Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛВРА

Формирование однородных обучающих выборок в задачах классификации

Выполнила:
студентка 4 курса 174 группы
Ефимова Ирина Валерьевна

Научный руководитель:
с.н.с ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва, 2015

Содержание

Аннотация

Рассматривается задача пополнения обучающей выборки. Имеются две размеченные выборки объектов двух классов. Первая выборка эталонная, вторая содержит неизвестную долю выбросов — объектов с неверной классификацией. Исследование состоит в построении алгоритма, позволяющего очищать вторую выборку от выбросов, для получения одной однородной выборки и повышения качества классификации. Предлагаются три метода: метод сближения AUC двух выборок, метод сближения ROC-кривых и метод выделения объектов, влияющих на переобучение. Исследуется обобщающая способность построенных алгоритмов.

Ключевые слова: *отсев выбросов, ROC-кривые, пополнение выборки, Precision-Recall кривые.*

Введение

Актуальность темы. В задачах статистического обучения в области медицинской диагностики может возникать проблема неоднородности обучающих выборок. Например, в тех случаях, когда диагнозы в основном ставятся терапевтами, и лишь для небольшой доли обследуемых диагнозы подтверждаются лабораторными и инструментальными исследованиями. В этой ситуации и возникает задача пополнения обучающей выборки.

Цель работы. Целью работы является нахождение вычислительно эффективного способа очистки одной из двух размеченных выборок от выбросов. Ожидается, что качество классификации при обучении на объединенной выборке, полученной в результате пополнения первой выборки очищенной второй, будет выше по сравнению с обучением только по первой.

Научная новизна. Предложены метод сближения ROC-кривых и метод выделения объектов, влияющих на переобучение, для очистки второй выборки от выбросов в задаче пополнения обучающей выборки.

Практическая ценность. Разработан программный модуль, который

- проверяет однородность выборок;
- по первой выборке очищает вторую выборку от выбросов;
- пополняет первую выборку очищенной второй выборкой;
- визуализирует результаты.

Положения, выносимые на защиту: Для решения задачи пополнения обучающей выборки предложены:

- метод сближения ROC-кривых;
- метод выделения объектов, влияющих на переобучение.

Апробация. Результаты работы докладывались:

- на 57-ой международной научной конференции МФТИ (24–29 ноября 2014 г., Москва-Долгопрудный-Жуковский);
- на Традиционной Школе «Управление, информация и оптимизация» (14–20 июня 2015 г., г. Солнечногорск Московской области).

Обзор методов отсева выбросов. Проблема обнаружения и отсева аномальных объектов (выбросов) возникает во многих задачах анализа данных, включая задачи классификации в области медицинской диагностики. Статистические методы для решения задачи отсева выбросов были предложены еще в 19 веке [1]. В связи с развитием технологий для сбора данных и их организации (систематизации) данная задача начала активно изучаться учеными в области компьютерных наук. За последние двадцать лет было разработано множество новых методов. Среди последних публикаций наиболее полный обзор по всем существующим методам сделан в [2, 4, 5].

Данные могут иметь метки, которые указывают является объект выбросом или нет. Следует отметить, что определение таких меток требует больших усилий и затрат. И в зависимости от того, метки каких объектов доступны, выделяют три основных подхода к решению задачи отсева выбросов:

1. *Обучение без учителя.* В данном подходе не требуется обучающая выборка, в связи с чем эта техника имеет широкое применение.
2. *Обучение с учителем.* Предполагается, что обучающая выборка полностью размечена, то есть для каждого объекта указано, является он выбросом или нет. Методы отсева выбросов, обучающиеся с учителем, являются особым случаем задач классификаций. Особенность заключается в том, что метки классов сильно не сбалансированы. Обычно процесс получения (сбор) представителей хороших объектов значительно проще по сравнению с получением представителей объектов-выбросов. В связи с чем и возникает дисбаланс классов в обучающей выборке (в машинном обучении эта проблема известна как rare class detection). Решение такой задачи с помощью стандартных классификаторов может привести к переобучению.

Существуют два подхода к решению проблемы дисбаланса классов:

- *Cost Sensitive Learning:* Функционал качества алгоритма классификации модифицируется таким образом, чтобы оценивать вклад ошибок для разных классов по-разному: штраф за неправильную классификацию выбросов больше, чем за неправильную классификацию хороших объектов [14].Существуют MetaCost методы [6] и Weighting Methods [7]. MetaCost методы основаны на смене меток классов: хорошие объекты, которые имеют достаточную вероятность (reasonable probability) быть классифицированными как выбросы, объявляются выбросами. Это позволяет добиться того, что доля ложно отрицательных классификаций будет выше доли ложно положительных (это важно с практической точки зрения). Под отрицательными классификациями понимаются объекты, которые классифицировались как выбросы, а под положительными — как хорошие объекты. MetaCost может быть применен к любому классификатору.При использовании Weighting Methods модификация алгоритма классификации заключается в неявном рассмотрении каждого объекта обучающей выборки с весом, равным штрафу ошибочной классификации. Как правило, такие модификации зависят от специфики алгоритма классификации. Существует множество алгоритмов классификаций, модифицированных согласно Weighting Methods, среди них байесовский классификатор [7], метод ближайших соседей [15], решающее дерево [16, 17], метод опорных векторов (SVM классификатор) [18, 19].
- *Адаптивный ресэмплинг (Adaptive Re-sampling)*: Вместо исходной выборки на вход алгоритму классификации подается сэмплированная выборка. Сэмплирование производится для того, чтобы увеличить относительную долю объектов-выбросов [12, 13]. Вероятности сэмплирования обычно выбираются пропорционально штрафам за неправильную классификацию.

Есть ряд методов, которые основаны на введении искусственных выбросов в случае отсутствия данных об объектах-выбросах [8–10].

3. *Частичное обучение.* В данном подходе предполагается, что имеется выбор-

ка из хороших объектов и большая неразмеченная выборка. Неразмеченная выборка содержит как хорошие объекты, так и объекты-выбросы, при этом в неизвестной пропорции. Существуют алгоритмы, которые используют двухшаговую стратегию: S-EM [20], PEBL [21], Roc-SVM [22].

Шаг 1: Определение в неразмеченной выборке надежных объектов-выбросов. На этом шаге S-EM использует технику Spy (Spy technique), PEBL — 1-DNF, Roc-SVM — алгоритм Rochio [23].

Шаг 2: Построение множества классификаторов, итеративно применяя алгоритмы классификации, и затем выбор наилучшего из них. На этом шаге S-EM использует EM (Expectation Maximization) алгоритм [24] с наивным байесовским классификатором, PEBL и Roc-SVM — SVM. Для S-EM и Roc-SVM существуют несколько способов выбора финального классификатора. PEBL в качестве финального классификатора использует последний, который может оказаться не лучшим.

Эти два шага могут рассматриваться, как итеративный способ увеличения числа объектов из неразмеченной выборки, классифицируемых как выбросы, приэтом безошибочно классифицируя первую выборку.

Существуют и другие методы решения данной задачи. В [25] предложен модифицированный наивный байесовский классификатор. Главный недостаток этого метода заключается в том, что он требует знания пользователя о вероятности класса хороших объектов. На практике же у пользователя не всегда есть возможность определить эту вероятность. В [35] предложен biased SVM. Результаты экспериментов показали, что данный метод работает лучше по сравнению со всеми существующими двухшаговыми техниками. В [36] получен интересный и фундаментальный результат. Если размеченная выборка состоит из хороших объектов, случайно выбранных из всего класса хороших объектов, то условные вероятности, которые получены моделью, обученной по размеченной и неразмеченной выборкам, отличаются в константу раз от условных вероятностей, которые получены моделью, обученной по полностью размеченной выборке,

состоящей из хороших объектов и объектов-выбросов. В [36] предложено два метода, как, используя данный результат, обучить классификатор по выборке из хороших объектов и неразмеченной выборке, и показано, что эти методы работают лучше biased SVM.

Также есть методы, которые не используют неразмеченную выборку и обучаются только на выборке из хороших объектов. Существуют два подхода к решению задач такого типа. Первый подход состоит в оценивании плотности вероятности, но, как известно, это сложная задача для многомерных данных. Второй подход состоит в использовании одноклассового SVM [26,27]. Недостатком данного подхода является его чувствительность к выбору параметров, для настройки которых пока нет эффективных методов.

Есть ряд техник, которые предполагают, что для обучения имеется только выборка из объектов-выбросов [32–34]. Такие методы используются нечасто, так как на практике очень сложно собрать всех представителей объектов-выбросов.

Возможны ситуации, когда имеются небольшая размеченная выборка и большая неразмеченная. Для решения такой задачи в [29] предлагают использовать наивный байесовский классификатор и EM алгоритм, в [30] — трансдуктивный SVM, в [31] — метод co-training.

1 Постановка задачи

1.1 Задача отсева выбросов

Дана выборка из хороших объектов P и неразмеченная выборка U, причем $U = Q \cup N$. Q образуют хорошие объекты, а N — выбросы. Мощности Q и N неизвестны.

Требуется построить алгоритм по выборкам P и U, способный классифицировать объекты на хорошие и на выбросы.

1.2 Задача пополнения выборки

Даны две выборки объектов $P = \{x_{pi}, y_{pi}\}_{i=1}^{l}$, $U = \{x_{ui}, y_{ui}\}_{i=1}^{k}$, где $y_{ti} \in \{0, 1\}$ — класс объекта x_{vi}, $v = \{p, u\}$.

Предполагается, что выполняются следующие положения:

- выборка P — эталонная: для объекта x_{pi} метка класса y_{pi} поставлена верно, $i = 1, ..., l$;
- выборка U содержит некоторую долю выбросов $N = \{x_{ni}, y_{ni}\}_{i=1}^{m} \subset U$ с инвертированной меткой y_{ni}.

Пусть $Q = U \setminus N$.

Требуется построить вычислительно эффективный алгоритм очистки выборки U от выбросов:

$$g : (P, U) \longrightarrow Q. \tag{1}$$

Обозначим через a алгоритм, осуществляющий классификацию объектов выборок P и U.

Построенный алгоритм (1) должен удовлетворять следующему критерию: качество классификации алгоритма a при обучении на объединенной выборке $P \cup Q$ выше по сравнению с обучением только по выборке P.

1.3 Задача проверки однородности двух выборок

Даны два набора объектов $X_1 = \{x_i\}_{i=1}^{l_1}$ и $X_2 = \{x_i\}_{i=1}^{l_2}$, взятых из неизвестных распределений $F_1(x)$ и $F_2(x)$ соответственно.

Необходимо проверить гипотезу о равенстве распределений $F_1(x) = F_2(x)$ при всех x, против альтернативы $F_1(x) \neq F_2(x)$ при некотором x.

1.4 Задача оценивания достаточности объема обучения

Дана выборка $D = \{x_i, y_i\}_{i=1}^{l}$, где y_i — класс объекта x_i, $i = 1, ..., l$. Пусть $D_{train} \subset D$ — обучающая выборка, $D_{test} \subset D$ — контрольная выборка, $D_{train} \cap D_{test} = \varnothing$.

Необходимо определить такой минимальный объем выборки D_{train}, что при его увеличении качество классификации алгоритма на D_{test}, обученного на D_{train}, не

улучшается.

2 Методы отсева выбросов и пополнения обучающих выборок

2.1 ROC-кривая

Пусть X — множество описаний объектов, $Y = \{0, 1\}$ — множество классов объектов, $a\colon X \to Y$ — классификатор, $a(x) = [f(x, w) \geq \theta]$. $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$ — выборка.

ROC-кривая является ступенчатой функцией, она характеризует зависимость доли верных положительных классификаций (True Positive Rate)

$$TPR = \frac{\sum\limits_{i=1}^{l} a(x_i)y_i}{\sum\limits_{i=1}^{l} y_i}$$

от доли ложных положительных классификаций (False Positive Rate)

$$FPR = \frac{\sum\limits_{i=1}^{l} a(x_i)(1 - y_i)}{\sum\limits_{i=1}^{l} (1 - y_i)}$$

при варьировании порога классификатора θ.

Качество классификатора a определяется площадью под ROC-кривой (AUC).

2.2 Алгоритм

Для решении задачи пополнения обучающей выборки в обычную процедуру построения классификатора добавляются шаги 2 и 3:

1. настроить классификатор по обучающей подвыборке P_{train};
2. отфильтровать выборку U определенным методом;
3. перенастроить классификатор по пополненной выборке $P_{train} \cup Q$;
4. классифицировать контрольную подвыборку P_{test}.

Пусть $P_{train} \subset P$ — обучающая выборка, $P_{test} \subset P$ — контрольная выборка, $P_{train} \cap P_{test} = \varnothing$, $P_{train} \cup P_{test} = P$.

2.3 Метод сближения AUC двух выборок

Введём обозначения для площадей под ROC-кривыми, вычисленными по следующим выборкам: AUC_1 — по выборке P_{train}, AUC_2 — по выборке U, $\mathrm{AUC}_{2,t}$ — по выборке U_t, полученной из U на t-ом шаге методом сближения AUC двух выборок.

Предполагается, что выборки P_{train} и U являются однородными, если площади под их ROC-кривыми (AUC) близки.

На каждом каждом шаге алгоритма исключается объект из выборки U, минимизирующий разность AUC:

$$(x_t, y_t) = \arg \min_{(x_d, y_d) \in U_{t-1}} |\mathrm{AUC}_1 - \mathrm{AUC}_{2,d}|.$$

В ходе эксперимента выяснилось, что данный метод не позволяет добиться однородности двух выборок.

2.4 Метод сближения ROC-кривых

Введём обозначения для ROC-кривых, вычисленных по следующим выборкам: ROC_1 — по выборке P_{train}, ROC_2 — по выборке U, $\mathrm{ROC}_{2,t}$ — по выборке U_t, полученной из U на t-ом шаге методом сближения ROC-кривых.

Для выравнивания ROC_1 и ROC_2 кривых предлагается на t-ом шаге алгоритма исключать объект из выборки U_{t-1}, минимизирующий площадь между ROC_1 и $\mathrm{ROC}_{2,t-1}$ кривыми:

$$(x_t, y_t) = \arg \min_{(x_d, y_d) \in U_{t-1}} \Delta(\mathrm{ROC}_1, \mathrm{ROC}_{2,d}). \quad (2)$$

Поиск на t-ом шаге алгоритма пары (x_t, y_t) (2), используя полный перебор по $(x_d, y_d) \in U_{t-1}$, занимает $O(k^2)$, $k = |U|$. Для вычисления (2) предлагается эффективный алгоритм, который вычисляет пару (x_t, y_t) за $O(k)$ (описан ниже в виде псевдокода). Он основан на анализе геометрической формы ROC-кривых как кусочно-постоянных монотонных функций, образуемых последовательностью прямоугольников. На t-ом шаге алгоритма осуществляется проход по $\mathrm{ROC}_{2,t-1}$-кривой слева напра-

во и каждый раз удаляется следующий объект и возвращается в выборку предыдущий. Это позволяет вычислить приращения площади от удаления каждого объекта за линейное время.

2.5 Метод выделения объектов, влияющих на переобучение

Введём обозначения для площадей под ROC-кривыми, вычисленными по следующим выборкам: AUC_1 — по выборке P_{train}, AUC_2 — по выборке U; $\mathrm{AUC}_{1,t}$ — по выборке $P_{train,t} = P \cup (x_t, y_t)$, $\mathrm{AUC}_{2,t}$ — по выборке $U_t = U \setminus (x_t, y_t),\ x_t \in U$.

Назовем переобученностью алгоритма a при заданных обучающей P_{train} и контрольной U выборках разность качества классификации на обучении и контроле:

$$d_0 = \mathrm{AUC}_1 - \mathrm{AUC}_2.$$

Пусть $d_t = \mathrm{AUC}_{1,t} - \mathrm{AUC}_{2,t}$.

Влияние каждого объекта выборки U на переобучение определяется формулой:

$$Impact_t = |d_0 - d_t|,\ t = 1, ..., |U|.$$

Выбросами объявляются объекты, для которых $Impact_t \geq \delta$, где δ — заданный порог.

3 Информационный анализ электрокардиосигналов

Известно, что импульсы, генерируемые сердцем, несут важную информацию о состоянии сердца и системы регуляции его функций. Современные электрокардиографы позволяют оценить состояние миокарда и функций сердца с большой точностью. Однако, опыт изучения вариабельности сердечного ритма на основе длительной регистрации электрокардиограммы свидетельствует о том, что электрокардиоимпульсы могут быть носителями информации также о состоянии системы регуляции основных функций организма в норме, при различных заболеваниях и в условиях воздействия на человека экстремальных факторов профессиональной деятельности и среды обитания [39,40]. Установлено, что импульсы, генерируемые сердцем, распространяются по всему организму и имеют свойства сигналов, которые и несут информацию о состоянии внутренних органов в норме и при различных патологиях. Эти результаты обосновывают рассмотрение сердца как информационного органа. Исследования

Алгоритм 1 Метод сближения ROC-кривых

Вход: P — первая выборка; U — вторая выборка;

$nOutliers$ — количество выбросов, которое надо удалить;

$f(x)$ — дискриминантная функция алгоритма классификации, обученного по выборке P;

Выход: N — множество объектов-выбросов;

отсортировать P и U по убыванию величин $f(x_i, \theta)$.
построить ROC_1-кривую по P;
R_1 = прямоугольники ROC_1-кривой (рис. 2 b);
$U_0 = U,\ N = \varnothing$;
для всех $t = 1, \ldots, nOutliers$
  для всех $m = 0, 1$ (объекты класса $y = m$)
    если $m = 1$ **то**
      удалить из U_{t-1} первый объект: $y = 1$;
    иначе
      удалить из U_{t-1} последн. объект: $y = 0$;
    построить $\mathrm{ROC}_{2,1}$-кривую по U_{t-1};
    S_1 = площадь между ROC_1 и $\mathrm{ROC}_{2,1}$ (рис. 2 a);
    R_2 = прямоугольники для $\mathrm{ROC}_{2,1}$ (рис. 2 b);
    вычислить $R_1 \cap R_2$;
    для всех $p : y_p = m,\ (x_p, y_p) \in U_{t-1}$
      S_p = площадь между ROC_1 и $\mathrm{ROC}_{2,p}$;
    $S_{r,m} = \min\limits_{\substack{(x_p,y_p)\in U_{t-1} \\ y_p = m}} S_p$;
  $x_r = \arg\min\limits_m S_{r,m}$;
  $U_t = U_{t-1} \backslash x_r,\ N = N \cup r$;

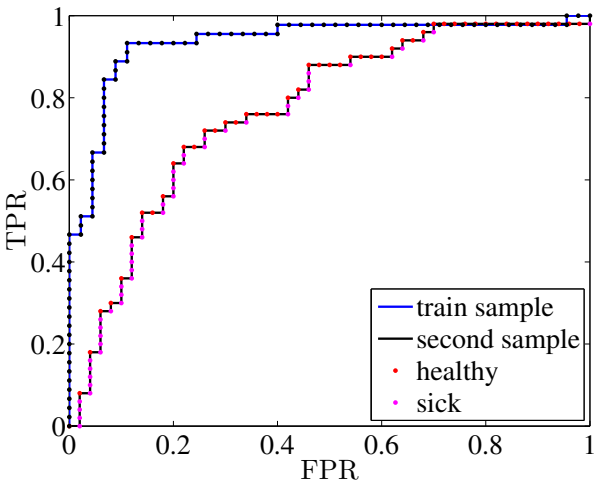
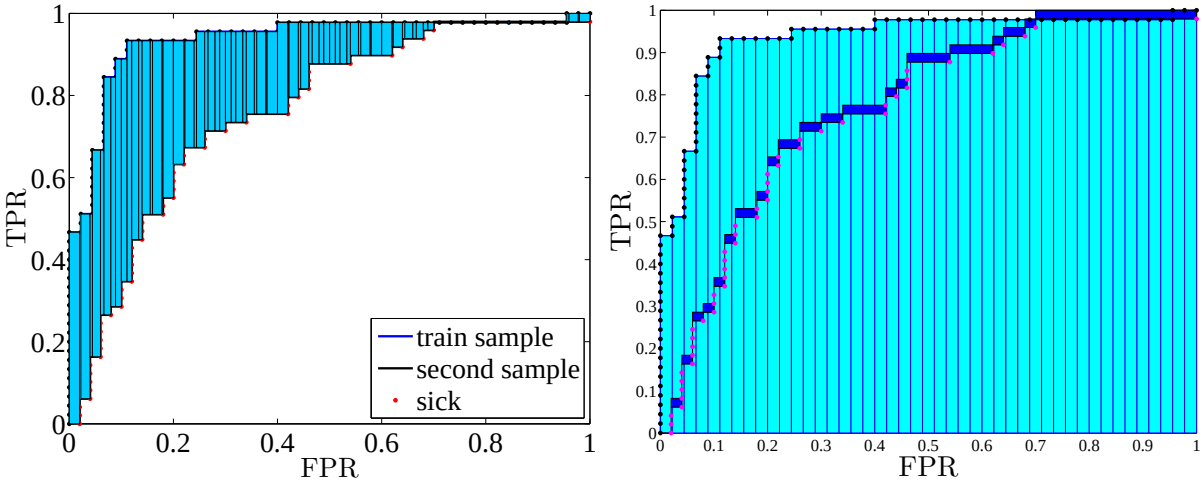


Рис. 1: Метод сближения ROC-кривых: ROC_1 и ROC_2 кривые до удаления выбросов из выборки U.



(a) Шаг 12

(b) Шаги 3, 13

Рис. 2: Метод сближения ROC-кривых: вычисление разницы площади между двумя ROC кривыми.

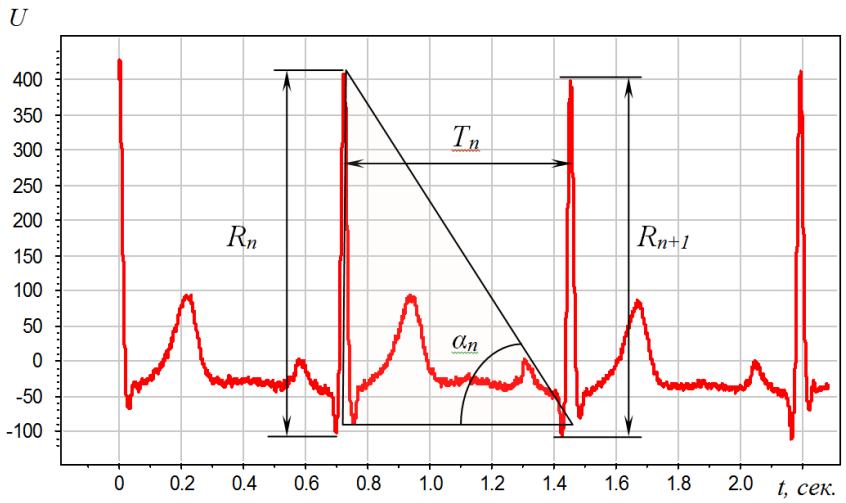


Рис. 3: Пример электрокардиограммы. Кардиоцикл с амплитудой R_n, интервалом T_n и углом α_n.

в этой области привели к созданию технологии информационного анализа электрокардиосигналов. На основе данной технологии была разработана диагностическая система «Скринфакс». В настоящее время «Скринфакс» позволяет диагностировать более 30 наиболее распространенных заболеваний внутренних органов. За 15 лет применения диагностической системы во врачебной практике накоплено более 20 тысяч записей электрокардиограмм.

3.1 Дискретизация ЭКГ-сигнала

Электрокардиограмма представляет собой квазипериодический сигнал, периоды которого называются кардиоциклами (рис. 3).

На первом этапе электрокардиосигнал преобразуется в последовательность амплитуд кардиоциклов R_n и их интервалов T_n. Также вводится арктангенс их отношения $\alpha_n = arctg\frac{R_n}{T_n}$ (рис. 3).

Предполагается, что диагностическую ценность несут знаки приращений величин $R_n,\ T_n,\ \alpha_n$. Возможны только 6 сочетаний изменений данных параметров, которые предлагается кодировать буквами шестисимвольного алфавита: {A, B, C, D, E, F}. В таблице «+» означает положительное приращение, «-» —отрицательное:

$R_{n+1} - R_n$	+	−	+	−	+	−
$T_{n+1} - T_n$	+	−	−	+	+	−
$\alpha_{n+1} - \alpha_n$	+	+	+	−	−	−
s_n	A	B	C	D	E	F

В результате дискретизации амплитудограмма и интервалограмма преобразуются в символьную последовательность $S = (s_n)_{n=1}^{N}$, состоящую из символов алфавита {A, B, C, D, E, F} и называемую кодограммой. Каждый символ кодирует тип взаимосвязи между двумя соседними кардиоциклами.

Важной особенностью данного подхода является то, что учитываются не только интервалограммы, как в современном анализе электрокардиосигналов, но и амплитудограммы. Затем на основе их совместного анализа автоматически выявляются паттерны заболеваний и строятся диагностические правила.

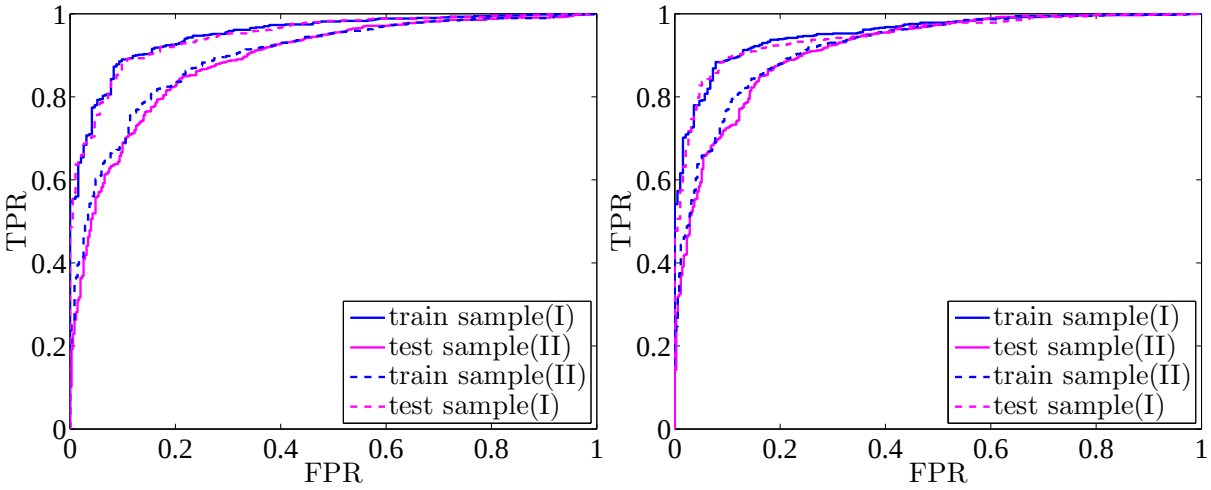
На последнем этапе происходит векторизация символьной последовательности S. Слово, образованное k последовательными буквами кодограммы, будем называть k-граммой. Частота k-граммы определяется как отношение числа её вхождений в кодограмму к длине кодограммы. Преобразование кодограммы в вектор частот k-грамм называется векторизацией. В данном исследовании используются 216-мерные векторы частот триграмм, $k = 3$.

Дискретизация и векторизация сохраняют значимую диагностическую информацию при сокращении объёма данных в несколько тысяч раз.

3.2 Постановка эксперимента

Эксперимент проводился на данных, полученных с помощью технологии информационного анализа электрокардиосигналов [37]. Каждому обследованию соответствует вектор из 216 числовых признаков и метка класса: 0 — здоров, 1 — болен. Данные по каждой болезни разбиты на две подвыборки. Первая подвыборка состоит из обследований с надёжно установленными диагнозами. Они используются для настройки параметров алгоритма классификации. Вторую подвыборку составляют случаи, когда диагнозы не были подтверждены лабораторными и инструментальными исследованиями, либо вызывали сомнения у врачей.

Цели эксперимента:



(a) Болезнь А (b) Болезнь Б

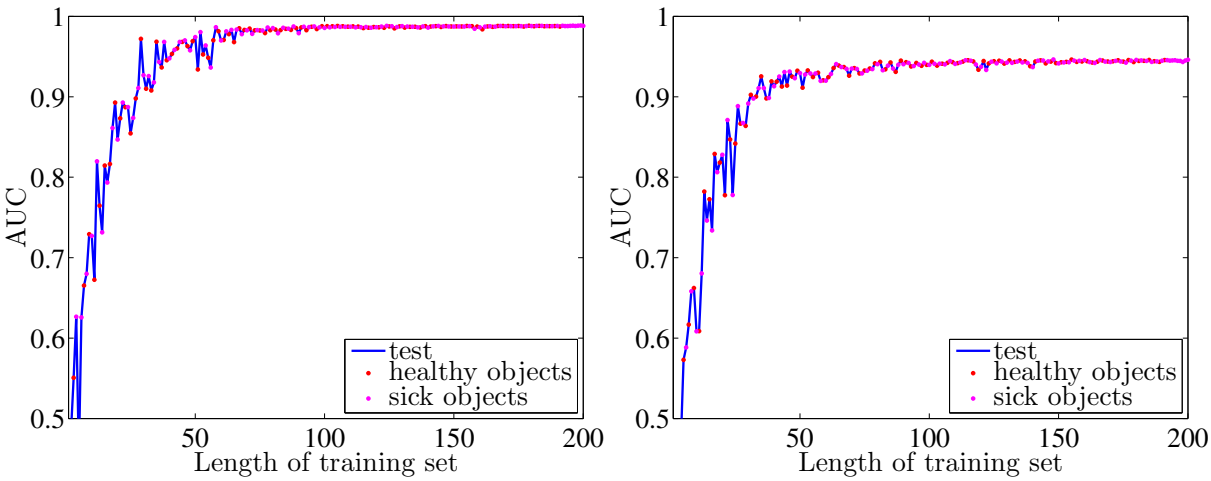
Рис. 4: Проверка гипотезы о более высоком уровне шума во второй выборке.

- оценить достаточную длину обучающей выборки;
- очистить вторую выборку от выбросов;
- повысить качество классификации на пополненной выборке;
- сравнить методы отсева выбросов и пополнения выборки.

Для классификации объектов использовался синдромный алгоритм — наивный байесовский классификатор с отбором признаков [38], так как известно, что на этих данных синдромный алгоритм дает хорошее качество классификации и имеет крайне низкий уровень переобучения.

Проверка гипотезы о более высоком уровне шума во второй выборке

Для проверки данной гипотезы проводились два эксперимента. В первом эксперименте классификатор настраивали по первой выборке и строили ROC-кривые по обеим выборкам. Выяснилось, ROC_2-кривые для всех болезней проходят существенно ниже ROC_1-кривых. Затем настройка классификатора производилась по второй выборке. Оказалось, что и в этом эксперименте ROC_2-кривые проходят ниже ROC_1-кривых (рис. 4). Эти результаты подтверждают гипотезу о более высоком шуме во второй выборке и говорят о робастности синдромного алгоритма.



(a) Болезнь А, $|P_{test}| = 235$

(b) Болезнь Б, $|P_{test}| = 470$

Рис. 5: Кривые обучения.

Оценка достаточной длины обучающей выборки

Для каждого заболевания были построены кривые обучения — зависимость качества классификации от длины обучающей выборки. Оказалось, что для всех болезней в первой выборке содержится достаточный объем данных для успешного обучения ≈ 50 объектов (рис. 5).

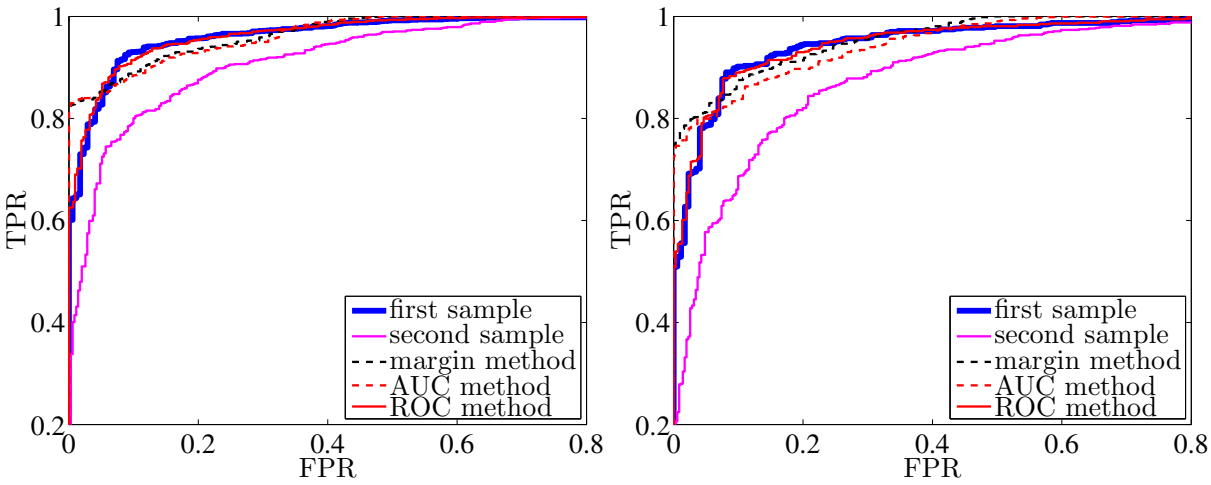
Сравнение методов отсева выбросов

Для фильтрации второй выборки к данным по каждой болезни были применены три метода:

- метод сближения ROC-кривых;
- метод сближения AUC двух выборок;
- метод, удаляющий объекты, отступы которых менее заданного порога.

На рис. 6 видно, что метод сближения AUC двух выборок не позволяет добиться однородности двух выборок, поскольку в первую очередь удаляет объекты с отрицательными отступами. Как и многие методы отсева выбросов, он подстраивается под конкретную модель классификатора, построенную по первой выборке.

На рис. 7 показаны ROC-кривые с нескольких промежуточных шагов метода сближения ROC-кривых.



(a) Болезнь А, $|P_{train}| = 1878,\ |U| = 1304$

(b) Болезнь Б, $|P_{train}| = 803,\ |U| = 1108$

Рис. 6: Сравнение трёх методов: метода сближения ROC-кривых, метода сближения AUC и метода отступов, который удаляет все объекты второй выборки, отступы которых менее заданного порога.

Видно, что только методу сближения ROC-кривых удается добиться идентичности ROC-кривых, построенных по первой и второй выборкам.

3.2.1 Полумодельные данные

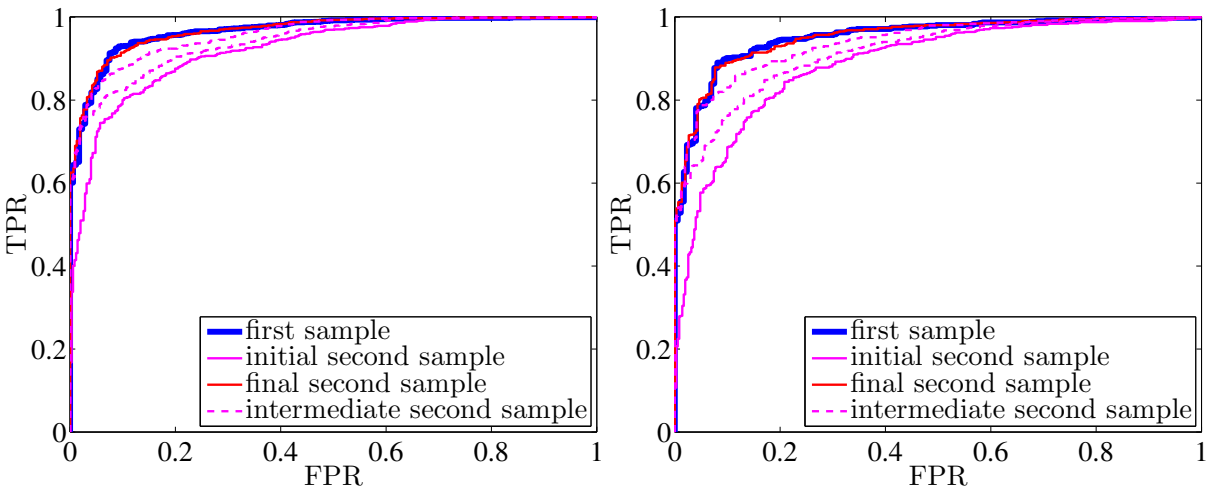
Пусть $P = P_0 \cup P_1, P_0$ — эталонная выборка здоровых, P_1 — эталонная выборка больных. Предлагается выборку P случайно разбить на две равные части со стратификацией классов: P^1, P^2, — и в выборке P^2 случайным образом переставить местами метки классов «больной/здоровый» (10-20%), затем P^1 использовать в качестве первой выборки ($P := P^1$), P^2 — в качестве второй ($U := P^2$).

Эксперимент на полумодельных данных важен, так как он позволяет проверить способность построенных алгоритмов идентифицировать и удалять из выборки U в первую очередь заранее известные выбросы.

3.2.2 Precision-Recall кривая

Пусть X — множество описаний объектов, $Y = \{0, 1\}$ — множество классов объектов, $b\colon X \to Y$ — классификатор, $b(x) = [f(x, w) \geq t]$. $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$ — выборка.

Precision-Recall кривая характеризует зависимость доли выбросов среди объек-



(a) Болезнь А, $|P_{train}| = 1878,\ |U| = 1304$ (b) Болезнь Б, $|P_{train}| = 803,\ |U| = 1108$

Рис. 7: Несколько промежуточных шагов метода сближения ROC-кривых.

тов, найденных классификатором

$$\text{Precision} = \frac{\sum\limits_{i=1}^{l} b(x_i)y_i}{\sum\limits_{i=1}^{l} b(x_i)}$$

от доли выбросов, найденных классификатором

$$\text{Recall} = \frac{\sum\limits_{i=1}^{l} a(x_i)y_i}{\sum\limits_{i=1}^{l} y_i},$$

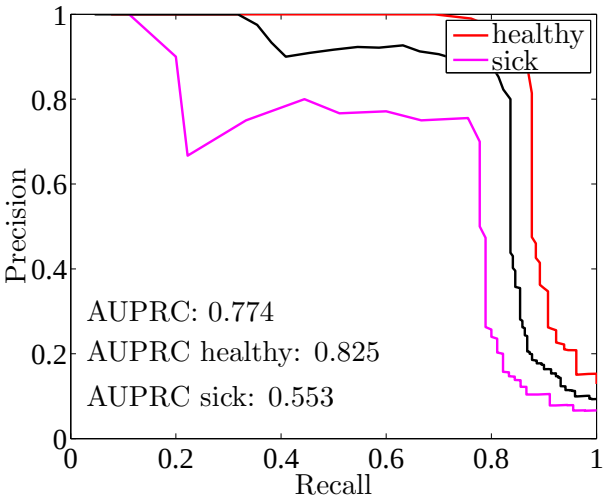
при варьировании порога t.

Точность обнаружения классификатором b выбросов ($y = 1$) определяется площадью под Precision-Recall кривой (AUPRC).

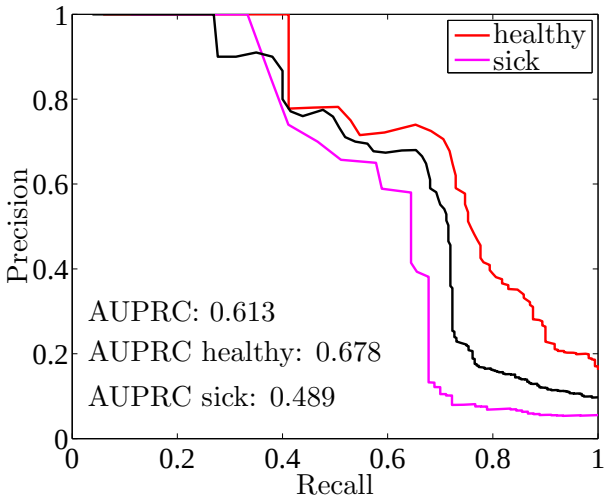
Анализ точности отсева выбросов

При выполнении экспериментов использовалась кросс-валидация по 10 блокам.

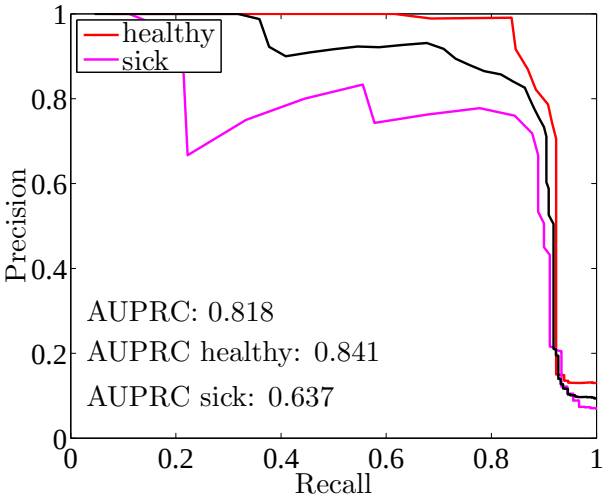
Для метода сближения ROC-кривых и метода выделения объектов, влияющих на переобучение, были построены кривые Precision-Recall. На рис. 8 видно, что данным методам удается определить большую часть выбросов сразу же, но некоторую часть найти не удается.



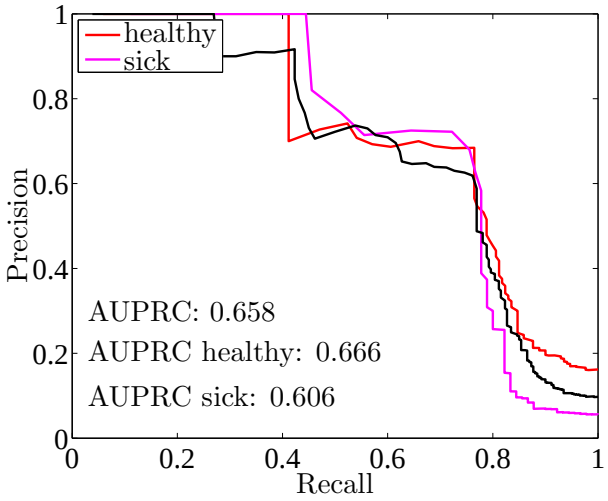
(a) Болезнь А, $|P_{train}| = 211, |P_{test}| = 25, |U| = 235$



(b) Болезнь Б, $|P_{train}| = 239, |P_{test}| = 28, |U| = 266$



(c) Болезнь А, $|P_{train}| = 211, |P_{test}| = 21, |U| = 235$



(d) Болезнь Б, $|P_{train}| = 239, |P_{test}| = 28, |U| = 266$

Рис. 8: Кривые Precision-Recall для метода сближения ROC-кривых (a,b), для метода сближения AUC двух выборок (c,d).

Для каждого метода были построены графики зависимости AUC на различных выборках, а именно на обучающей и контрольной подвыборках, на пополненной обучающей подвыборке, на второй и на отфильтрованной второй выборках (рис. 9). По графикам видно, что изначально качество классификации на второй выборке достаточно низкое, за 20–60 шагов работы алгоритма качество классификации на отфильтрованной второй выборке становится сравнимым с качеством классификации на обучающей подвыборке первой выборки. Качество классификации на контрольной подвыборке практически не изменяется при удалении выбросов из второй выборки и добавлении оставшихся объектов второй выборки в первую. Отсюда можно сделать вывод, что длина обучающей выборки достаточна. Таким образом, при достаточной длине обучающих выборок, данные методы работают как методы отсева выбросов.

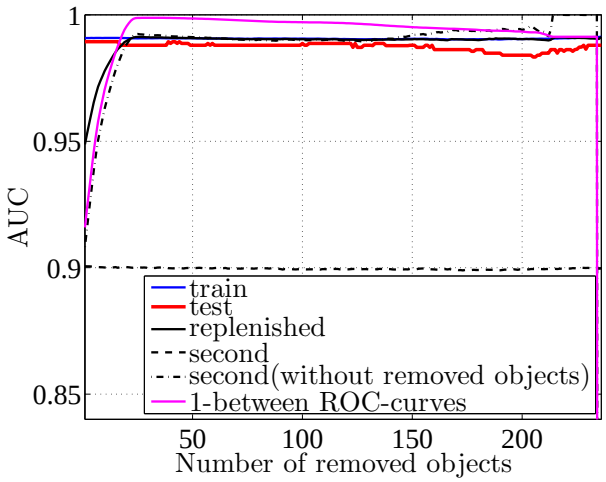
Для каждого метода были построены графики зависимости AUC на различных выборках от числа удаленных объектов, когда классификатор настраивался на обучающей выборке небольшого объема (рис. 10). Но здесь не наблюдается значимых улучшений качества классификации на контрольной выборке.

В таблице 1 видно, что пока не для всех болезней удается очищать вторую выборку от выбросов с большой точностью.

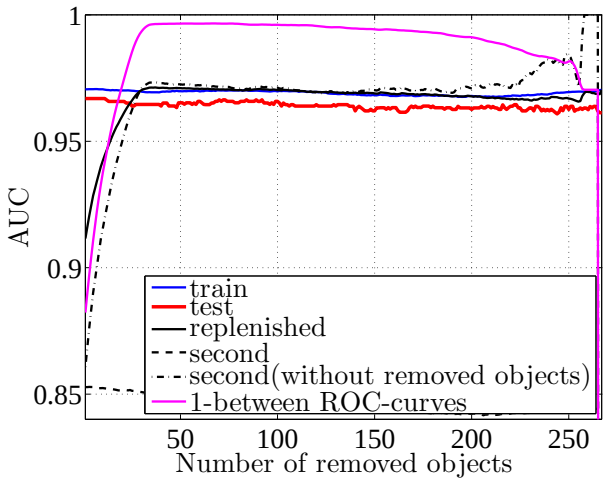
Заключение

Задача пополнения обучающих выборок отличается от обычных методов фильтрации выбросов тем, что имеется априори более надёжный источник исходных данных — «первая выборка». Предложен метод выделения объектов, влияющих на переобучение и метод выравнивания ROC-кривых, основанный на анализе геометрической формы ROC-кривых как кусочно-постоянных монотонных функций, образуемых последовательностью прямоугольников. Пока данные методы не для всех болезней позволяют очищать вторую выборку от выбросов с большой точностью. Поэтому планируется сделать модификации предложенных методов и исследовать зависимость качества классификации на независимой контрольной выборке от объема обучающей выборки.

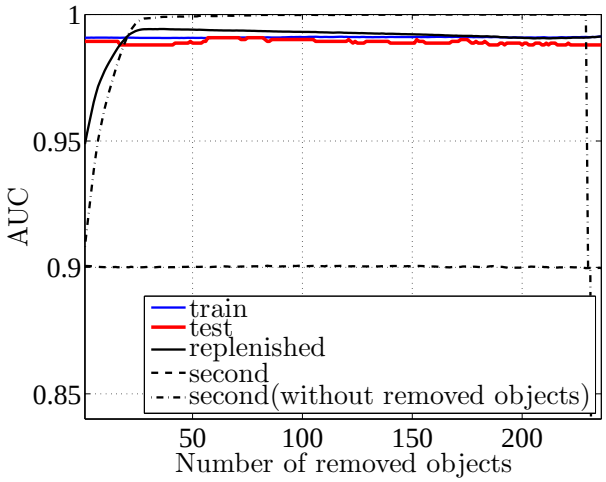
Разработанные методы предполагается использовать в рамках технологии ин-



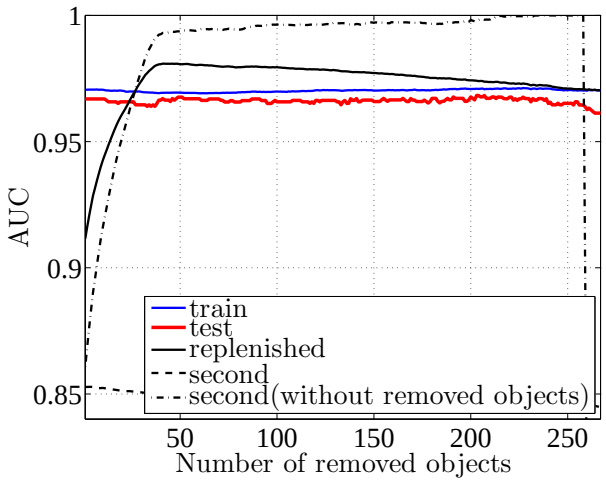
(a) Болезнь А, $|P_{train}| = 211, |P_{test}| = 25, |U| = 235$



(b) Болезнь Б, $|P_{train}| = 239, |P_{test}| = 28, |U| = 266$

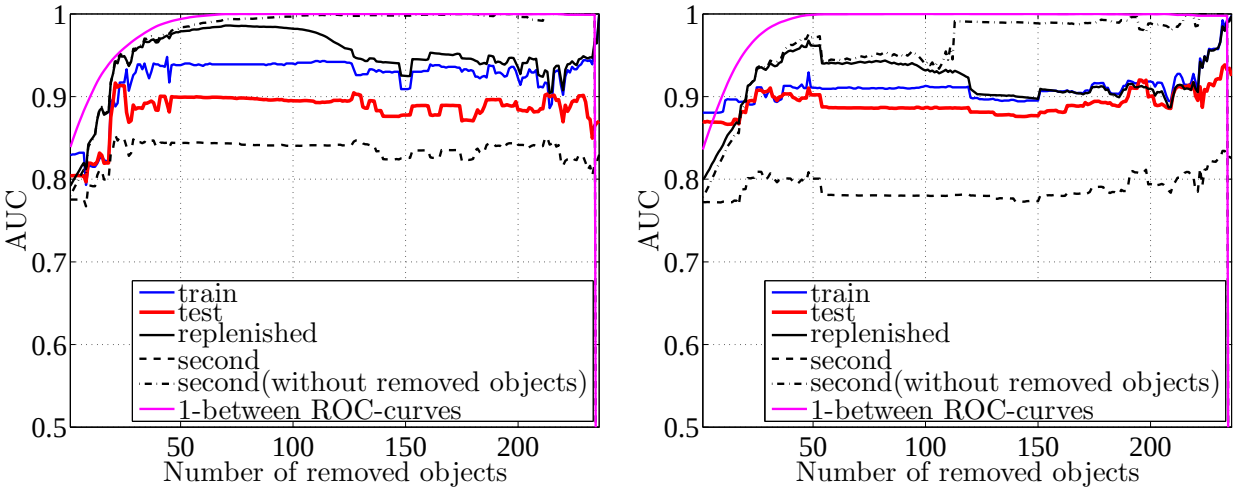


(c) Болезнь А, $|P_{train}| = 211, |P_{test}| = 21, |U| = 235$

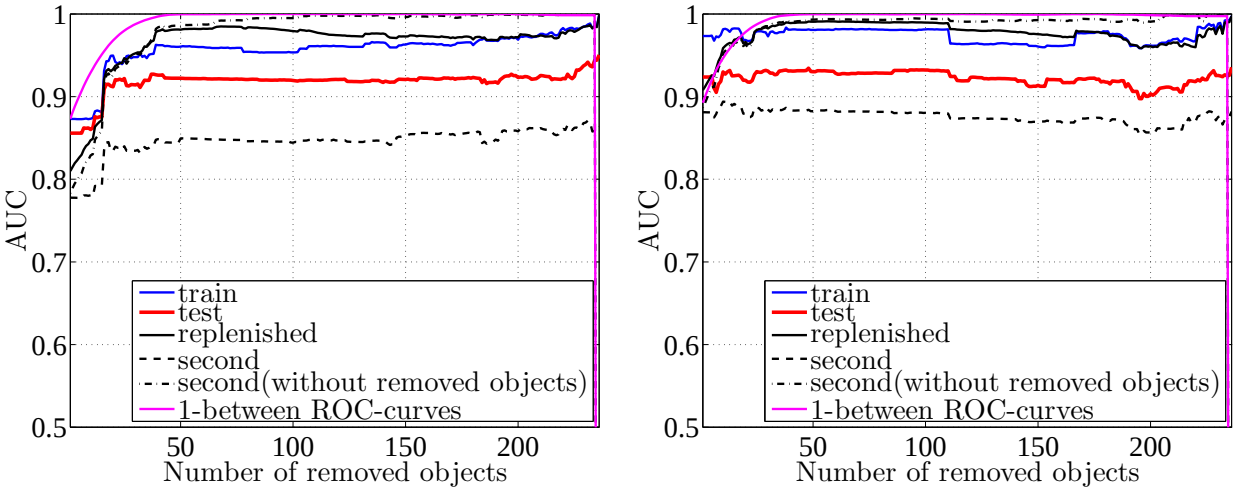


(d) Болезнь Б, $|P_{train}| = 239, |P_{test}| = 28, |U| = 266$

Рис. 9: Зависимость значения AUC на различных выборках от числа удаленных объектов из второй выборки для метода сближения ROC-кривых (a,b), для метода выделения объектов, влияющих на переобучение (c,d).



(a) $|P_{train}| = 20$ (b) $|P_{train}| = 30$



(c) $|P_{train}| = 40$ (d) $|P_{train}| = 60$

Рис. 10: Зависимость значения AUC на различных выборках от числа удаленных объектов из второй выборки для метода сближения ROC-кривых, $|P_{test}| = 118$, $|\mathrm{U}| = 235$ (болезнь A) .

Таблица 1: Точность обнаружения выбросов.

Болезнь	AUPRC	
	Метод сближения ROC-кривых	Метод выделения объектов, влияющих на переобучение
А	0.774	**0.818**
Б	0.613	**0.658**
В	0.661	**0.796**
Г	0.429	**0.460**
Д	0.617	**0.862**
Е	0.565	**0.626**
Ж	**0.854**	0.819
З	**0.817**	0.804
И	0.668	**0.778**
К	0.776	**0.803**
Л	0.696	**0.783**
М	0.591	**0.720**
Н	0.484	**0.568**
О	0.738	**0.786**

формационного анализа электрокардиосигналов [37] для повышения качества диагностики заболеваний путём фильтрации выбросов и формирования однородных обучающих выборок.

Список публикаций

1. Ефимова И. В. Формирование однородных обучающих выборок для задач медицинской диагностики // Труды 57-ой международной научной конференции МФТИ, 2014, С. 91–92.

2. Успенский В. М., Воронцов К. В., Целых В. Р., Бунаков В. А., Ефимова И. В., Полежаев В. А. Информационный анализ электрокардиосигналов для диагно-

стики многих заболеваний внутренних органов по одной электрокардиограмме // Интеллектуализация обработки информации (ИОИ-2014): Тезисы докл. – Москва: Торус Пресс, 2014. С. 172–173.

Список литературы

[1] Edgeworth F. Y. On discordant observations // Philosophical Magazine, 1887. Vol. 23, No. 5, Pp. 364 –375.

[2] Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys (CSUR), July 2009. Vol. 41(3), No. 15.

[3] Hodge V. and Austin J. A survey of outliers detection methodologies // Artificial Intelligence Review, 2004. Vol. 22(2), Pp. 85–126.

[4] Zhang J. Advancements of Outlier Detection: A Survey // ICST Transactions on Scalable Information Systems, 2013. Vol. 13(1), Pp. 1–26.

[5] Aggarwal C. C. Outlier Analysis // Springer, 2013.

[6] Domingos P. MetaCost: A General Framework for Making Classifiers Cost-Sensitive // ACM KDD Conference, 1999.

[7] Zadrozny B., Langford J., Abe N. Cost-Sensitive Learning by Cost-Proportionate Example Weighting // ICDM Conference, 2003.

[8] Theiler J., Cai D. M. Resumpling approach for anomaly detection in miltispectral images // In Proceedings of SPIE 5093, 2003. Pp. 230–240.

[9] Abe N., Zadrozny B., Langford J. Outlier detection by active learning. In Proceedings of the 12th ACM SIGKDD Internetional Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, USA, Pp. 504–509.

[10] Steinwart I., Hush D., Scovel C. A classificational framework for anomaly detection // Journal of Maching Learning Research 6, 2005, Pp 211-232.

[11] Drummond C., Holte R. C4.5, Class Imbalance, and Cost Sensitivity: Why Undersampling beats Oversampling // ICML Workshop on Learning from Imbalanced Data Sets, 2003.

[12] Chan P. K., Stolfo S. J. Toward Scalable Learning with Nonuniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection // KDD Conference, 1998. Pp. 164–168.

[13] Kubat M. and Matwin S. Addressing the Curse of Imbalanced Training Sets: One Sided Selection // ICML Conference, 1997.

[14] Elkan C. The Foundations of Cost-Sensitive Learning // IJCAI, 2001.

[15] Zhang J., Mani I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction // Proceedings of the ICML Workshop on Learning from Imbalanced Datasets, 2003.

[16] Ting K. M. An Instance-weighting Method to Induce Costsensitive Trees // IEEE Transaction on Knowledge and Data Engineering, 2002, Vol. 14, Pp. 659–665.

[17] Weiss G., Provost F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction // Journal of Artificial Intelligence Reserach, 2003, Vol. 19, Pp. 315–354.

[18] Tang Y., Zhang Y. Q., Chawla N. V., Krasser S. SVMs Modeling for Highly Imbalanced Classification // IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, 2009, Vol. 39(1), Pp. 281–288.

[19] Wu G., Chang E. Y. Class-boundary Alignment for Imbalanced Dataset Learning // Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets, 2003.

[20] Liu B., Lee W. S., Yu P., Li X. Partially supervised classification of text documents // ICML-02, 2002.

[21] Yu H., Han J., Chang K. PEBL: Positive example based learning for Web page classification using SVM // KDD-02, 2002.

[22] Li X., Liu B. Learning to classify text using positive and unlabeled data // IJCAI-03, 2003.

[23] Rocchio J. Relevant feedback in information retrieval.In G. Salton (ed.). The smart retrieval system- experiments in automatic document processing, Englewood Cliffs, NJ, 1971.

[24] Bockhorst J., Craven M. (2002) Exploiting relations among concepts to acquire weakly labeled training data // ICML-02, 2002.

[25] Denis F., Gilleron R., Tommasi, M. Text classification from positive and unlabeled examples // IPMU, 2002.

[26] SchËolkopf B., Platt J. C., Shawe-Taylor J., Smola A. J., Williamson R. C.. Estimating the support of a high-dimensional distribution // Neural Computation, 2001, Vol. 13(7), Pp. 1443–1471.

[27] Tax D. M. J., Duin R. P. W. Support vector data description // Machine Learning, 2004, Vol. 54(1), Pp. 45–66.

[28] Liu B., Dai Y., Li X., Lee W. S., Yu P. S. Building text classifiers using positive and unlabeled examples // In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 2003, Pp. 179–188.

[29] Nigam K., McCallum A., Thrun S., Mitchell T. Learning to classify text from labeled and unlabeled documents. AAAI-98 (Pp. 792–799). Madison, US: AAAI Press, Menlo Park, US.

[30] Joachims T. Transductive inference for text classification using support vector machines // Proceedings of ICML-99, 16th International Conference on Machine Learning, 1999, Pp. 200–209.

[31] Blum A., Mitchell T. Combining labeled and unlabeled data with co-training // COLT: Proceedings of the Workshop on Computational Learning Theory, 1998.

[32] Dasgupta D., Nino F. A comparison of negative and positive selection algorithms in novel pattern detection // In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2000, Vol. 1. Nashville, TN, Pp. 125–130.

[33] Dasgupta D., Majumdar N. Anomaly detection in multidimensional data using negative selection algorithm // In Proceedings of the IEEE Conference on Evolutionary Computation, Hawaii, 2002, 1039–1044.

[34] Forrest S., Warrender C., Pearlmutter B. Detecting intrusions using system calls: Alternate data models // In Proceedings of the 1999 IEEE ISRSP. IEEE Computer Society, Washington, DC, USA, 1999, Pp. 133–145.

[35] Liu B., Dai Y., Li X., Lee W. S., Yu P. S. Building text classifiers using positive and unlabeled examples // In Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 2003, Pp. 179–188.

[36] Elkan C., Noto K. Learning Classifiers from only Positive and Unlabeled Data // ACM KDD Conference, 2008

[37] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информация 2008. 116 с.

[38] Uspenskiy V. M., Vorontsov K. V., Tselykh V. R., Bunakov V. A. Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System // in Advances in Mathematical and Computational Tools in Metrology and Testing X (vol.10), Series on Advances in Mathematics for Applied Sciences, vol. 86, World Scientific, Singapore (2015). Pp. 377–384.

[39] Баевский Р. М., Иванов Г. Г. Вариабельность сердечного ритма: теоретические аспекты и возможности клинического применения // Ультразвуковая и функциональная диагностика. 2001, №3, С. 108–127.

[40] Баевский Р. М., Иванов Г. Г., Чирейкин Л. В., Гаврилушкин А. П., Довгалевский П. Я., Кукушкин Ю. А., Миронова Т. Ф., Прилуцкий Д. А., Семенов Ю. Н., Федоров В. Ф., Флейшман А. Н., Медведев М. М. Анализ вариабельности сердечного ритма при использовании различных электрокардиографических систем (методические рекомендации) // Вестник аритмологии. 2001, №24, С. 65–87.