154.

[2] Hawkins, D. 1980. *Identification of outliers*. London, U.K.: Chapman and Hall.

[3] Aggarwal, C. C. 2013. *Outlier analysis*. Springer.

[4] Barnett, V., and T. Lewis. 1994. *Outliers in statistical data*. New York, NY: John Wiley.

[5] Knorr, E., and R. Ng. 1998. Algorithms for mining distance-based outliers in large datasets. *24th Conference (International) on Very Large Data Bases (VLDB) Proceedings*. 392—403.

[6] Zagoruiko, N. G., and O. A. Kutnenko. 2013. Censoring of a train dataset. *Vectnik Tomskogo gosydarstvennogo yniversiteta. Upravlenie, vychislitel'naya tekhnika i informatika* 1(22):66–73.

[7] Arkad'ev, A. G., and E. M. Braverman. 1964. *Machine learning to pattern recognition*. Moscow: Nauka.

[8] Subbotin, S. A. 2010. Complex characterization and comparison criteria of training samples for diagnostics and pattern recognition. *Matematichni mashini i sistemi* 1:25–39.

[9] Zagoruiko, N. G., I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko. 2013. A construction of a compressed description of data using a function of rival similarity. *J. Appl. Ind. Math.* 7(2):275–286.

[10] Zagoruiko, N. G., I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko. 2008. Methods of recognition based on the function of rival similarity. *Pattern Recognition Image Anal.* 18(1):1–6.

[11] Zagoruiko, N. G., O. A. Kutnenko, A. O. Ziranov, and D. A. Levanov. 2014. Learning to recognition without overfittinig. *Mashinnoe obychenie i analiz dannykh* 1(7):891–901. (In Russian.)