Математическая биология и биоинформатика
2018. Т. 13. № 1. С. 38–49. doi: 10.17537/2018.13.38

========== ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ==========

УДК: 519.95

Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства

Борисова И.А.*, Кутненко О.А.**

Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия

Аннотация. В работе рассматривается задача цензурирования данных из области медицинской диагностики. Предполагается, что в анализируемой выборке могут встречаться ошибочно диагностированные объекты. Подобные объекты оказывают негативное влияние на процедуру анализа данных и поиск содержащихся в них закономерностей, что замедляет процесс получения результатов и ведет к их искажению. Предложенная процедура цензурирования позволяет отыскивать такие объекты и либо удалять их, либо исправлять ошибки в диагностическом (целевом) признаке. Исправление ошибок предпочтительнее в том случае, когда исходная выборка мала, так как это позволяет максимально сохранить полезную информацию, содержащуюся в выборке. Для решения поставленной задачи используется функция конкурентного сходства, с помощью которой оценивается локальное сходство объектов со своими ближайшими соседями. Будучи усредненными по всей выборке, величины локального сходства дают представление о том насколько сильно различаются классы объектов с разными диагнозами на основе имеющихся данных. При этом предполагается, что если в выборке присутствуют неверно диагностированные объекты, то их сходство с ближайшими аналогами из своего класса низкое, и их исключение или коррекция целевого признака позволит увеличить общую разделимость выборки. Процедура коррекции-фильтрации неверно диагностированных объектов основана на наблюдении за изменениями в оценке разделимости классов, вычисленной до и после внесения исправлений в выборку. Процесс цензурирования продолжается до достижения точки перегиба функции разделимости. Для тестирования предложенного метода использовался ряд модельных задач различной сложности. Кроме того этот метод применялся к задачам диагностики диабета, рака груди по результатам биопсии, болезни Паркинсона по нарушениям речи. Предложенный метод показал высокую чувствительность по отношению к ошибочно диагностированным объектам, а исправление таких ошибок позволило улучшить качество классификации при незначительном сокращении объема обучающей выборки.

Ключевые слова: *распознавание образов, функция конкурентного сходства, компактность образов, разделимость классов, цензурирование объектов.*

1. ВВЕДЕНИЕ

Развитие технологий в современном мире приводит к экспоненциальной скорости роста объема информации в самых разных областях науки, в том числе и в биологии. Большое количество различных исследовательских групп по всему миру решают сходные задачи, собирают данные по близким тематикам. Агрегация такого рода

*biamia@mail.ru
**olga@math.nsc.ru

данных в единые базы с одной стороны открывает новые возможности для решения важнейших задач генетики и медицины. Но, с другой стороны, все сложнее становится контролировать источники происхождения этих данных, их качество и надежность.

Таким образом, все выше становится риск появления разного рода ошибок в анализируемых данных, что, как правило, приводит к значительному ухудшению качества получаемых на основе этих данных выводов, снижению подтверждаемости обнаруженных закономерностей. Проблема редактирования и очистки данных (Data filtering, Data cleaning) приобретает все большую актуальность при решении самых разных прикладных задач [1–3]. В большинстве стандартных пакетов анализа данных для улучшения качества входных данных используются алгоритмы фильтрации шумовых объектов.

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 [4, 5]. В других алгоритмах предусмотрена предобработка данных, в процессе которой шумовые объекты с помощь некоторого критерия выявляются и отфильтровываются [6–11]. В некоторых случаях предпринимается попытка корректировки отдельных признаков объекта-выброса с целью преобразовать его в типичный объект [4, 12, 13]. Большинство процедур выявления шумовых объектов основано на использовании правила ближайшего соседа, признавая выбросами объекты, неверно распознаваемые по своим ближайшим соседям [14–16]. Среди последних публикаций достаточно полный обзор по существующим методам цензурирования шумовых объектов сделан в [17].

Стратегия фильтрации испорченных объектов себя оправдывает, если данных много, и даже значительное уменьшение объемов выборки после процедуры фильтрации не сказывается на ее представительности. Однако на практике довольно часто исследователям приходится иметь дело с задачами, количество наблюдений в которых невелико. В такой ситуации фильтрация может привести к значительному снижению представительности выборки. Поэтому если объект-выброс стал таковым в результате ошибки измерения одной или нескольких характеристик, а исходная выборка мала, то такие объекты целесообразно не отфильтровывать, а корректировать значения соответствующих признаков, и уже исправленные объекты использовать для дальнейшей работы.

В работе рассматривается специфичная модель шума, представленная объектами-выбросами, образованными искажением номинального целевого признака. Подобные ошибки в данных могут возникать из-за сложности и недостаточной изученности решаемых проблем, низкой квалификации экспертов, принимающих решения, из-за сбоев в работе измерительных приборов. Для очистки данных от такого рода шумов предлагается новый подход, основанный на использовании функции конкурентного сходства. Его универсальность заключается в том, что объекты-выбросы, которые поддаются коррекции, исправляются, а шумовые объекты, которые скорректировать не удалось, отфильтровываются. Фильтрация или исправление объектов осуществляется на основе изменения оценки разделимости классов до и после коррекции-фильтрации объектов. Процесс цензурирования останавливается при достижении максимального значения оценки разделимости выборки или при отсутствии объектов-выбросов в оставшейся части выборки.

Для тестирования предложенной методики коррекции-фильтрации ошибочно диагностированных объектов использовались как модельные, так и реальные задачи, связанные с анализом медицинских данных.

2. ЦЕНЗУРИРОВАНИЕ (КОРРЕКЦИЯ-ФИЛЬТРАЦИЯ) ОБЪЕКТОВ-ВЫБРОСОВ С ПОМОЩЬЮ ФУНКЦИИ КОНКУРЕНТНОГО СХОДСТВА

2.1. Функция конкурентного сходства

В распознавании образов существует целый пласт эвристических алгоритмов, оперирующих понятием близости объектов, сходства объектов с некими классами. В их основе лежит очевидный человеческий принцип классификации: «на какой класс больше похож, к тому классу и относишься». Но чтобы этот принцип работал максимально эффективно, необходима мера оценки сходства максимально приближенная к той, что человек использует в своей когнитивной деятельности.

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 его расстояние до образа A, но и расстояние до конкурирующего образа B. Следовательно, сходство в распознавании образов является категорией не абсолютной, а относительной.

Для вычисления конкурентного сходства объекта z с объектом a в конкуренции с объектом b с опорой на некоторую метрику r, определяющую расстояния между этими объектами, предлагается использовать тернарную относительную меру, которая называется функцией конкурентного сходства или FRiS-функцией (Function of Rival Similarity) [18]:

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

Данная функция обладает следующими свойствами:

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

и хорошо согласуется с механизмом восприятия сходства (различия), которым пользуется человек [19].

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

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

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. Очевидно, что независимо от используемого расстояния $F(z, A\,|\,B)$ обладает теми же свойствами, что и $F(z,a|b)$.

2.2. FRiS-компактность. Оценка разделимости классов

Понятие компактности образов [20, 21] в том или ином виде используется во многих алгоритмах распознавания. При этом в зависимости от модели образы признаются компактными при выполнении одного из нижеперечисленных условий: если они имеют простую форму, разделяются границей простой формы, объекты одного образа похожи друг на друга и не похожи на объекты других образов. Для получения количественной оценки компактности каждого образа в отдельности предлагается использовать описанную выше FRiS-функцию. С ее помощью формализуется представление о компактности образа, в соответствии с которым

Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.38

«внутреннее» сходство его объектов друг с другом велико, а «внешнее» сходство с объектами других образов мало.

Действительно, для произвольного объекта $a \in A$ мера конкурентного сходства этого объекта 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. Поэтому при решении задачи распознавания FRiS-функция может интерпретироваться как оценка вероятности принадлежности объекта z к образу A. Если усреднить значения FRiS-функции из (1) по всем объектам образов A и B, то можно вычислить важную характеристику решаемой задачи распознавания – некую эмпирическую оценку надежности распознавания образов, аналогом которой в других источниках (см., например, [22]) выступает отделимость классов, компактность, сложность выборки и т. д.:

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

В данной работе эта величина дальше будет называться FRiS-компактностью выборки.

Предположение, легшее в основу предложенного метода, заключается в том, что если в выборке присутствуют неверно диагностированные объекты, то их сходство со своим классом низкое, и их исключение или коррекция целевого признака позволит увеличить общую компактность выборки.

Однако напрямую использовать величину компактности F_{AB} для целей цензурирования нельзя, так как с ростом числа исключенных или исправленных объектов неизбежно повышается переобученность метода, и результаты становятся все менее достоверными. Для учета этого эффекта используется нормирующий коэффициент M^*/M, где M – исходное число объектов в выборке, а M^* – число объектов, оставшихся неизменными после процедуры коррекции-фильтрации. В результате итоговая оценка разделимости классов A и B выглядит следующим образом:

$$G_{AB} = \frac{M^*}{M} \times F_{\mathrm{AB}} = \frac{M^*}{M} \times \frac{\sum_{a \in \mathrm{A}} F(a, \mathrm{A} \mid \mathrm{B}) + \sum_{b \in \mathrm{B}} F(b, \mathrm{B} \mid \mathrm{A})}{|\,\mathrm{A} \cup \mathrm{B}|}, \qquad (3)$$

где $\mathrm{A} \cup \mathrm{B}$ – множество объектов, получившееся в результате проведения коррекции-фильтрации на исходном множестве $A \cup B$, $M = |\,A \cup B\,|$.

2.3. Постановка задачи

Дана выборка $\{(x_i, y_i)\}_{i=\overline{1,M}}$ размерности M, где $y_i \in \{-1, 1\}$ – номинальный целевой признак. Тогда образ A можно записать как множество объектов $\{(x_i : y_i = 1)\}$, образ B, соответственно, как $\{(x_i : y_i = -1)\}$; $Y = \{y_i\}_{i=\overline{1,M}}$.

Учитывая указанное выше 3-е свойство FRiS-функции, запишем (2) в следующем виде:

$$F_{AB} = \frac{\sum_{i=1}^{M} F(x_i, A \mid B) y_i}{M}.$$

При решении задачи поиска и исправления либо удаления ошибочно диагностированных объектов все изменения касаются только целевого признака. Поэтому целью является нахождение множества $\bar{Y} = \{\bar{y}_i\}_{i=\overline{1,M}}$, где $\bar{y}_i \in \{-1, 0, 1\}$; $\bar{y}_i = 0$ означает, что i-ый объект выборки исключен. Тогда (3) можно записать как

Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.38

$$G_{AB} = \frac{M^*}{M} \times \frac{\sum_{i=1}^{M} F(x_i, A \mid B)\bar{y}_i}{|\mathrm{A} \cup \mathrm{B}|}.$$

Таким образом, для поиска объектов-выбросов при операции коррекции-фильтрации требуется найти множество $\bar{Y}$, на котором достигается максимум оценки разделимости классов A и B:

$$\bar{Y} = \arg \max_{\substack{T=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} G_{AB} = \arg \max_{\substack{T=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} \frac{M^*}{M} \times \frac{\sum_{i=1}^{M} F(x_i, A \mid B) t_i}{|\mathrm{A} \cup \mathrm{B}|} =$$

$$= \arg \max_{\substack{T=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} \sum_{i=1}^{M} (t_i^2 + t_i y_i) \times \frac{\sum_{i=1}^{M} F(x_i, \{x_i : t_i = 1\} \mid \{x_i : t_i = -1\}) t_i}{\sum_{i=1}^{M} t_i^2}.$$

Отметим, что здесь $\sum_{i=1}^{M} t_i^2 = |\mathrm{A} \cup \mathrm{B}|$ – число объектов выборки после проведения операции коррекции-фильтрации равное M минус число удаленных объектов; $\sum_{i=1}^{M} t_i y_i$ – число объектов выборки, не участвовавших в операции коррекции-фильтрации, минус число скорректированных объектов (если $t_i = y_i$, то $t_i y_i = 1$, и соответственно, если $t_i = -y_i$, тогда $t_i y_i = -1$). Отсюда следует, что $M^* = \frac{1}{2} \times \sum_{i=1}^{M} (t_i^2 + t_i y_i)$.

При операции чистой фильтрации для поиска объектов-выбросов требуется найти, соответственно, множество $\bar{Y}$:

$$\bar{Y} = \arg \max_{\substack{T=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} G_{AB} = \arg \max_{\substack{T=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} \sum_{i=1}^{M} F(x_i, \{x_i : t_i = 1\} \mid \{x_i : t_i = -1\}) t_i .$$

2.4. Процедура коррекции-фильтрации ошибочно диагностированных объектов

При работе со специфичным видом искажений – ошибками в целевом признаке, появившимися в результате неверной диагностики объектов, появляется возможность вместо полной фильтрации объектов корректировать их и тем самым сохранять значимую информацию об обучающей выборке. Так как на практике не всегда известно, явилась ли ошибка диагностики причиной появления того или иного объекта-выброса, алгоритм коррекции целевого признака должен срабатывать только в случае улучшения качества описания классов в сравнении со стандартным удалением объекта-выброса из выборки.

Процедура коррекции-фильтрации выглядит следующим образом. Поочередно перебираются наиболее нетипичные объекты выборки. Нетипичными будем считать объекты, для которых сходство F со своим образом не больше порогового значения F^*. Для выбранного нетипичного объекта вычисляется величина компактности выборки в случае изменения класса объекта и величина компактности выборки в случае удаления этого объекта. Если корректировка объекта повышает компактность выборки в сравнении с фильтрацией, то целевой признак исправляется, иначе объект удаляется из выборки. Процесс останавливается при достижении точки перегиба в характеристике разделимости выборки или при отсутствии нетипичных объектов в оставшейся части выборки.

2.5. Алгоритмы

Для решения поставленной задачи разработаны алгоритм коррекции-фильтрации FRiS-CFL (FriS-Correction-Filtering Local) и алгоритм фильтрации FRiS-FL (FriS-

Filtering Local), основанные на локальном сходстве объектов выборки с k ближайшими соседями.

Опишем эвристический алгоритм коррекции-фильтрации на примере двух образов A и B. Пусть размерность выборки равна $M = |A \cup B|$. Через G_{AB}^{i} обозначим оценку разделимости классов, полученную на i-ой итерации алгоритма. Под итерацией понимается однократное выполнение шагов с 1-го по 5-ый. В качестве расстояния r от объекта до образа берется среднее расстояние до k ближайших объектов образа.

Алгоритм FRiS-CFL:

Шаг 0. Положим $i = 1$, $M^* = M$, $\mathsf{A} = A$, $\mathsf{B} = B$, $G_{AB}^{0} = -1$.

Шаг 1. По (3) вычисляется оценка разделимости классов G_{AB}^{i}. Если $G_{AB}^{i} < G_{AB}^{i-1}$, то переход на шаг 4.

Шаг 2. Для всех объектов $z \in \mathsf{A} \cup \mathsf{B}$ по (1) вычисляется конкурентное сходство F со своим образом и определяется объект $z^* \in \mathsf{A} \cup \mathsf{B}$ с минимальным значением F, $F \le F^*$: $z^* = \arg\min\limits_{z \in \mathsf{A} \cup \mathsf{B}} \left\{ F : \forall z \in \mathsf{A}\ F(z, \mathsf{A} \mid \mathsf{B}) \le F^*,\ \forall z \in \mathsf{B}\ F(z, \mathsf{B} \mid \mathsf{A}) \le F^* \right\}$. Если такого объекта нет, то переход на шаг 5.

Шаг 3. К выбранному объекту z^* применяется процедура коррекции-фильтрации описанная в разделе 2.4: по (2) вычисляется величина компактности выборки в случае удаления этого объекта и величина компактности в случае изменения класса объекта. Если корректировка объекта повышает компактность выборки в сравнении с фильтрацией, то целевой признак исправляется и z^* остается в выборке, иначе объект удаляется: $\mathsf{A} \cup \mathsf{B} := \{\mathsf{A} \cup \mathsf{B}\} \setminus \{z^*\}$. Положим $i := i + 1$, $M^* := M^* - 1$. Переход на шаг 1.

Шаг 4. Фиксируется набор $\mathsf{A} \cup \mathsf{B}$, построенный на ($i$–1)-ой итерации алгоритма. Переход на шаг 6.

Шаг 5. Фиксируется набор $\mathsf{A} \cup \mathsf{B}$, построенный на i-ой итерации алгоритма.

Шаг 6. Алгоритм заканчивает работу.

Алгоритм чистой фильтрации FRiS-FL отличается от FRiS-CFL только действиями на шаге 3:

Шаг 3. Выбранный объект z^* удаляется: $\mathsf{A} \cup \mathsf{B} := \{\mathsf{A} \cup \mathsf{B}\} \setminus \{z^*\}$. Положим $i := i + 1$, $M^* := M^* - 1$. Переход на шаг 1.

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Цель данной работы – разработка алгоритма коррекции-фильтрации неверно диагностированных объектов. Предлагаемый метод должен давать результаты соизмеримые с методами полной фильтрации. Учитывая, что на практике исследователь, как правило, не знает, содержат ли анализируемые данные ошибки диагностики и каков уровень этой ошибки, применение разработанного алгоритма не должно существенно ухудшать характеристики информативности диагностического метода и качество распознавания при применении его к «чистой» (незашумленной) выборке.

Тестирование разработанного алгоритма проводилось как на модельных, так и на реальных медико-биологических данных [23–25]. При этом рассматривались процедуры коррекции-фильтрации (FRiS-CFL) и чистой фильтрации (FRiS-FL).

Для проверки качества разработанных алгоритмов вычислялись следующие характеристики: надежность распознавания (P) до и после цензурирования, чувствительность (Sns) и специфичность (Spc). Уровень ошибки диагностики (De) в целевом признаке составлял 0 %, 10 %, 20 %, 30 %. Использовались обучающие выборки различного объема (по M/2 объектов первого и второго образов, M = 20, 40,

80, 160, 320). В каждом эксперименте выборка 100 раз случайным образом делилась на обучающую (M объектов первого и второго образов) и контрольную (оставшиеся объекты). Число ближайших соседей $k = 3$. Пороговое значение $F^* = 0$.

К показателям информативности диагностических методов относятся:
чувствительность – способность метода давать правильный результат (доля истинно положительных результатов):

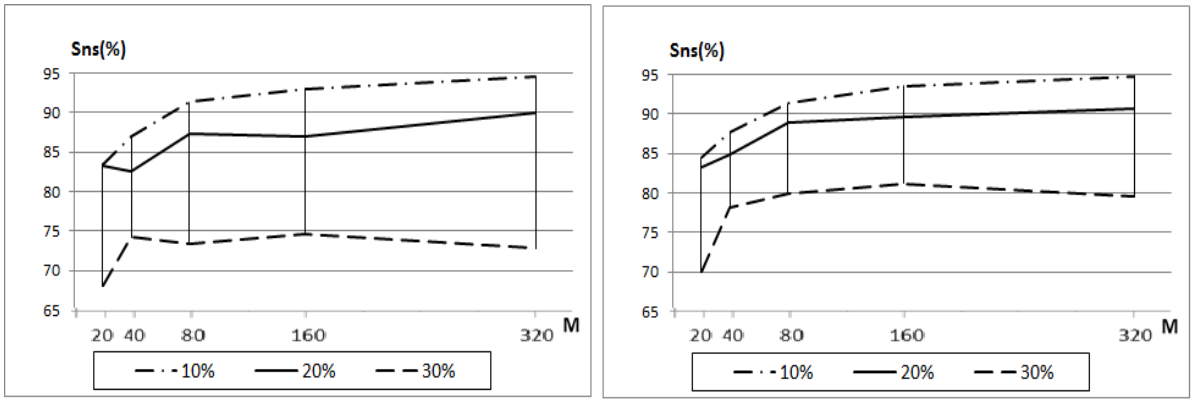
$$Sns = \frac{TP}{TP + FN} \times 100\%; \quad (4)$$

специфичность – способность метода не давать ложноположительных результатов (доля истинно отрицательных результатов):

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

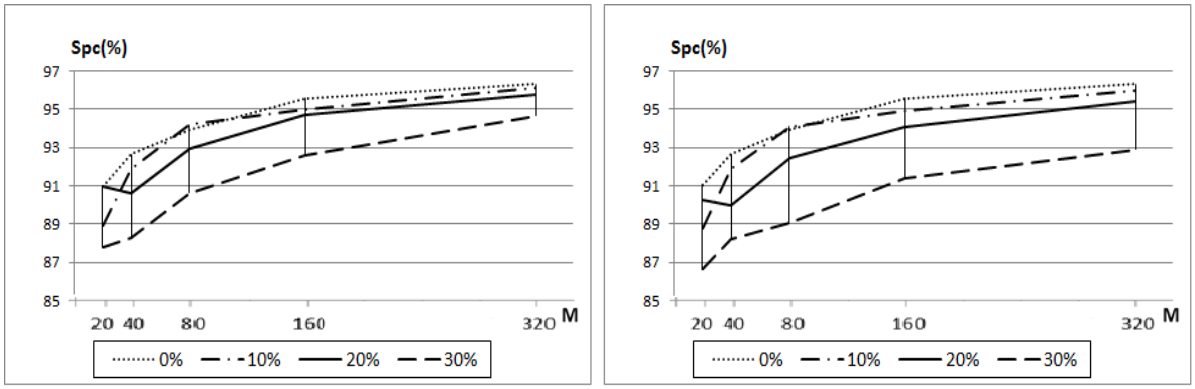
где TP – количество истинно положительных результатов, FN – число ложноотрицательных результатов, TN – количество истинно отрицательных результатов, FP – число ложноположительных результатов.

Ниже на рисунках 1–4 представлены результаты тестирования разработанного метода на биомедицинской задаче диагностики рака груди по результатам биопсии [23]. Выборка состоит из 569 объектов, описанных в пространстве 31 признака. Для моделирования ошибок диагностики целевой признак (диагноз) случайным образом менялся для части объектов.



а) б)

Рис. 1. Зависимость чувствительности от уровня ошибки диагностики *De* и размера выборки *M* после применения алгоритма коррекции-фильтрации **(а)** и алгоритма фильтрации **(б)**.



а) б)

Рис. 2. Зависимость специфичности от уровня ошибки диагностики *De* и размера выборки *M* после применения алгоритма коррекции-фильтрации **(а)** и алгоритма фильтрации **(б)**.

Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.38

При использовании алгоритма коррекции-фильтрации объектов-выбросов чувствительность метода незначительно ниже чувствительности, полученной при применении алгоритма чистой фильтрации; специфичность, напротив, незначительно выше (см. рисунки 1 и 2). При уровне ошибки в целевом признаке $De = 0\ \%$ чувствительность метода согласно (4) не определена.

На рисунке 3 показаны ожидаемый объем выборки $\tilde{M}$ и ожидаемая ошибка диагностики данных $\tilde{D}e$ после применения процедуры коррекции-фильтрации и процедуры полной фильтрации. Ошибка диагностики формируется как сумма оставшихся исходных ошибок и ошибок, привнесенных коррекцией целевых признаков объектов-выбросов. Как коррекция-фильтрация, так и чистая фильтрация, существенно снижают уровень ошибки диагностики. При этом применение процедуры коррекции-фильтрации сохраняет большую часть анализируемой выборки при практически том же уровне ожидаемой ошибки диагностирования, что и применение процедуры фильтрации. Например, при $M = 320$ и $De = 30\ \%$ для процедуры коррекции-фильтрации $\tilde{M} \approx 290.06$, $\tilde{D}e \approx 10.48$; для процедуры фильтрации – $\tilde{M} \approx 227.61$, $\tilde{D}e \approx 8.59$.

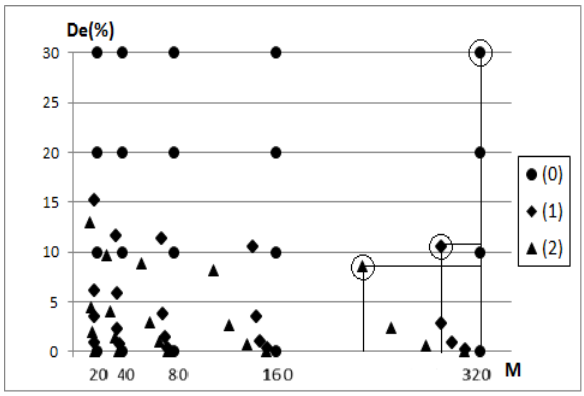
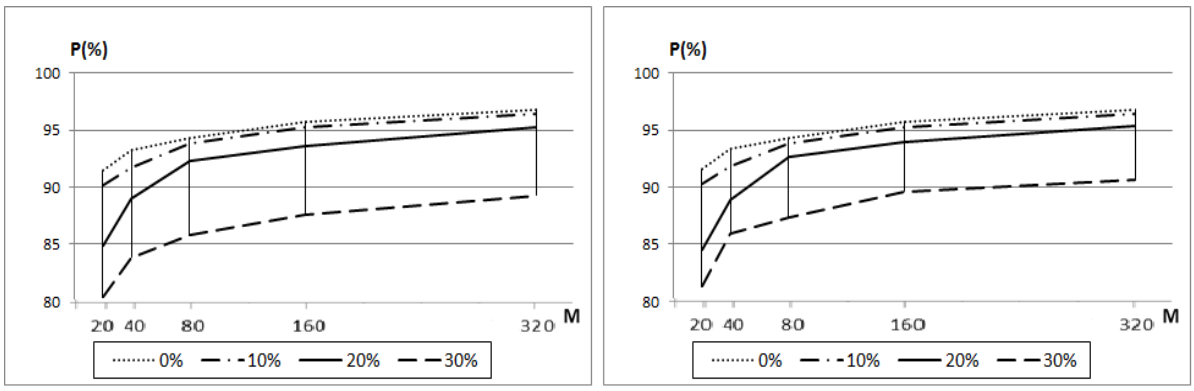


Рис. 3. Ожидаемый объем выборки и ожидаемая ошибка диагностики исходных данных (**0**) после цензурирования: (**1**) – процедура коррекции-фильтрации, (**2**) – процедура фильтрации.

В случае наличия в выборке неверно диагностированных объектов как коррекция-фильтрация, так и чистая фильтрация объектов-выбросов, приводят к существенному улучшению качества распознавания по сравнению с надежностью распознавания на исходной выборке; в случае не зашумленной выборки результаты распознавания по k ближайшим соседям на исходных данных, как и ожидаемо, незначительно лучше. Результаты распознавания процедур коррекции-фильтрации и чистой фильтрации сопоставимы (см. рис. 4).



а) б)

Рис. 4. Зависимость надежности распознавания P от уровня ошибки диагностики De и размера выборки M после применения алгоритма коррекции-фильтрации **(а)**, алгоритма фильтрации **(б)**.

В таблице 1 представлена относительная разница в надежности распознавания после применения алгоритма коррекции-фильтрации (ΔP_1) и алгоритма полной фильтрации (ΔP_2) с надежностью распознавания по исходным данным.

Таблица 1. Относительная разница в надежности распознавания

De	**0 %**		**10 %**		**20 %**		**30 %**	
$M/\Delta P$	ΔP_1	ΔP_2	ΔP_1	ΔP_2	ΔP_1	ΔP_2	ΔP_1	ΔP_2
20	–1.99	–1.86	–1.25	–1.16	1.79	1.51	2.95	3.89
40	–1.11	–1.06	0.00	0.11	2.99	2.92	7.80	9.88
80	–0.70	–0.69	1.34	1.30	5.71	6.10	9.06	10.57
160	–0.06	–0.03	1.47	1.50	6.69	7.03	11.44	13.48
320	0.35	0.35	2.48	2.48	7.87	7.99	12.92	14.28

Проведенные эксперименты при разных значениях порога F^* (–0.3, –0.1, 0, 0.1, 0.3) показали, что повышение порога $F^*>0$ не улучшает такие характеристики метода как надежность распознавания, чувствительность, качество очищенной выборки (ожидаемые длина выборки и ошибка диагностики), при этом незначительно ухудшается специфичность метода. Понижение порога $F^*< 0$ позволяет улучшить специфичность метода, но при этом существенно ухудшает выше перечисленные характеристики.

На рисунке 5 представлены результаты тестирования стандартных алгоритмов цензурирования ENN и RENN [26] и предложенных в работе алгоритмов коррекции-фильтрации и чистой фильтрации на указанной выше биомедицинской задаче. Доля ошибочно диагностированных объектов во всех экспериментах составила 30 %.

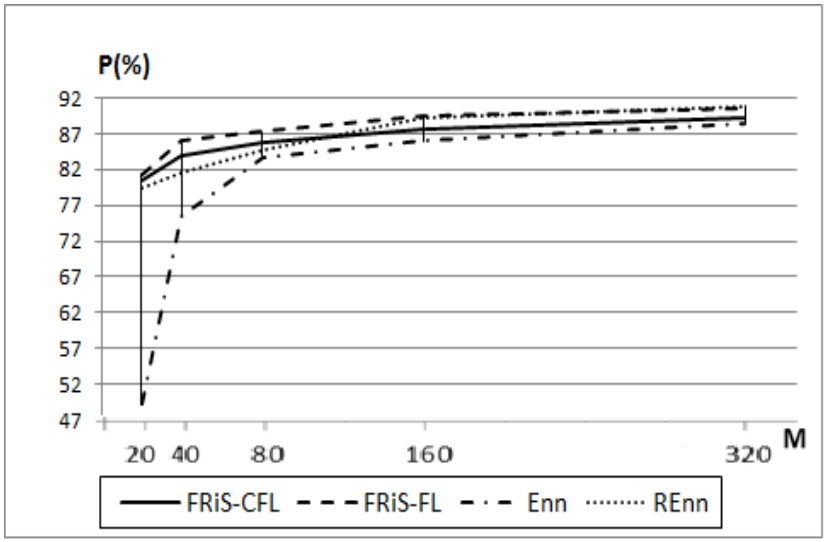


Рис. 5. Сравнение алгоритмов цензурирования – надежность распознавания P.

В таблице 2 представлены результаты тестирования разработанного алгоритма чистой фильтрации (FRiS-FL), алгоритмов цензурирования ENN, RENN и алгоритма knn, см., например, [27] (k nearest neighbors, k = 3) для следующих задач: распознавания диабета [24] (выборка состоит из 538 объектов, описанных в пространстве девяти признаков), распознавания болезни Паркинсона [25] (выборка состоит из 195 объектов, описанных в пространстве 23-х признаков), модельной задачи распознавания двух образов (выборка состоит из 1000 объектов, описанных в пространстве шести признаков). Доля ошибочно диагностированных объектов в данных экспериментах составляла 20 %.

Таблица 2. Надежность распознавания

	Диабет				Паркинсон				Модельная задача			
Алг. M	Enn	Renn	FRiS FL	knn	Enn	Renn	FRiS-FL	knn	Enn	Renn	FRiS-FL	knn
20	39.60	65.71	**67.08**	65.30	46.13	68.68	63.48	**67.42**	57.39	55.09	63.33	**67.19**
40	60.80	66.01	**68.56**	65.30	70.95	69.39	66.63	**72.35**	66.89	62.80	69.33	**70.63**
80	65.49	66.57	**69.47**	65.45	73.83	71.17	70.77	**76.42**	71.89	70.47	**74.00**	72.17
160	68.29	70.30	**74.53**	67.02	–	–	–	–	74.97	74.26	**77.45**	73.44
320	67.72	69.73	**82.35**	68.22	–	–	–	–	76.93	77.34	**79.32**	73.80

Более детальный анализ полученных результатов показал, что предложенный метод обладает сопоставимой чувствительностью и значительно более высокой степенью специфичности по сравнению с рассмотренными алгоритмами цензурирования.

ЗАКЛЮЧЕНИЕ

Проблема представительности данных актуальна для многих прикладных задач, в том числе и для задач, связанных с анализом биомедицинской информации. Предложенный в данной статье подход коррекции-фильтрации является эффективным инструментом повышения представительности выборки как за счет удаления из нее объектов-выбросов, искажающих представление о скрытых закономерностях, так и за счет исправления отдельных ошибок, что позволяет сохранить максимальное количество информации о задаче.

Разработанный на основе FRiS-методологии подход к решению задачи цензурирования продемонстрировал свою конкурентоспособность в сравнении со стандартными методами, используемыми для этих целей. Проведенные исследования показали, что предложенные алгоритмы обладают высокой чувствительностью по отношению к объектам, содержащим ошибки в целевом признаке. Это позволяет использовать как алгоритм коррекции-фильтрации, так и алгоритм полной фильтрации, для снижения уровня ошибки диагностики данных. Учитывая, что корректирующие методы в отличие от фильтрующих позволяют сохранить значительно большую часть выборки, можно сделать вывод о потенциальной эффективности применения алгоритма коррекции-фильтрации в ситуациях, требующих минимального сокращения объема выборки в процессе предобработки. Предложенные методы могут быть использованы при проектировании систем интеллектуального анализа данных для повышения как достоверности их работы, так и качества обучения.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 16-07-00168.

СПИСОК ЛИТЕРАТУРЫ

1. de Waal T., Pannekoek J., Scholtus S. *Handbook of Statistical Data Editing and Imputation.* John Wiley and Sons, Inc. Hoboken, New Jersey, 2011. 456 p. doi: 10.1002/9780470904848.ch1
2. Jason W. Osborne. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* 1st Edition. SAGE Publication, Inc. Los Angeles, 2013. 296 p. doi: 10.4135/9781452269948
3. Luca Greco. *Robust Methods for Data Reduction Alessio Farcomeni*. Chapman and Hall/CRC, 2015. 297 p.

4. Teng C.M. A comparison of noise handling techniques. In: *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference.* 2001. P. 269–273.
5. Quinlan J.R. Induction of decision trees. *Machine Learning.* 1986. P. 81–106. doi: 10.1023/A:1022643204877, doi: 10.1007/BF00116251.
6. Frйnay B., Verleysen M. Classification in the Presence of Label Noise: a Survey. *IEEE Transactions on neural networks and learning systems.* 2014. V. 25. № 5. P. 845–869. doi: 10.1109/TNNLS.2013.2292894
7. Segata N., Blanzieri E. Noise Reduction for Instance-Based Learning with a Local Maximal Margin Approach. *Journal of Intelligent Information Systems 35 (October).* 2010. doi: 10.1007/s10844-009-0101-z
8. Massie S., Craw S., Wiratunga N. 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. 2007. P. 92–106. doi: 10.1007/978-3-540-74141-1_7
9. Son S.-H., Kim J.-Y. 2006. Data Reduction for Instance-Based Learning Using Entropy-Based Partitioning. In: *Proceedings of the International Conference on Computational Science and Its Applications.* 2006. P. 590–599. doi: 10.1007/11751595_63
10. Delany S.J., Segata N., Mac Namee B. Profiling Instances in Noise Reduction. *Knowledge-Based Systems 31 (July).* 2012. P. 28–40. doi: 10.1016/j.knosys.2012.01.015
11. Борисова И.А., Кутненко О.А. Цензурирование ошибочно классифицированных объектов выборки. *Машинное обучение и анализ данных.* 2015. Т. 1. № 11. С. 1632–1641.
12. Yang Y., Wu X., X. Zhu. 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. 2004. doi: 10.1007/978-3-540-30116-5_43
13. Brodley C.E, Friedl M.A. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research.* 1999. № 11. P. 131–167.
14. Wilson D.R., Martinez T.R. Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning.* 2000. V. 38. № 3. P. 257–286. doi: 10.1023/A:1007626913721
15. Jankowski N., Grochowski M. Comparison of Instances Seletion Algorithms I. Algorithms Survey. *Artificial Intelligence and Soft Computing.* 2004. P. 1–6. doi: 10.1007/978-3-540-24844-6_90
16. Brighton H., Mellish C. Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery 6.* 2002. P. 153–172. doi: 10.1023/A:1014043630878
17. Aggarwal C.C. Outlier analysis. *Data Mining.* Springer International Publishing. 2015. P. 237–263. doi: 10.1007/978-3-319-14142-8_8
18. 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 Analysis.* 2008. V. 18. № 1. P. 1–6. doi: 10.1134/S105466180801001X
19. Загоруйко Н.Г. *Когнитивный анализ данных.* Новосибирск: Академическое изд-во ГЕО, 2013. 186 с.
20. Загоруйко Н.Г. *Прикладные методы анализа данных и знаний.* Новосибирск: Изд. ИМ СО РАН, 1999. 270 с.

21. Загоруйко Н.Г., Борисова И.А., Кутненко О.А., Дюбанов В.В. Построение сжатого описания данных с использованием функции конкурентного сходства. *Сибирский журнал индустриальной математики.* 2013. Т. XVI. № 1(53). С. 29–41.
22. 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.
23. *Breast Cancer Wisconsin (Diagnostic) Data Set*. URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (accessed July 2016).
24. *Pima Indians Diabetes Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes (accessed July 2016).
25. *Parkinsons Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Parkinsons (accessed July 2016).
26. Wilson D.R., Martinez T.R. Reduction Techniques for Instance-Based Learning Algorithms. *Machine learning*. 2000. V0. 38. № 3. P. 257–286. doi: 10.1023/A:1007626913721
27. Фукунага К. *Введение в статистическую теорию распознавания образов.* М.: Наука, 1979. 368 с. (Перевод с англ. Keinosuke Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, 1972)

Рукопись поступила в редакцию 31.01.2018.
Дата опубликования 27.03.2018.