УДК 519.95

Н.Г. Загоруйко, О.А. Кутненко

ЦЕНЗУРИРОВАНИЕ ОБУЧАЮЩЕЙ ВЫБОРКИ[1]

Предлагается количественная мера компактности образов, основанная на использовании функции конкурентного сходства (FRiS-функции). Рассматривается метод цензурирования обучающей выборки путем исключения «шумящих» объектов, что повышает компактность образов и приводит к улучшению качества распознавания контрольной выборки. Состав исключаемых объектов определяется автоматически. Эффективность алгоритма цензурирования иллюстрируется решением модельной задачи распознавания двух образов.

Ключевые слова: *функция конкурентного сходства*, *компактность*, *цензурирование.*

Цензурирование обучающей выборки состоит в исключении из нее объектов, которые понижают компактность образов. Это могут быть как «случайные» объекты, свойства которых сильно отличаются от свойств остальных объектов своих образов, так и объекты, находящие в зоне пересечения с объектами других образов. Такие данные разрушают компактность образов и усложняют решающие правила, что ведет к увеличению числа ошибок при распознавании контрольной выборки. Назовем подобные объекты «шумящими» и будем исключать их из обучающей выборки.

Предлагаемый метод повышения компактности основывается на использовании новой меры для оценки сходства между объектами – функции конкурентного сходства, с помощью которой можно описывать любые распределения образов набором эталонных объектов («столпов»). Использование столпов позволяет оценить вклад в компактность образов каждого объекта выборки и получить количественную меру компактности всей совокупности образов или любого отдельного образа, а также выбрать объекты, вносящие отрицательный вклад в компактность образов. В итоге решающее правило, построенное по очищенной обучающей выборке, обеспечивает повышение качества распознавания контрольных объектов.

1. Функция конкурентного сходства

Сформулируем следующие требования, которым должна удовлетворять мера $F(z,a|b)$ сходства объекта z с объектом a в конкуренции с объектом b.

1. Свойство *нормированности*. Если оценивается мера сходства объекта z с объектом a в конкуренции с объектом b, то при совпадении объектов z и a мера $F(z,a|b)$ должна иметь максимальное значение равное 1, а при совпадении z с b – минимальное значение равное –1. Во всех остальных случаях мера конкурентного сходства принимает значения от –1 до 1.

2. Свойство *антисимметричности*. Значения сходства z с a в конкуренции с b и сходства z с b в конкуренции с a связаны соотношением $F(z,a|b) = -F(z,b|a)$. При

[1] Работа выполнена при финансовой поддержке РФФИ, проект № 11-01-00156.

одинаковых расстояниях $r(z,a)$ и $r(z,b)$ объект z в равной степени будет похожим на объекты a и b и $F(z,a|b) = F(z,b|a) = 0$.

3 Свойство *инвариантности*. Значение $F(z,a|b)$ должно сохраняться при аффинных преобразованиях пространства признаков: при сдвиге начала координат, повороте координатных осей, а также при умножении всех координат на одно и то же число.

Предлагаемая нами функция конкурентного сходства **FRiS** (**F**unction of **Ri**val **S**imilarity) [1]

$$F(z,a \mid b) = \frac{r(z,b) - r(z,a)}{r(z,b) + r(z,a)} \tag{1}$$

удовлетворяет всем этим требованиям.

Как расстояния r между объектами, так и сходство F между ними не зависит от аффинных преобразований пространства признаков. Но независимые изменения масштабов разных координат меняют вклад, вносимый отдельными характеристиками в оценку и расстояний, и сходства. Меняя веса характеристик, можно подчеркнуть сходство или различие между заданными объектами, что обычно и делается при выборе информативных признаков и построении решающих правил в распознавании образов.

Сходство в шкале порядка, используемое в методе k ближайших соседей, отвечает на вопрос: «На объект какого образа z похож больше всего?». Конкурентное сходство, измеряемое с помощью FRiS-функции, отвечает на этот вопрос и, кроме того, на такой вопрос: «Какова абсолютная величина сходства z с $a \in A$ в конкуренции с $b \in B$?» Оказалось, что дополнительная информация, которую дает абсолютная шкала по сравнению со шкалой порядка, позволяет существенно улучшить методы анализа данных (АД). Функция конкурентного сходства используется нами в алгоритмах решения широкого круга как известных, так и новых задач АД [2].

Определим меру сходства $F(z,A|B)$ объекта z с образом A в конкуренции с образом B как $F(z,a|b)$, где a (b) – ближайший к z объект образа $A(B)$, т.е. помимо указанных выше свойств мера сходства $F(z,A|B)$ удовлетворяет свойству *локальности*: $F(z,A|B)$ зависит не от характера распределения всего множества объектов образов A и B, а от особенностей распределения объектов в окрестности z. Окрестностью объекта будем называть сферу минимального радиуса, содержащую объекты анализируемых образов. Отметим, что в зависимости от рассматриваемой задачи образы могут быть представлены как непосредственно своими объектами, так и своими эталонами (столпами).

2. Выбор эталонных объектов

Для распознавания образов необходимо выбрать объекты-эталоны (столпы), с которыми будут сравниваться контрольные объекты. 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 $F^* = 0$. Здесь описано решающее правило, которое основано на использовании FRiS-функции и строится с помощью алгоритма FRiS-Stolp. Этот алгоритм работает при любом соотношении количества объектов к количеству признаков и при произвольном виде распределения образов.

В качестве столпов выбираются объекты, которые обладают высокими значениями двух свойств: обороноспособности по отношению к объектам своего образа и толерантности по отношению к объектам других образов. Чем выше обороноспособность эталона, тем меньше будет ошибок типа «пропуск цели». Чем выше толерантность эталона, тем меньше будет ошибок типа «ложная тревога». В результате для каждого образа выбираются такие столпы, на которые свои объекты похожи больше, чем на объекты конкурирующих образов.

Алгоритм выбирает эталоны для произвольного количества образов, но объяснять его работу будем на примере распознавания двух образов – $A = \{a_1, \ldots, a_{M_A}\}$ и $B = \{b_1, \ldots, b_{M_B}\}$, представленных наборами из M_A и M_B объектов обучающей выборки соответственно. Поясним алгоритм FRiS-Stolp с помощью рис. 1.

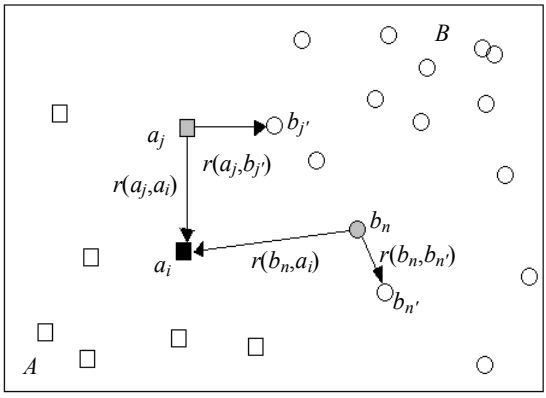

Рис. 1. Оценка обороноспособности и толерантности объекта $a_i \in A$

Начнем с выбора первого столпа для образа A.

1. Оценим качество исполнения роли столпа всеми объектами a_i, $i = 1, \ldots, M_A$, по очереди. Вначале проверим, хорошо ли объект a_i защищает объекты a_j, $j = 1, \ldots, M_A$, образа A. Для каждого объекта a_j, $j = 1, \ldots, M_A$, определим расстояния $r(a_j, a_i)$ и $r(a_j, b_{j'})$, где $b_{j'} \in B$ является ближайшим соседом объекта a_j, т. е. $j' = \arg \min\limits_{m=1,\ldots,M_B} r(a_j, b_m)$, и по формуле (1) получим значение $F(a_j, a_i \mid b_{j'})$ функции сходства объекта a_j с $a_i \in A$ в конкуренции с $b_{j'} \in B$ (см. рис. 1).

2. Выделим $m(a_i)$ объектов $a_j \in A$, сходство которых с a_i не меньше заданного порога F^*: $F_j^+ = F(a_j, a_i \mid b_{j'}) - F^* \geq 0$, $j \in \{1, \ldots, M_A\}$. Эти объекты надежно защищены a_i. Получим оценку $D(a_i)$ обороноспособности объекта a_i:

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

3. Теперь оценим толерантность a_i, т. е. меру несходства с a_i объектов образа B. Для каждого $b_n \in B$, $n = 1, \ldots, M_B$, 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 $F(b_n, b_{n'} \mid a_i)$ объекта b_n с $b_{n'}$ в конкуренции с a_i (см. рис. 1).

4. Выделим те объекты образа B, у которых $F_n^- = F(b_n, b_{n'} \mid a_i) - F^* < 0$, $n \in \{1,\ldots,M_B\}$. Эти объекты больше похожи на a_i, чем на ближайшие объекты своего образа, что отрицательно влияет на оценку a_i. Получим оценку $T(a_i)$ «нетолерантности» объекта a_i:

$$T(a_i) = \sum_{n=1}^{M_B} F_n^- \mid_{F_n^- < 0}.$$

5. Качество выполнения объектом a_i роли столпа образа A оценивается величиной

$$S(a_i) = D(a_i) + T(a_i). \qquad (2)$$

6. Первым столпом образа A становится объект a_i, набравший наибольшее значение величины $S(a_i)$, $i = 1,\ldots,M_A$. Данный столп защищает $m(a_i)$ объектов своего образа.

7. Если в образе A не все объекты надежно защищены выбранным столпом a_i, т. е. $m(a_i) < M_A$, то для оставшихся незащищенными объектов повторяем пункты 1–6, предварительно заменив исходное количество объектов величиной $M_A - m(a_i)$. В результате будет выбран следующий столп. Процесс выбора столпов повторяется до момента, когда сходство всех M_A объектов образа A со своими столпами будет не меньше порога F^*.

8. Тем же способом выбираются столпы и для образа B.

9. Кластеры возникали поочередно, и формирование состава каждого следующего кластера осуществлялось в условиях «отсутствия» многих исходных объектов. По этой причине делается уточнение состава кластеров: объект включается в кластер, образованный ближайшим к нему столпом своего образа. Теперь каждый из столпов стоит в центре своего кластера, т. е. подмножества объектов, которые на него похожи больше, чем на любой другой столп.

Если количество образов K больше двух, то при построении столпов для образа A_k, $k \in \{1,\ldots,K\}$, объекты всех остальных образов объединяются в один виртуальный образ $B_k = \bigcup_{\substack{i=1,\ldots,K \\ i \neq k}} A_i$.

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 надежностью. 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.

Процесс распознавания с опорой на столпы очень прост и состоит в следующем.

1. Находятся расстояния от контрольного объекта *z* до двух ближайших столпов, принадлежащих разным образам.

2. Объект *z* будет принадлежать тому образу, чей стоп оказался ближайшим.

3. По данным расстояниям определяется значение функции конкурентного сходства *F* объекта с образом. По величине *F* можно судить о надежности принятого решения.

3. Гипотеза компактности

Практически все алгоритмы распознавания основаны на использовании гипотезы компактности [3]. К сожалению, строгой формулировки гипотезы и способа количественной оценки компактности образов в литературе нет. Иногда простыми или компактными называются такие образы, которые отделяются друг от друга «*не слишком вычурными*» границами. Описание образов столпами позволяет предложить количественную меру компактности образов.

Компактность образа зависит от того, насколько сильно его объекты похожи на свои столпы и насколько сильно они отличаются от столпов других образов. Эти две характеристики можно определить для каждого объекта в отдельности и тем самым оценить вклад этого объекта в компактность своего образа [4].

В случае двух образов $A = \{a_1, \ldots, a_{M_A}\}$ и $B = \{b_1, \ldots, b_{M_B}\}$ предлагается следующий вариант оценки 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. Обозначим через I_A, $I_A \subseteq \{1, \ldots, M_A\}$, – множество индексов элементов образа *A*, являющихся столпами.

2. Для каждого элемента $a_i \in A$, не являющегося столпом образа *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*:

$$C_{A|B} = \frac{1}{c_A M_A} \sum_{i=1}^{M_A} F(a_i, s_A(a_i) \mid s_B(a_i))|_{i \notin I_A}. \qquad (3)$$

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

4. Далее получим оценку компактности образов *A* и *B* как геометрическое усреднение величин $C_{A|B}$ и $C_{B|A}$.

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

$$C = \sqrt[K]{\prod_{k=1}^{K} C_{A_k|B_k}}. \qquad (4)$$

4. Метод повышения компактности

При построении столпов наряду с объектами, хорошо отражающими структуру образов, принимали участие и шумящие объекты и даже мелкие кластеры таких объектов, влияние которых было бы целесообразно исключить. Для их цензурирования можно применять алгоритм FRiS-Compactor, использующий в качестве критерия, управляющего процессом повышения компактности обучающей выборки, меру FRiS-компактности образов и включающий как составную часть алгоритм FRiS-Stolp.

Опишем алгоритм FRiS-Compactor на примере двух образов *A* и *B*, представленных наборами из M_A и M_B объектов, $M = M_A + M_B$. Компактность образов *C* вычисляется по формулам (3), (4). Через M^* обозначим число объектов обучающей выборки, оставшихся после очередного этапа сокращения выборки. Величину $\left(\frac{M^*}{M}\right)^{\alpha}$, $\alpha \geq 0$, будем использовать в качестве штрафа за исключение объектов из обучающей выборки. С учетом этого компактность H_{AB} образов на каждом шаге сокращения выборки будем оценивать следующим образом:

$$H_{AB} = \left(\frac{M^*}{M}\right)^{\alpha} \sqrt{C_{A|B} C_{B|A}} \,. \qquad (5)$$

Выбор оптимального значения параметра α осуществляется путем сравнения результатов работы алгоритма FRiS-Compactor при разных значениях α. Определим пороги сокращения обучающей выборки: $0 \leq d < 1$ – максимальная доля объектов обучающей выборки, которые можно исключить; m^* – максимальное количество объектов в удаляемом кластере. Положим $M^* = M$.

1. Алгоритмом FRiS-Stolp строятся столпы, стоящие в центрах своих кластеров. По формуле (5) вычисляется компактность H_{AB} образов и заносится в список оценок компактности. Если $M - M^* = dM$, то переход на пункт 7.

2**.** Кластеры с количеством объектов $m \leq m^*$ заносятся в список из *L* кластеров – кандидатов на удаление. Если в выборке нет таких кластеров, то переход на пункт 7. Через $l = 1,\ldots,L$ обозначим номер кластера в сформированном списке. Положим $l = 1$.

3. Из выборки исключается *l*-й кластер, который входит в список и состоит из $m(l)$ объектов. Для оставшихся объектов алгоритмом FRiS-Stolp строятся столпы и вычисляется компактность $H_{AB}(l)$. Элементы *l*-го кластера возвращаются в выборку. Положим $l := l+1$. Если $l \leq L$, то пункт 3 повторяется.

4. После прохода всех *L* кластеров списка выбирается кластер l^*, исключение которого обеспечивало максимальное значение компактности $H_{AB}(l)$: $l^* = \arg\max_{l=1,\ldots,L} H_{AB}(l)$.

5. Если при исключении кластера l^* оказывается превышен порог сокращения обучающей выборки, т. е. $M - M^* + m(l^*) > dM$, то переход на пункт 7.

6. Объекты l^*-го кластера удаляются из выборки, корректируется количество элементов, оставшихся после сокращения выборки. Переход на пункт 1.

7. По списку оценок компактности выбирается вариант, соответствующий максимуму величины H_{AB}. Набор столпов, который был зафиксирован при этом, служит основой решающего правила, используемого для распознавания контрольной выборки. Алгоритм заканчивает работу.

5. Тестирование алгоритма FRiS-Compactor

Алгоритм тестировался 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 дисперсией кластеров, координатами их математических ожиданий и количеством объектов в кластерах, что отражалось на величине FRiS-компактности образов. 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). Таким образом, общее количество экспериментов при различных численных реализациях исходных данных было равно 1000. Максимальное число элементов в удаляемом кластере $m^* = 4$, допустимая доля исключаемых объектов $d = 0{,}15$, т. е. из 100 объектов обучающей выборки разрешалось удалять не более 15 объектов.

По результатам машинного эксперимента было найдено, что оптимальное значение α равно 5. 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.

Обобщенные результаты распознавания контрольной выборки приведены на рис. 2. По оси ординат отложено абсолютное число экспериментов N (из 1000), в которых была достигнута данная надежность P(%). Кривая *1* соответствует надежности без цензурирования, среднее значение равно 91,6 %. Кривая *2* соответствует надежности распознавания с использованием цензурирования. Здесь среднее значение равно 95,9 %. Количество ошибок уменьшилось более чем в два раза.

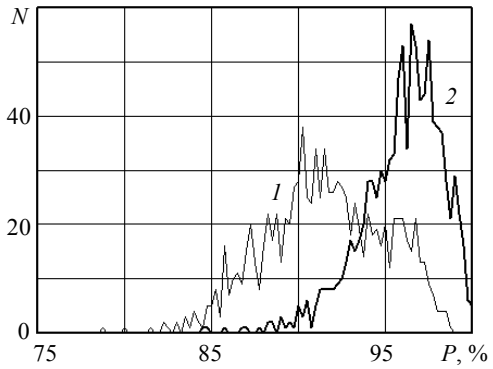

Рис. 2. Распределения надежности распознавания контрольной последовательности

Трудоемкость алгоритма зависит от исходной компактности образов H_0. Чем выше H_0, тем короче список из L претендентов на исключение и тем меньше времени требуется для выбора наилучшего варианта цензурирования. Однако при одном и том же значении H_0 при разных распределениях образов доля исключенных объектов d^*, при которой достигалось максимальное значение компактности образов, меняется в очень больших пределах и предсказать значение d^* по величине H_0 невозможно. Среднее по 1000 экспериментам значение d^* было равно 12,7 %.

Заключение

В работе рассматривается количественная мера компактности образов, основанная на функции конкурентного сходства. Показана полезность применения данной меры сходства для решения задачи цензурирования обучающей выборки. Эффективность предлагаемого метода повышения компактности образов иллюстрируется решением модельной задачи распознавания двух образов. Как показали эксперименты, удаление шумящих объектов из обучающей выборки заметно улучшает результаты распознавания контрольных объектов. Поэтому рекомендуется применять цензурирование выборки при построении решающих правил в задачах распознавания образов.

ЛИТЕРАТУРА

1. *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. No. 1. P. 1–6.
2. *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. Berlin – Heidelberg: Springer-Verlag, 2011. P. 256–270.
3. *Браверман Э.М.* Эксперименты по обучению машины распознаванию зрительных образов // Автоматика и телемеханика. 1962. Т. 23. № 3. С. 349–365.
4. *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.

Загоруйко Николай Григорьевич
Кутненко Ольга Андреевна
Институт математики им. С.Л. Соболева СО РАН (г. Новосибирск)
E-mail: zag@math.nsc.ru; olga@math.nsc.ru
Поступила в редакцию 4 мая 2012 г.

Zagoruiko Nikolay G., *Kutnenko Olga A.* (Sobolev Institute of Mathematics of Siberian Branch of the Russian Academy of Sciences). **Training dataset censoring.**

Keywords: function of rival similarity, compactness, censoring.

The proposed method of compactness increasing is based on the new measure of similarity between objects – function of rival similarity (FRiS-function) – which allows to describe any type of probability distribution with the set of standards. One can estimate contribution of every object of the dataset into compactness of its class, calculate the quantitative measure of compactness of each class separately and compactness of the whole dataset. As well objects, which influence negatively on the compactness value, can be selected. Main idea of proposed method of training dataset censoring consists in removing such objects. As a result the decision rule, constructed on censored dataset, has a better recognition quality. The set of excluded objects is detected automatically. Effectiveness of the censoring algorithm is illustrated by a model task of two classes recognition.