Математическая биология и биоинформатика
2019. Т. 14. № 2. С. 464–476. doi: 10.17537/2019.14.464

==========ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
==========ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ==========

УДК: 519.95

Очистка данных от диагностических ошибок в признаковых пространствах большой размерности

Борисова И.А.*, Кутненко О.А.**

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

Аннотация. В статье предлагается новый подход к цензурированию данных, позволяющий очищать выборки от диагностических ошибок в целевом признаке в случае, когда эти выборки описаны в признаковых пространствах большой размерности. Рассмотрение данного случая как отдельной задачи объясняется тем, что в пространствах большой размерности перестают работать большинство методов цензурирования и очистки данных, как статистических, так и метрических. При этом для задач медицинской диагностики, учитывая сложность изучаемых объектов и явлений, большое количество описывающих характеристик является скорее нормой, чем исключением. Для решения поставленной задачи предложен подход, ориентированный на локальное сходство между собой объектов выборки и использующий в качестве меры сходства функцию конкурентного сходства (FRiS-функцию). В предложенном подходе для эффективной очистки данных от ошибок происходит выбор наиболее информативного и релевантного решаемой задаче признакового подпространства малой размерности, в котором разделимость классов после их корректировки будет максимальна. Под разделимостью классов понимается похожесть объектов одного класса друг на друга и их непохожесть на объекты другого классов. Очистка от ошибок может выражаться как в их исправлении, так и в удалении испорченных объектов из выборки. Описанный метод был реализован в виде алгоритма FRiS-LCFS (FRiS Local Censoring with Feature Selection) и протестирован на модельных и реальных биомедицинских задачах, в том числе и на задаче диагностики рака простаты по результатам измерения генной активности. Разработанный алгоритм показал свою конкурентоспособность по сравнению со стандартными методами, фильтрации данных в пространствах большой размерности.

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

1. ВВЕДЕНИЕ

Новейшие методы и технологии обработки данных Data Mining позволяют исследователям работать с выборками, размерность и сложность которых ранее делала их практически непригодными для анализа. Использование этих технологий открывает новые возможности для решения важнейших задач биоинформатики, генетики, медицины и других естественных наук, где из года в год наблюдается непрерывный рост объема накапливаемой информации.

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

Сложность задач, возникающих при анализе биомедицинских данных, объясняется как свойствами рассматриваемых объектов, так и их слабой изученностью. Множество факторов, которые потенциально могут оказывать влияние на изучаемые явления и процессы, существенно увеличивают размерность задач, маскируют скрытые в данных закономерности и затрудняют попытки их решения.

С другой стороны, важность решаемых проблем приводит к тому, что данными задачами занимается большое количество несвязанных между собою групп исследователей, потому все сложнее становится контролировать качество и надежность информации, собираемой для анализа. Таким образом, все выше становится риск появления разного рода ошибок в данных, что, как правило, приводит к значительному ухудшению качества получаемых на их основе выводов, снижению подтверждаемости обнаруженных закономерностей. Ситуация усугубляется, когда такие ошибки имеются в выборках, включающих избыточное количество описывающих характеристик, часть из которых содержит дублирующую информацию или нерелевантна решаемой задаче.

В стандартных пакетах анализа данных для улучшения качества входных данных используются различные алгоритмы обнаружения объектов-выбросов. Эти алгоритмы условно можно разделить на две большие группы: статистические [1, 2], делающие выводы о наличии ошибок в описании объекта на основе статистических методов проверки гипотез, и метрические [3–5], обнаруживающие шумы в данных, основываясь на сходстве и расстояниях между объектами. Однако обе группы методов перестают работать на выборках большой размерности. Потому для решения таких задач используют третью группу методов – эвристические алгоритмы обнаружения объектов-выбросов в пространствах большой размерности [6]. Костяк этой группы составляют гибридные алгоритмы, снижающие размерность задачи с помощью стандартных алгоритмов выбора признаков, например, Recursive Feature Elimination (RFE) [7], поочередно исключающий самые неинформативные признаки из системы, чтобы затем применять для решения задачи стандартные алгоритмы фильтрации в пространстве малой размерности. К таким, среди прочих, относится алгоритм Local Outlier Factor (LOF) [8], который выявляет объекты-выбросы на основе изменения оценок плотности. Кроме того есть и специфичные алгоритмы, например, Isolation Forest [9], использующий для поиска выбросов длины случайных деревьев, отделяющих друг от друга все объекты выборки, или ABOD [10], предполагающий, что периферийные объекты-выбросы можно выявлять, сравнивая углы между объектами.

После обнаружения в выборке объектов-выбросов наиболее распространенной является стратегия фильтрации таких объектов, однако у нее есть свои слабые стороны. Так например, она может оказаться неоптимальной, если объект-выброс стал таковым в результате ошибки измерения одной или нескольких характеристик. При малых объемах исходной выборки такие объекты целесообразно не отфильтровывать, а корректировать значения соответствующих признаков и исправленные объекты использовать для повышения представительности выборки [11, 12].

В рамках данной работы предлагается новый подход к задаче коррекции-фильтрации объектов-выбросов в пространствах большой размерности, основанный на использовании функции конкурентного сходства. Его универсальность заключается в том, что очистка данных от объектов-выбросов и признаков осуществляется параллельно, объекты-выбросы, которые поддаются коррекции, исправляются а остальные отфильтровываются. Корректируются только значения целевого признака, так как именно ошибки в нем оказывают наибольшее влияние на качество получаемых в процессе анализа данных закономерностей. Эффективность предложенной стратегии для задачи очистки данных в пространствах малой размерности была проиллюстрирована в [13], где на примере задачи диагностики рака груди по результатам биопсии было показано, что при использовании процедуры коррекции в

сравнении с фильтрацией удается сохранить большую часть выборки, при этом надежность распознавания у обоих подходов сопоставима.

В предлагаемом подходе обнаружение и исправление объектов в фиксированном признаковом пространстве осуществляется на основе изменения оценки разделимости классов [13], основанной на функции конкурентного сходства, вычисляемой до и после корректировки или фильтрации объектов. Конкурентоспособность этого подхода в сравнении со стандартными алгоритмами очистки данных для пространств малой размерности подтверждается результатами экспериментов в [13, 14] как на модельных задачах, так и на задаче распознавания диабета. Для расширения области применения используемой методики очистки данных на пространства большой размерности предлагается параллельно с очисткой данных от ошибок осуществлять выбор признакового подпространства малой размерности, в котором разделимость классов после очистки будет максимальна.

Для тестирования предложенной методики коррекции-фильтрации, или цензурирования, ошибочно диагностированных объектов в признаковых пространствах большой размерности использовались как модельные так и реальные задачи, связанные с анализом медицинских данных. В работе приведены результаты экспериментов для задачи диагностики рака простаты [15] по результатам измерения генной активности, содержащей 12625 признаков.

2. ЦЕНЗУРИРОВАНИЕ ОБЪЕКТОВ-ВЫБРОСОВ С ПОМОЩЬЮ ФУНКЦИИ КОНКУРЕНТНОГО СХОДСТВА

2.1. Функция конкурентного сходства

В распознавании образов существует достаточно много эвристических алгоритмов, оперирующих понятием близости объектов, сходства объектов с некими классами. При измерении таких характеристик объектов, как вес, длина, сопротивление и т. д. обычно используются эталонные объекты. Результат измерения определяется свойствами только этих двух объектов и не зависит от свойств других объектов, то есть результат измерения имеет характер абсолютной величины. Но объекты могут описываться и такими характеристиками, для которых не существует эталонов, например: «близок-далек», «добрый-злой» и т. д. Чтобы ответить на вопрос «сходны» или «не сходны», «близки» или «далеки» два объекта с не совпадающими свойствами, нужно знать ответ на вопрос «по сравнению с чем?» Хорошо известная фраза «все познается в сравнении» отражает фундаментальный закон познания, то есть адекватная мера сходства должна зависеть от особенностей конкурентного окружения рассматриваемого объекта и должна быть максимально приближенная к той, что человек использует в своей когнитивной деятельности.

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 τ_X, определяющую расстояния между этими объектами в признаковом пространстве X, предлагается использовать тернарную относительную меру, которая называется функцией конкурентного сходства или FRiS-функцией (Function of Rival Similarity) [16]:

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

Данная функция хорошо согласуется с механизмом восприятия сходства (различия), которым пользуется человек [17].

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

$$F_X(z, A \mid B) = \frac{\tau_X(z, B) - \tau_X(z, A)}{\tau_X(z, B) + \tau_X(z, A)}. \quad (1)$$

Отметим, что в зависимости от особенностей решаемой задачи расстояние от объекта до образа может вычисляться по-разному. В качестве него может использоваться и расстояние до ближайшего объекта образа, и среднее расстояние до всех объектов образа, и среднее расстояние до k ближайших объектов образа, и расстояние до центра тяжести образа и т. д.

2.2. FRiS-компактность. Оценка разделимости классов

Понятие компактности образов в том или ином виде используется во многих алгоритмах распознавания. Например, представленное в [18] определение компактности опирается на соотношение количества «внутренних» и «граничных» точек, отражающих объекты образов в пространстве описывающих характеристик (признаков). В [19] вместо одной количественной характеристики вычисляется «профиль компактности», отражающий зависимость числа объектов «своего образа» в локальной окрестности каждого объекта выборки от радиуса этой окрестности. В [20] компетентность кластеров оценивается усреднением квадрата расстояния от объектов до центров кластеров, к которым они относятся.

Для получения количественной оценки компактности каждого образа в отдельности предлагается использовать описанную выше FRiS-функцию [21]. Действительно, для произвольного объекта $a \in A$ мера конкурентного сходства этого объекта со своим образом в конкуренции с образом B в признаковом пространстве X показывает, насколько этот объект похож на свой образ и не похож на образ B в этом пространстве. Если эта величина для всех объектов образа A положительна, то можно считать данный образ компактным, так как подобная ситуация хорошо согласуется с интуитивным представлением о компактности, как о сходстве объектов внутри образа и несходстве их с объектами конкурирующего образа. 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 в признаковом пространстве X, аналогом которой в других источниках (см., например, [22]) выступает отделимость классов, компактность, сложность выборки и т. д.:

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

Назовем эту величину FRiS-компактностью выборки. Заметим, что она может выступать как оценкой разделимости выборки, так и оценкой пригодности признакового пространства X для классификации этой выборки.

Предположение, легшее в основу предложенного метода выявления объектов-выбросов в пространствах большой размерности, заключается в том, что если в выборке присутствуют объекты-выбросы, то их сходство со своим классом низкое даже в информативных признаковых подпространствах небольшой размерности, для которых общая характеристика разделимости выборки, вычисляемая по формуле (2),

высока. Потому их исключение или коррекция позволит увеличить общую компактность выборки в этих подпространствах.

Однако напрямую использовать величину компактности $F_{AB}(X)$ 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 исправления целевого признака или удаления объектов-выбросов. Если обозначить A^* и B^* – новый состав образов A и B, итоговая оценка разделимости классов после цензурирования в пространстве X выглядит следующим образом:

$$G_{AB}(X, A^*, B^*) = \frac{M^*}{M} F_{A^*B^*}(X). \quad (3)$$

2.3. Постановка задачи

Даны два непересекающихся конечных множества объектов A и B, $|A \cup B| = M$; множество признаков X, $|X| = N$, и метрика τ_X, позволяющая вычислять расстояние между любой парой объектов из $A \cup B$ как в пространстве X, так и в любом его подпространстве $X^* \subseteq X$, $|X^*| = n^* \le N$. Для удобства изложения предполагается, что пространство X нормировано таким образом, что расстояние между любой парой объектов не превышает 1. Требуется найти такие непересекающиеся множества $A^* \subseteq A \cup B$, $B^* \subseteq A \cup B$ и множество $X^* \subseteq X$, $|X^*| = n^*$, которые обеспечат, согласно (2) и (3), максимум функции $G_{AB}(X^*, A^*, B^*)$.

Обозначим через $\mathbf{x_i} = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iN})$, $i = \overline{1, M}$, описание i-го объекта в пространстве X. Множество $A \cup B$ можно записать как выборку $\{(\mathbf{x_i}, y_i)\}_{i=\overline{1,M}}$, где $y_i \in \{-1, 1\}$ – номинальный целевой признак. Образ A запишем как множество объектов $\{(\mathbf{x_i}: y_i = 1)\}$, образ B, соответственно, как $\{(\mathbf{x_i}: y_i = -1)\}$; $\mathbf{y} = \{y_i\}_{i=\overline{1,M}}$.

Учитывая свойство асимметричности FRiS-функции: $F(z, a \mid b) = -F(z, b \mid a)$, запишем (2) в следующем виде:

$$F_{AB}(X) = \frac{\sum_{i=1}^{M} F(\mathbf{x_i}, \{(\mathbf{x_i}: y_i = 1)\} \mid \{(\mathbf{x_i}: y_i = -1)\}) y_i}{M}.$$

При решении задачи поиска и исправления ошибочно диагностированных объектов в некотором фиксированном признаковом пространстве X все изменения касаются только целевого признака. Поэтому целью является нахождение множества значений целевого признака $\mathbf{t} = \{t_i\}_{i=\overline{1,M}}$, где $t_i \in \{-1, 0, 1\}$; $t_i = 0$ означает, что i-ый объект выборки исключен. Тогда (3) можно записать как

$$G_{AB}(X, A^*, B^*) = \frac{M^*}{M} \times F_{A^*B^*} = \frac{\sum_{i=1}^{M} (t_i^2 + t_i y_i)}{2M} \times \frac{\sum_{i=1}^{M} F(\mathbf{x_i}, \{(\mathbf{x_i}: t_i = 1)\} \mid \{(\mathbf{x_i}: t_i = -1)\}) t_i}{\sum_{i=1}^{M} t_i^2}.$$

Отметим, что здесь $\sum_{i=1}^{M} t_i^2$ – число объектов в выборке 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 $M^* = \frac{1}{2} \times \sum_{i=1}^{M} (t_i^2 + t_i y_i)$.

Каждому подпространству $X^* \subseteq X$, $|X^*| = n^*$, можно взаимно-однозначно поставить в соответствие вектор $\mathbf{s} = \{s_j\}_{j=\overline{1,N}}$, $\sum_{j=1}^{N} s_j = n^*$, где $s_j = \begin{cases} 1, & \text{если признак } j \in X^* \\ 0, & \text{если признак } j \notin X^* \end{cases}$.

Далее через $X(\mathbf{s})$ будем обозначать признаковое подпространство, заданное вектором $\mathbf{s}$, через $\mathbf{x_i}(\mathbf{s}) = \{s_j x_{ij}\}_{j=\overline{1,N}}$ – описание объекта $\mathbf{x_i}$, $i = \overline{1,M}$, в пространстве $X(\mathbf{s})$. Тогда образ A в данном признаковом пространстве обозначим через $A(\mathbf{s})$ и запишем как множество объектов $\{(\mathbf{x_i}(\mathbf{s}): y_i = 1)\}$, образ B, соответственно, обозначим через $B(\mathbf{s})$ и запишем как $\{(\mathbf{x_i}(\mathbf{s}): y_i = -1)\}$.

Таким образом, для решения задачи коррекции ошибок диагностики в данных, описанных в некотором фиксированном подпространстве $X(\mathbf{s})$, требуется найти множество $\mathbf{t}^*(\mathbf{s})$, на котором достигается, согласно (2) и (3), максимум оценки разделимости классов $A(\mathbf{s})$ и $B(\mathbf{s})$:

$$\mathbf{t}^*(\mathbf{s}) = \arg \max_{\substack{\mathbf{t}=\{t_i\}_{i=\overline{1,M}} \\ t_i \in \{-1,0,1\}}} \sum_{i=1}^{M} (t_i^2 + t_i y_i) \times \frac{\sum_{i=1}^{M} F(\mathbf{x_i}(\mathbf{s}), \{\mathbf{x_i}(\mathbf{s}): t_i = 1\} | \{\mathbf{x_i}(\mathbf{s}): t_i = -1\}) t_i}{\sum_{i=1}^{M} t_i^2},$$

где $\mathbf{x_i}(\mathbf{s}) = \{s_j x_{ij}\}_{j=\overline{1,N}}$ – описание объекта $\mathbf{x_i}$, $i = \overline{1,M}$, в пространстве $X(\mathbf{s})$.

В итоге для решения поставленной задачи коррекции ошибок диагностики в данных, описанных в пространствах большой размерности, требуется найти множества $\mathbf{s}^* = \{s_j^*\}_{j=\overline{1,N}}$, $s_j^* \in \{0,1\}$, $\sum_{j=1}^{N} s_j^* = n^*$ и $\mathbf{t}^* = \{t_i^*\}_{i=\overline{1,M}}$, $t_i^* \in \{-1,0,1\}$, на которых достигается максимум оценки разделимости классов A и B:

$$\{\mathbf{s}^*, \mathbf{t}^*\} =$$

$$= \arg \max_{\substack{\mathbf{s}=\{s_j\}_{j=\overline{1,N}} \\ \mathbf{t}=\{t_i\}_{i=\overline{1,M}}}} \sum_{i=1}^{M} (t_i^2 + t_i y_i) \times \frac{\sum_{i=1}^{M} F(\{s_j x_{ij}\}_{j=\overline{1,N}}, \{\{s_j x_{ij}\}_{j=\overline{1,N}}: t_i = 1\} | \{\{s_j x_{ij}\}_{j=\overline{1,N}}: t_i = -1\}) t_i}{\sum_{i=1}^{M} t_i^2}.$$

При операции чистой фильтрации

$$\{\mathbf{s}^*, \mathbf{t}^*\} = \arg \max_{\substack{\mathbf{s}=\{s_j\}_{j=\overline{1,N}} \\ \mathbf{t}=\{t_i: t_i y_i \in \{0,1\}\}_{i=\overline{1,M}}}} \sum_{i=1}^{M} F(\{s_j x_{ij}\}_{j=\overline{1,N}}, \{\{s_j x_{ij}\}_{j=\overline{1,N}}: t_i = 1\} | \{\{s_j x_{ij}\}_{j=\overline{1,N}}: t_i = -1\}) t_i,$$

где условие $t_i y_i \in \{0,1\}$ исключает операцию коррекции.

В качестве примера выпишем максимизируемую функцию в частном случае. Пусть X подмножество N-мерного Евклидова пространства, т.е. $\tau_X(x,y) = \sqrt{\sum_{j=1}^{N} |x_j - y_j|^2}$; в качестве расстояния от объекта z до образа возьмем расстояние до ближайшего объекта данного образа. С учетом введенных выше обозначений будут справедливы следующие равенства:

$$\tau_X(z,B) = \min_{\substack{l=\overline{1,M} \\ l \neq i}} \left\{ \sqrt{\sum_{j=1}^{N} |z_j - x_{ij}|^2} + y_l \right\} + 1, \qquad \tau_X(z,A) = \min_{\substack{l=\overline{1,M} \\ l \neq i}} \left\{ \sqrt{\sum_{j=1}^{N} |z_j - x_{ij}|^2} - y_l \right\} + 1.$$

Математическая биология и биоинформатика. 2019. Т. 14. № 2. doi: 10.17537/2019.14.464

Тогда (1) запишется как

$$F_X(z, A \mid B) = \frac{\min\limits_{\substack{l=\overline{1,M}\\ l\neq i}}\left\{\sqrt{\sum_{j=1}^{N}|\, z_j - x_{ij}\,|^2} + y_l\right\} - \min\limits_{\substack{l=\overline{1,M}\\ l\neq i}}\left\{\sqrt{\sum_{j=1}^{N}|\, z_j - x_{ij}\,|^2} - y_l\right\}}{\min\limits_{\substack{l=\overline{1,M}\\ l\neq i}}\left\{\sqrt{\sum_{j=1}^{N}|\, z_j - x_{ij}\,|^2} + y_l\right\} + \min\limits_{\substack{l=\overline{1,M}\\ l\neq i}}\left\{\sqrt{\sum_{j=1}^{N}|\, z_j - x_{ij}\,|^2} - y_l\right\} + 2}.$$

При переходе в признаковое подпространство, заданное вектором $\mathbf{s} = \{s_j\}_{j=\overline{1,N}}$, где $s_j \in \{0,1\}$, расстояние между объектами $\mathbf{x_i}(\mathbf{s}) = \{s_j x_{ij}\}_{j=\overline{1,N}}$, $i = \overline{1,M}$, и $\mathbf{x_l}(\mathbf{s}) = \{s_j x_{lj}\}_{j=\overline{1,N}}$, $l = \overline{1,M}$, в пространстве $X(\mathbf{s})$ запишется как $\tau_{X(\mathbf{s})}(\mathbf{x_i}(\mathbf{s}), \mathbf{x_l}(\mathbf{s})) = \sqrt{\sum_{j=1}^{N} s_j \mid x_{ij} - x_{lj}\mid^2}$.

Таким образом, для решения поставленной задачи требуется найти множества $\mathbf{s}^* = \{s_j^*\}_{j=\overline{1,N}}$, $s_j^* \in \{0,1\}$, $\sum_{j=1}^{N} s_j^* = n^*$ и $\mathbf{t}^* = \{t_i^*\}_{i=\overline{1,M}}$, $t_i^* \in \{-1,0,1\}$, которые обеспечат максимум следующей функции:

$$H(\{(\mathbf{x_i}, y_i)\}_{i=\overline{1,M}}, \mathbf{s}, \mathbf{t}) =$$

$$= \frac{\sum_{i=1}^{M}(t_i^2 + t_i y_i)}{\sum_{i=1}^{M} t_i^2} \times \sum_{i=1}^{M} \frac{\min\limits_{\substack{l=\overline{1,M}\\ l:\, l\neq i,\, t_l\neq 0}}\left\{\sqrt{\sum_{j=1}^{N} s_j \mid x_{ij} - x_{lj}\mid^2} + t_l\right\} - \min\limits_{\substack{l=\overline{1,M}\\ l:\, l\neq i,\, t_l\neq 0}}\left\{\sqrt{\sum_{j=1}^{N} s_j \mid x_{ij} - x_{lj}\mid^2} - t_l\right\}}{\min\limits_{\substack{l=\overline{1,M}\\ l:\, l\neq i,\, t_l\neq 0}}\left\{\sqrt{\sum_{j=1}^{N} s_j \mid x_{ij} - x_{lj}\mid^2} + t_l\right\} + \min\limits_{\substack{l=\overline{1,M}\\ l:\, l\neq i,\, t_l\neq 0}}\left\{\sqrt{\sum_{j=1}^{N} s_j \mid x_{ij} - x_{lj}\mid^2} - t_l\right\} + 2} t_i,$$

где условие $t_l \neq 0$ означает, что в поиске минимального расстояния от i-го объекта до ближайшего объекта своего образа удаленные объекты не участвуют.

2.4. Приближенный алгоритм FRiS-LCFS (FRiS Local Censoring with Feature Selection) коррекции-фильтрации ошибочно классифицированных объектов в пространствах большой размерности

Для решения поставленной задачи предлагается приближенный жадный алгоритм, который на каждом шаге выбирает максимально информативную признаковую систему. Критерием информативности полученной системы выступает согласно (3) величина разделимости классов в данной системе после цензурирования. Наращивание признаковой системы осуществляется по принципам, используемым в алгоритме AddDel [23], который сначала пошагово добавляет в текущую признаковую подсистему признаки, максимально увеличивающие ее общую информативность, а через определенное количество шагов осуществляет откаты, пошагово удаляя самые неинформативные признаки, чье исключение из текущей системы минимально ухудшает ее качество. Для вычисления информативности рассматриваемых систем признаков в каждой из них запускается описанный в [13] алгоритм цензурирования данных в режиме коррекции (FRiS-LC(C)) или фильтрации (FRiS-LC(F)), осуществляющий выбор кандидатов на исключение из выборки или для исправления целевого признака, основываясь на локальном сходстве объектов выборки с k ближайшими соседями.

Модификация алгоритма AdDel заключается в том, что на текущий шаг алгоритма подается выборка, полученная в результате процедуры коррекции-фильтрации исходной выборки в признаковом пространстве, сформированном на предыдущем шаге алгоритма.

Опишем разработанный алгоритм более подробно на примере двух образов A и B, заданных в признаковом пространстве X размерности N. На вход алгоритма подается

обучающая выборка $O = A \bigcup B$. Параметры алгоритма: n^+ и n^-, $n^- < n^+$, – количество шагов на прямом и обратном ходе алгоритма; n^*, $n^* \le N$, – заданное количество признаков в информативной подсистеме. Очевидно, что минимальное число n признаков в текущей системе, позволяющее получить все возможные наборы заданной размерности n^*, будет при достижении значения $n = \min\{n^* + n^+, N\}$ на входе на шаг добавления очередного признака.

Переменные алгоритма: n – размерность текущей признаковой системы X', $O' = A' \bigcup B'$ – отцензурированная выборка с разбивкой на классы в текущей системе признаков, G' – качество текущей системы: $G' = G_{AB}(X', A', B')$. Рекорды задачи: O^* – цензурированная выборка, X^* – признаковая подсистема, G^* – качество.

Алгоритм FRiS-LCFS:

Шаг 0. Положим n:=0, $O' := O$, $X' := \varnothing$, $G' = -1$, $X^* := X'$, $O^* := O', G^* := G'$.

Шаг 1. (Add) i := 1.

Шаг 1.1. Если $n = N$ или $n = n^* + n^+$, то переход на шаг 3. В набор признаков X' добавляется наиболее информативный признак x^* из множества $X \setminus X'$. Информативность признака определяется оценкой разделимости классов $G'(X' \bigcup \{x\}, O')$: $x^* = \arg\max_{x \in X \setminus X'} G(X' \bigcup \{x\}, O')$. Положим $n := n + 1$, $X' := X' \bigcup \{x^*\}$. На полученном множестве признаков исходная выборка O корректируется процедурой FRiS-LC и получается выборка O' с качеством G'. Если $n = n^*$ и $G' > G^*$, то положим $X^* := X'$, $O^* := O', G^* := G'$. Положим $i := i + 1$. Если $i < n^+$, то переход на шаг 1.1.

Шаг 2. (Del) i := 1.

Шаг 2.1. Из набора признаков X' удаляется наименее информативный признак $x^* = \arg\max_{x \in X'} G(X' \setminus \{x\}, O')$. Положим $n := n - 1$, $X' := X' \setminus \{x^*\}$. На полученном множестве признаков исходная выборка O корректируется процедурой FRiS-LC и получается выборка O' с качеством G'. Если $n = n^*$ и $G' > G^*$, то положим $X^* := X'$, $O^* := O', G^* := G'$. Положим $i := i + 1$. Если $i \le n^-$, то переход на шаг 2.1, иначе переход на шаг 1.

Шаг 3. Алгоритм останавливается, а на выход подается O^* – цензурированная выборка, X^* – лучшая признаковая подсистема размерности n^*.

В зависимости от того, какая модификация алгоритма FRiS-LC используется, выборка после цензурирования будет либо с исправлениями целевого признака, либо без.

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Тестирование предложенного алгоритма FRiS-LCFS коррекции-фильтрации неверно классифицированных объектов в пространствах большой размерности проводилось как на модельных, так и на реальных медико-биологических данных. Для проверки качества разработанного алгоритма вычислялись надежность распознавания целевого признака (*P*), чувствительность (*Sns*) и специфичность (*Spc*) по отношению к ошибкам в целевом признаке. Параметры алгоритма: $n^+ = 2$, $n^- = 1$; строились наиболее информативные подсистемы признаков размерности $n^* = 1, \ldots, 10$; число ближайших соседей k = 5.

Ниже на рисунках 1–3 представлены результаты тестирования разработанного метода на задаче диагностики рака простаты [15]. Выборка состояла из 102 объектов, описанных в пространстве 12625 признаков (генов), из них 50 объектов – здоровые

клетки (Normal), 52 объекта представляют клетки опухоли (Tumour). В каждом эксперименте выборка 10 раз случайным образом делилась на обучающую и контрольную. Использовались обучающие выборки объема $M = 80$ объектов (по $M/2$ объектов первого и второго образов). Для моделирования ошибок диагностики целевой признак (диагноз) случайным образом менялся для части объектов обучающей выборки. Уровень ошибки диагностики (*De*) в целевом признаке составлял 0 %, 10 %, 20 %.

В первой серии экспериментов сравнивались между собой различные сочетания описанных в более ранних работах алгоритмов AdDel [23], FRiS-LC(C) [13], FRiS-LC(F) [14] и предложенные для работы в пространствах большой размерности алгоритмы цензурирования в режиме коррекции FRiS-LCFS(C) и в режиме фильтрации FRiS-LCFS(F). Для этого вычислялась надежность распознавания тестовой выборки по правилу k – ближайших соседей для следующих алгоритмов:

A1 – распознавание на неочищенных данных в пространстве информативных признаков, выбранных алгоритмом AdDel,

A2 – распознавание на данных, очищенных алгоритмом FRiS-LC(F) в пространстве информативных признаков, выбранных алгоритмом AdDel,

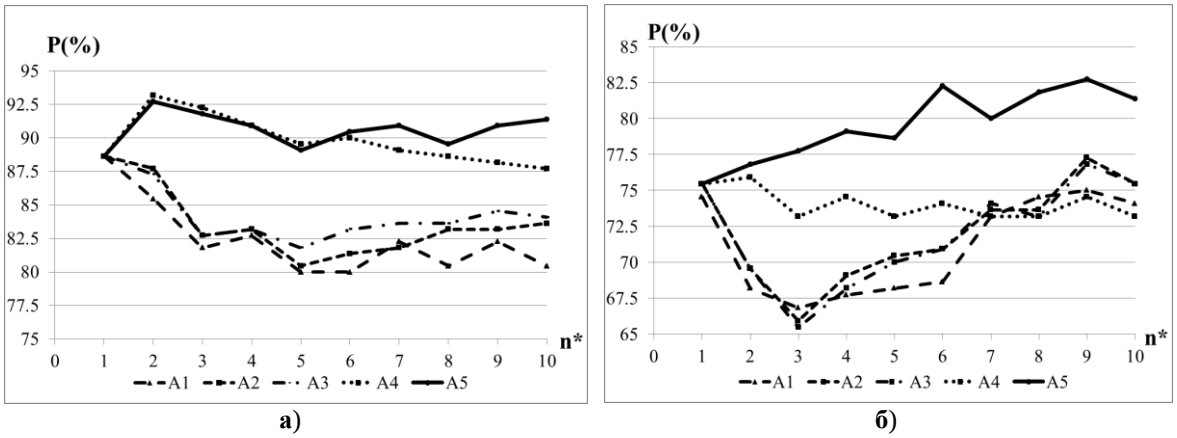
A3 – распознавание на данных, очищенных алгоритмом FRiS-LC(C) в пространстве информативных признаков, выбранных алгоритмом AdDel,

A4 – FRiS-LCFS(C),

A5 – FRiS- LCFS(F).

На рисунке 1 представлены результаты проведенного сравнения алгоритмов при уровне ошибки в целевом признаке *De*=10 %, 20 %. По оси абсцисс откладывается количество признаков в выбираемой признаковой подсистеме, по оси ординат – качество распознавания тестовой выборки в режиме перекрестной проверки (*P*).

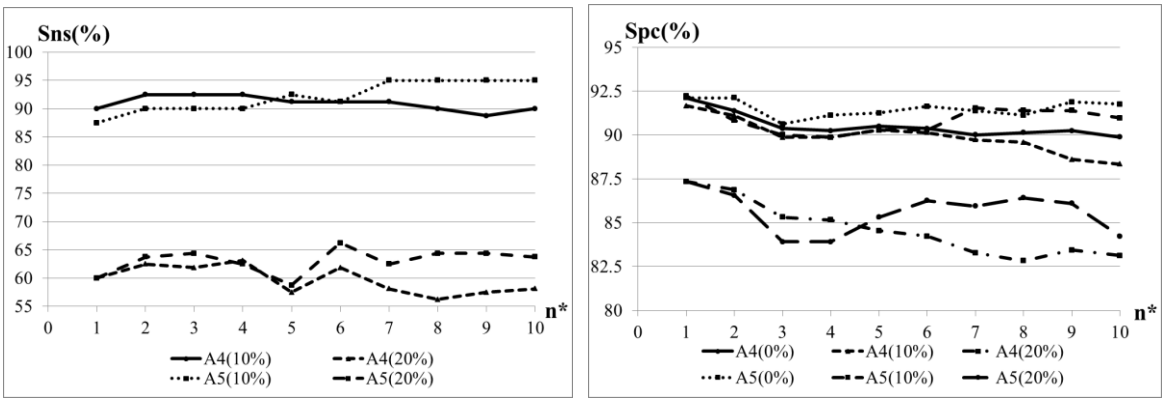
Анализ результатов проведенных экспериментов показывает, что предложенный алгоритм дает сопоставимые результаты в режиме коррекции и фильтрации для небольшого уровня ошибки в целевом признаке (10 %), увеличение же этого уровня до 20 % приводит к тому, что попытка корректировать объекты-выбросы сказывается негативно на общем качестве распознавания.



а) **б)**

Рис. 1. Зависимость надежности распознавания P от размера выбранного признакового пространства при уровне ошибки диагностики: **а**) *De* = 10 %, **б**) *De* = 20 %.

Для более детального сравнения двух режимов (коррекции и фильтрации) предложенного алгоритма очистки данных в пространствах большой размерности вычислялась чувствительность и специфичность к ошибкам в целевом признаке при уровне ошибки *De* = 0 %, 10 %, 20 %. Результаты данных экспериментов приведены на рисунке 2.



а) **б)**

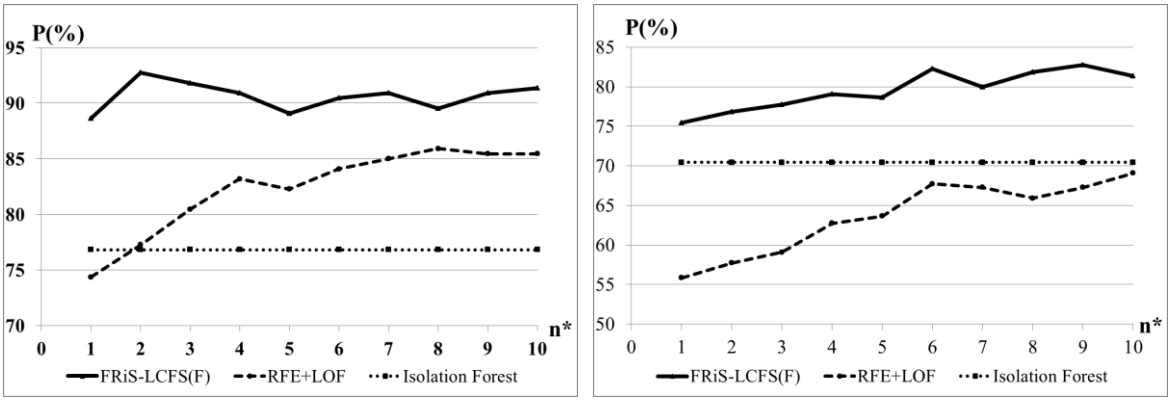
Рис. 2. Зависимость чувствительности **а**) и специфичности **б**) от уровня ошибки диагностики *De*, указанной в скобках, и размера выбираемого признакового пространства.

В таблице 1 приведены данные при использовании алгоритмов FRiS-LCFS(C) (обозначенный как A4) и FRiS-LCFS(F) (обозначенный как A5), отражающие качество выборки после применения, соответственно, процедуры коррекции-фильтрации и процедуры полной фильтрации – ожидаемый объем выборки $\tilde{M}$ и ожидаемый уровень ошибок в целевом признаке $\tilde{De}$. Ошибка диагностики $\tilde{De}$ вычисляется как сумма оставшихся исходных ошибок в целевом признаке и ошибок, привнесенных неправильной коррекцией целевых признаков. Уровень ошибки в целевом признаке исходной выборки составлял 10 % и 20 %. Как коррекция-фильтрация, так и чистая фильтрация, существенно снижают уровень ошибки в целевом признаке. Применение процедуры коррекции-фильтрации сохраняет большую часть анализируемой выборки при большем уровне ожидаемой ошибки диагностирования, чем применение процедуры фильтрации. Таким образом, применение процедуры фильтрации более качественно «очищает» исходную обучающую выборку, и если сохранение объема обучающей выборки не является приоритетным, то в пространствах большой размерности рекомендуется использовать именно ее.

Таблица 1. Ожидаемый объем выборки и ожидаемая ошибка диагностики исходных данных

	A4(10 %)		A5(10 %)		A4(20 %)		A5(20 %)	
n^*	$\tilde{M}$	$\tilde{De}$	$\tilde{M}$	$\tilde{De}$	$\tilde{M}$	$\tilde{De}$	$\tilde{M}$	$\tilde{De}$
1	77.4	6.72	67.4	1.48	75.8	17.28	62.3	10.27
2	75.4	5.57	66.2	1.21	72.4	15.06	61.2	9.48
3	73.8	5.56	65.6	1.22	72.6	15.29	59.4	9.60
4	73.7	5.29	65.5	1.22	72.4	15.61	59.7	10.05
5	74.7	6.02	65.6	0.91	73.0	16.44	61.2	10.78
6	75.3	6.64	65.7	1.07	73.0	15.89	60.6	8.91
7	74.0	5.68	66.3	0.60	72.0	17.08	61.0	9.84
8	74.6	6.43	66.2	0.60	72.5	17.79	61.0	9.34
9	73.9	6.36	66.2	0.60	72.3	17.01	60.8	9.38
10	73.5	6.53	65.9	0.61	72.1	16.64	59.7	9.72

В последней серии экспериментов предложенный алгоритм FRiS-LCFS(F) сравнивался с двумя из существующих аналогов, используемых для работы в пространствах большой размерности. Один из них – гибридный, состоящий из выбора признаков алгоритмом Recursive Feature Elimination (RFE) [7] с последующей очисткой данных от шумовых объектов алгоритмом Local Outlier Factor (LOF) [8], а второй – алгоритм для очистки данных в пространствах большой размерности Isolation Forest [9]. На рисунке 3 представлены результаты сравнения этих алгоритмов с предложенным алгоритмом FRiS-LCFS в режиме фильтрации объектов для уровня ошибки в целевом признаке 10 % и 20 %.



а) **б)**

Рис. 3. Зависимость надежности распознавания P от размера выбранного признакового пространства при уровне ошибки в целевом признаке: **а**) De = 10 %, **б**) De = 20 %.

Представленные результаты демонстрируют, что алгоритм FRiS-LCFS(F) значительно лучше рассмотренных аналогов справляется с очисткой данных от ошибок в целевом признаке в пространстве большой размерности.

ЗАКЛЮЧЕНИЕ

Наличие диагностических ошибок в данных, особенно для случая, когда эти данные описываются в пространствах большой размерности, значительно усложняет процесс отыскания скрытых в них закономерностей. Разработанный на основе FRiS-методологии подход к решению задачи очистки таких данных от ошибок в пространствах большой размерности продемонстрировал свою высокую эффективность в сравнении со стандартными методами, используемыми для этих целей. Использование предложенного в статье алгоритма FRiS-LCFS как в режиме коррекции, так и в режиме фильтрации, существенно повышает уровень надежности классификации по очищенным данным, при этом происходит выбор наиболее информативного и релевантного решаемой задаче признакового пространства малой размерности. Таким образом, параллельно осуществляется дополнительная очистка данных и от шумовых признаков.

Проведенные исследования показали, что результаты коррекции-фильтрации и чистой фильтрации сопоставимы при малых значениях уровня искажения целевого признаки. Учитывая, что корректирующие методы в отличие от фильтрующих позволяют сохранять значительно большую часть выборки, можно сделать вывод о потенциальной эффективности применения данного подхода в ситуациях, требующих минимального сокращения объема выборки в процессе предобработки. Однако с ростом уровня искажения целевого признака применение процедуры фильтрации начинает более качественно «очищать» исходную обучающую выборку, и если сохранение объема обучающей выборки не является приоритетным, а уровень

возможных диагностических ошибок велик, то в пространствах большой размерности рекомендуется использовать именно ее.

Предложенный алгоритм FRiS-LCFS и легший в его основу метод могут быть использованы при проектировании систем интеллектуального анализа данных для повышения как качества выборок, так и качества обучения.

Работа выполнена при поддержке программы ФНИ РАН, проект № 0314-2019-0015.

СПИСОК ЛИТЕРАТУРЫ

1. de Waal T., Pannekoek J., Scholtus S. *Handbook of Statistical Data Editing and Imputation.* Hoboken, New Jersey: John Wiley and Sons, Inc., 2011. 456 p. doi: 10.1002/9780470904848.ch1.
2. Barnett V., Lewis T. *Outliers in Statistical Data.* Chichester: John Wiley and Sons, 1994. 584 p.
3. 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.
4. Luca Greco. *Robust Methods for Data Reduction Alessio Farcomeni*. Chapman and Hall/CRC, 2015. 297 p.
5. 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.
6. Aggarwal C.C., Yu P.S. Outlier detection for high dimensional data. In: *Proc. ACM SIGMOD Int. Conf. on Management of Data.* California, USA. 2001. doi: 10.1145/375663.375668.
7. Guyon I., Weston J., Barnhill S., Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning.* 2002. V. 46. № 1. P. 389–422.
8. Breunig M.M., Kriegel H.-P., Ng R.T., Sander J.R. LOF: Identifying Density-based Local Outliers. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.* 2000. P. 93–104.
9. Liu F.T., Ting K.M., Zhou Z.-H. Isolation forest. In: *Proceedings of ICDM'08. Eighth IEEE International Conference on Data Mining*. 2008. P. 413–422.
10. Kriegel H.P., Schubert M., Zimek A. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.* 2008. P. 444–452. doi: 10.1145/1401890.1401946.
11. Yang Y., Wu X., Zhu X. 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.* Springer, 2004. doi: 10.1007/978-3-540-30116-5_43.
12. Brodley C.E., Friedl M.A. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research.* 1999. V. 11. P. 131–167.
13. Борисова И.А., Кутненко О.А. Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства. *Математическая биология и биоинформатика*. 2018. Т. 13. № 1. С. 38–49.
14. Борисова И.А., Кутненко О.А. Цензурирование ошибочно классифицированных объектов выборки. *Машинное обучение и анализ данных.* 2015. Т. 1. № 11. С. 1632–1641.
15. *Prostate Cancer Dataset.* URL: http://www.bioinf.ucd.ie/people/ian/Singh.txt (accessed January 2019).

16. 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.
17. Загоруйко Н.Г. *Когнитивный анализ данных.* Новосибирск: Академическое изд-во ГЕО, 2013. 186 с.
18. Аркадьев А.Г., Браверман Э.М. *Обучение машины классификации объектов.* М.: Наука, 1971. 112 с.
19. Воронцов К.В., Колосков А.О. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации. *Искусственный интеллект*. 2006. № 2. С. 30–33.
20. Шлезингер М.И. О самопроизвольном разделении образов. *Читающие автоматы и распознавание образов: сб. науч. трудов*. Киев: Наукова думка, 1965. С. 46–61.
21. Загоруйко Н.Г. *Прикладные методы анализа данных и знаний.* Новосибирск: Изд. ИМ СО РАН, 1999. 270 с.
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. Zagoruiko N.G., Kutnenko O.A. Recognition methods based on the AdDel algorithm. *Pattern Recognition and Image Analysis*. 2004. V. 14. № 2. P. 198–204.

Рукопись поступила в редакцию 04.07.2019, переработанный вариант поступил 04.10.2019.
Дата опубликования 07.10.2019.