Формирование однородных обучающих выборок в задачах классификации

Ефимова Ирина

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. К. В. Воронцов

1 июля 2015

Цель исследования

Имеются две размеченные выборки двух классов. *Первая выборка* эталонная, *вторая выборка* содержит неизвестную долю выбросов — объектов с неверной классификацией.

Цель исследования: построить алгоритм, позволяющий очищать вторую выборку от выбросов, для получения одной однородной выборки.

Литература

- Aggarwal C. C. Outlier Analysis. Springer, 2013.
- Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys (CSUR), July 2009. Vol.41(3), No.15.
- Hodge V. and Austin J. A survey of outliers detection methodologies //Artificial Intelligence Review, 2004. Vol. 22(2), Pp. 85–126.
- Успенский В.М. Информационная функция сердца // Клиническая медицина. 2008. Т.86, №5. С. 4–13.

Задача пополнения обучающей выборки

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

Предположения:

- 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:

$$g : (P, U) \longrightarrow Q.$$

Задача пополнения обучающей выборки

Критерий

Качество классификации на независимой контрольной выборке при обучении на объединенной выборке $P \cup Q$ выше по сравнению с обучением только по выборке P.

Пример задачи из области медицинской диагностики:

первая выборка состоит из обследований с надежно установленными диагнозами; во *второй выборке* диагнозы не были подтверждены лабораторными и инструментальными исследованиями.

Алгоритм

Для осуществления пополнения выборки P объектами из выборки U в обычную процедуру построения классификатора добавляются шаги 2 и 3:

1. обучиться на P_{train};
2. отфильтровать выборку U, то есть получить Q;
3. обучиться на $P_{train} \cup Q$;
4. классифицировать P_{test}.

ROC-кривая

Качество классификатора $a\colon X \to Y$, $Y = \{0, 1\}$, $a(x) = [f(x, w) \geq \theta]$ определяется площадью под ROC-кривой (AUC).
$X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$ — выборка.

ROC-кривая — зависимость доли верных положительных классификаций (True Positive Rate) от доли ложных положительных классификаций (False Positive Rate) при варьировании порога классификатора θ.

$$TPR = \frac{\sum_{i=1}^{l} a(x_i) y_i}{\sum_{i=1}^{l} y_i}, \quad FPR = \frac{\sum_{i=1}^{l} a(x_i)(1 - y_i)}{\sum_{i=1}^{l} (1 - y_i)}$$

Предлагаемые методы

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

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.

Для выравнивания кривых ROC_1 и ROC_2 предлагается на t-ом шаге алгоритма исключать объект из выборки U_{t-1}, минимизирующий разность AUC:

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

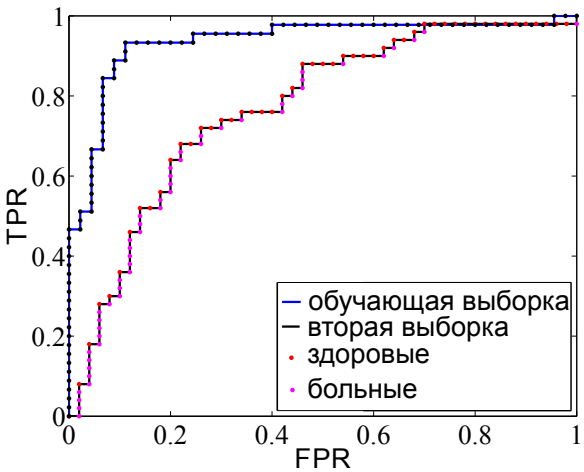
Метод сближения ROC-кривых

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.

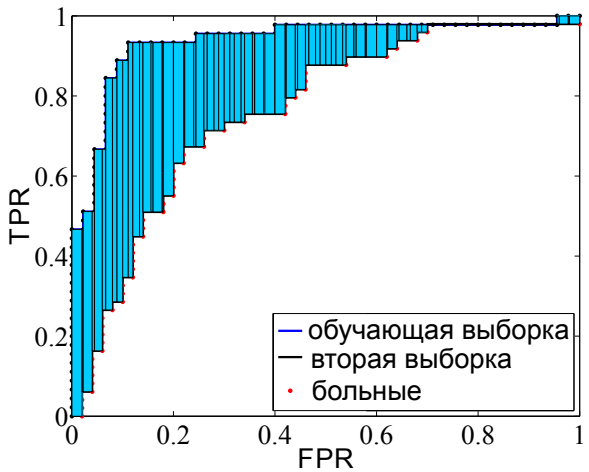
Для выравнивания ROC_1 и ROC_2 кривых предлагается на t-ом шаге алгоритма исключать объект из выборки U_{t-1}, минимизирующий площадь между ROC_1 и $\mathrm{ROC}_{2,t-1}$ кривыми:

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

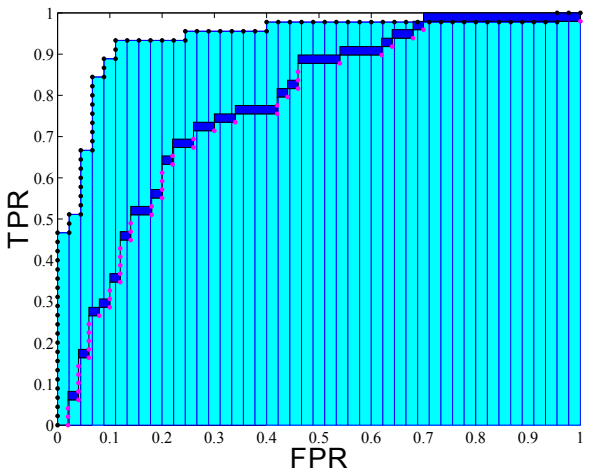
Эффективный способ вычисления площади между ROC-кривыми



(a)



(b)



(c)

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;
$\text{AUC}_{1,t}$ — по выборке $P_{train,t} = P \cup (x_t, y_t)$,
$\text{AUC}_{2,t}$ — по выборке $U_t = U \setminus (x_t, y_t),\ x_t \in U$.

Переобученность:
$d_0 = \text{AUC}_1 - \text{AUC}_2,\ d_t = \text{AUC}_{1,t} - \text{AUC}_{2,t}$.

Влияние на переобучение:

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

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

Цели эксперимента

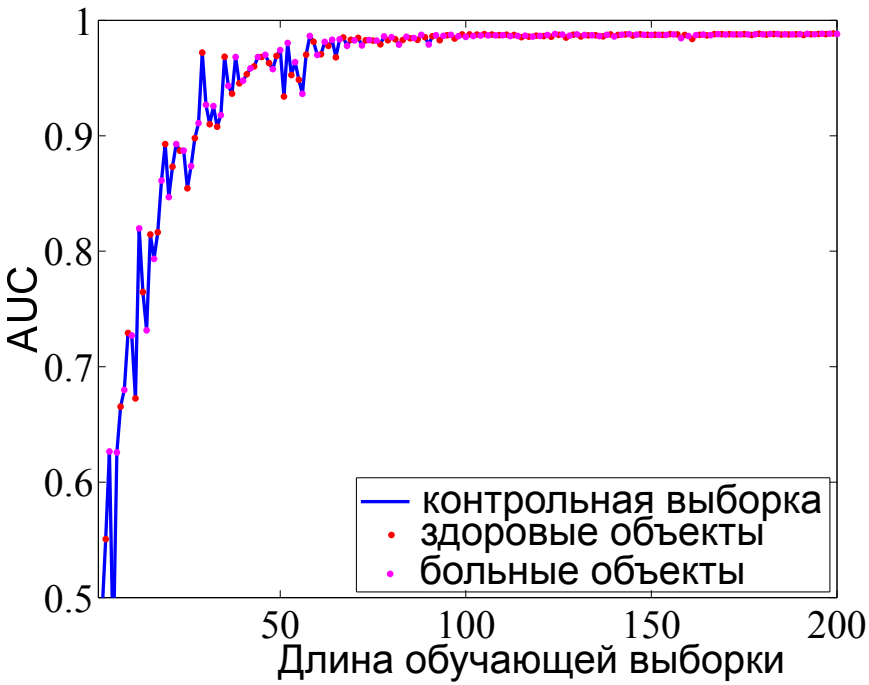
- Оценить достаточную длину обучающей выборки.
- Очистить вторую выборку от выбросов.
- Повысить качество классификации на пополненной выборке.
- Сравнить методы отсева выбросов и пополнения выборки.

Задача диагностики заболеваний по ЭКГ

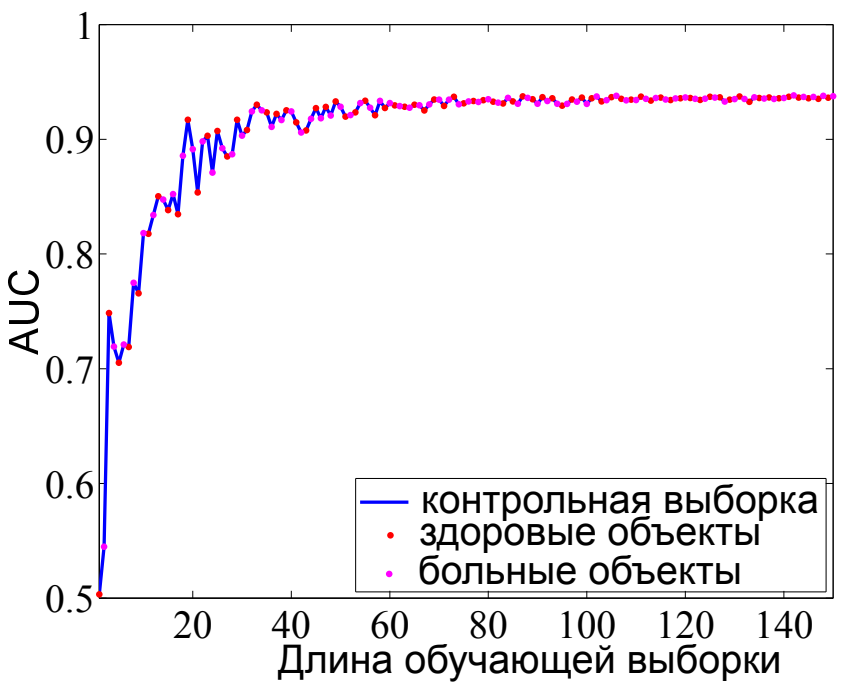
- Использовались данные, полученные с помощью технологии информационного анализа ЭКГ-сигналов.
- Каждому обследованию соответствует вектор из 216 числовых признаков и метка класса:
  0 — здоров, 1 — болен.
- Известно, что синдромный алгоритм с отбором признаков дает хорошее качество классификации, не переобучается.

Кривые обучения

Для обучения достаточно ≈ 50 объектов.

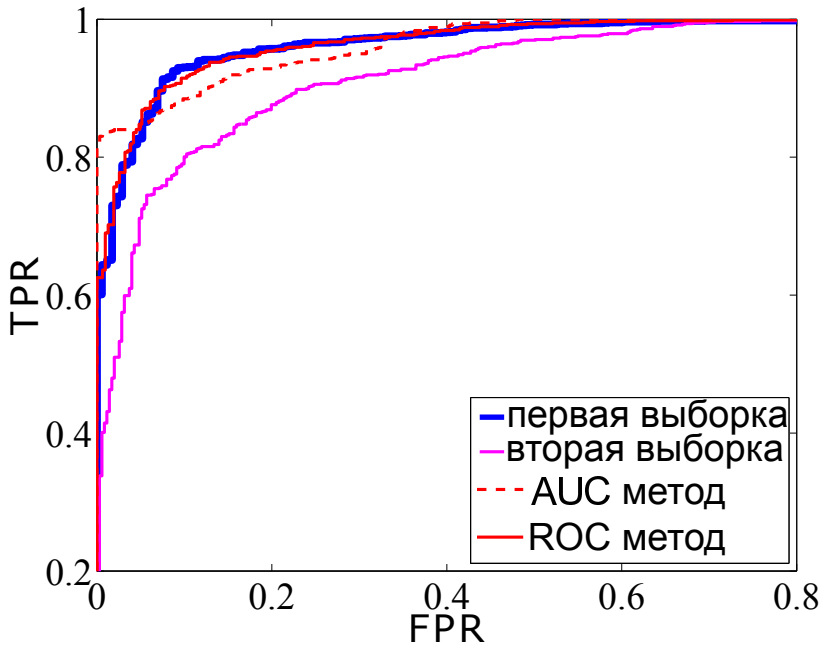


(d) Болезнь А, $|P_{test}| = 235$

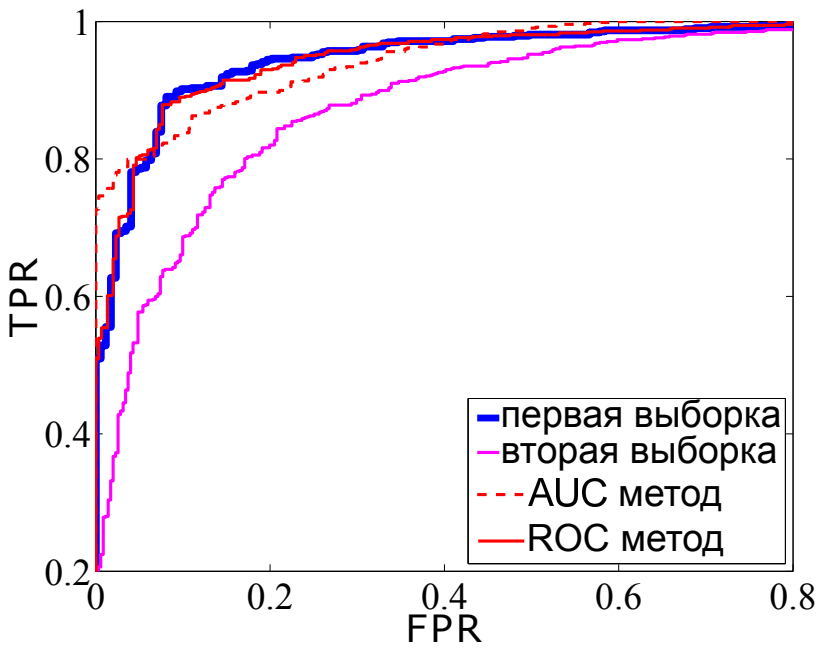


(e) Болезнь Б, $|P_{test}| = 470$

Сравнение методов отсева выбросов



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

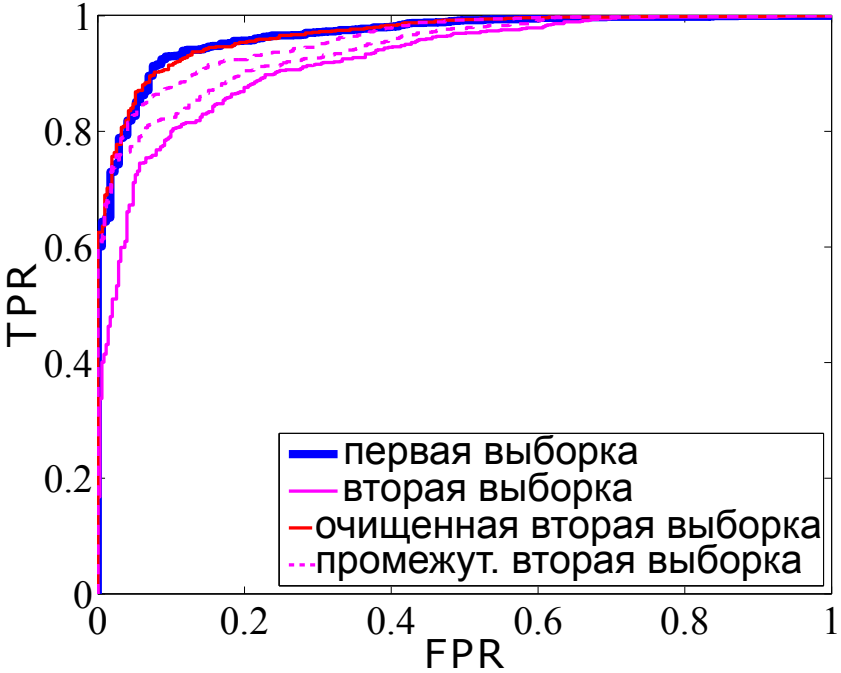


(g) Болезнь Б,
$|P_{train}| = 803, |U| = 1108$

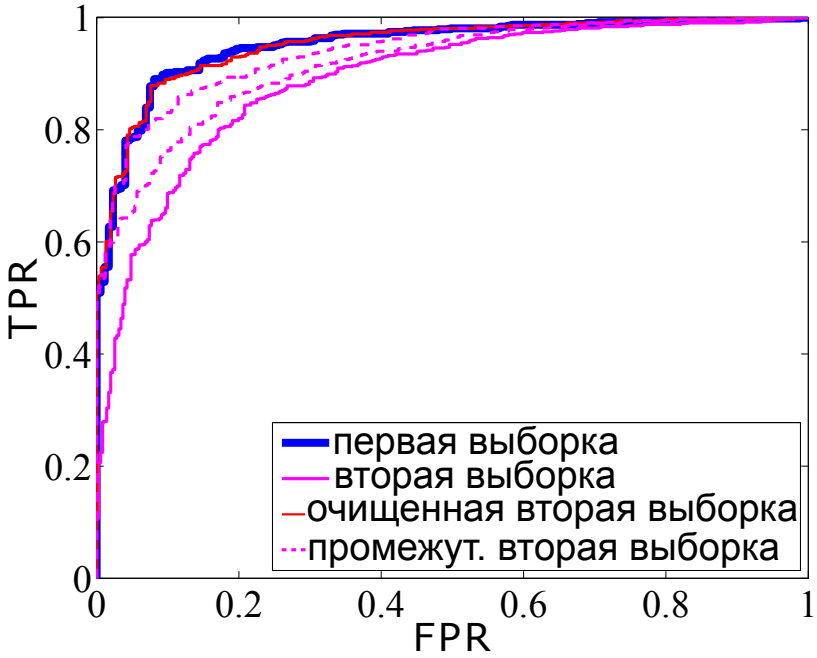
Вывод: Метод сближения AUC приводит к явной неоднородности двух выборок, так как в первую очередь удаляет объекты с отрицательными отступами.

Метод сближения ROC-кривых

Несколько промежуточных шагов.



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



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

Вывод: Данному методу удается добиться идентичности ROC-кривых двух выборок.

Полумодельные данные

$P = P_0 \cup P_1$, P_0 — эталонная выборка здоровых,
P_1 — эталонная выборка больных.

Получение:

- P случайно разбить на две равные части со стратификацией классов: P^1, P^2;
- в P^2 случайным образом переставить местами метки классов «больной/здоровый» (10-20%);
- использовать P^1 в качестве первой выборки ($P := P^1$), P^2 — в качестве второй ($U := P^2$).

Precision-Recall кривая

Точность обнаружения классификатором $b\colon X \to Y$, $Y = \{0, 1\}$, $b(x) = [f(x, w) \geq t]$ выбросов ($y = 1$) определяется площадью под Precision-Recall кривой (AUPRC).
$X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$ — выборка.

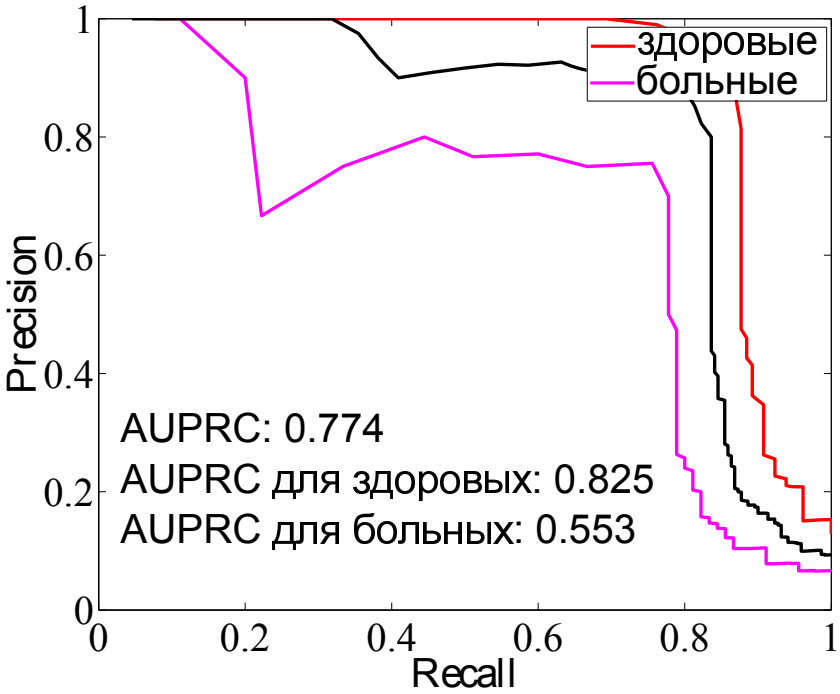
Precision — доля выбросов среди объектов, найденных классификатором;

Recall — доля выбросов, найденных классификатором.

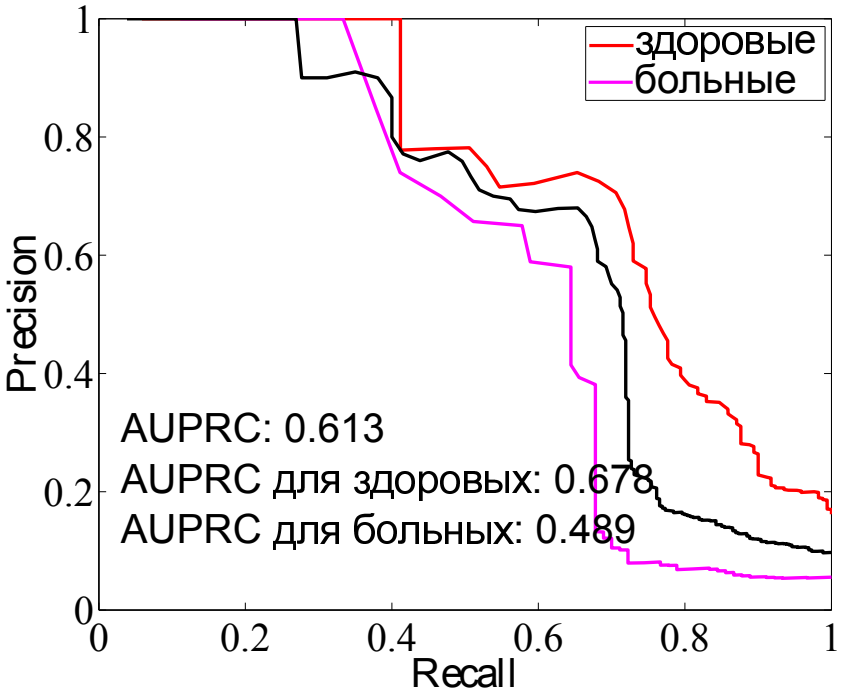
$$\text{Precision} = \frac{\sum_{i=1}^{l} b(x_i) y_i}{\sum_{i=1}^{l} b(x_i)}, \quad \text{Recall} = \frac{\sum_{i=1}^{l} b(x_i) y_i}{\sum_{i=1}^{l} y_i}.$$

Метод сближения ROC-кривых

Точность обнаружения выбросов. Кривые Precision-Recall.



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

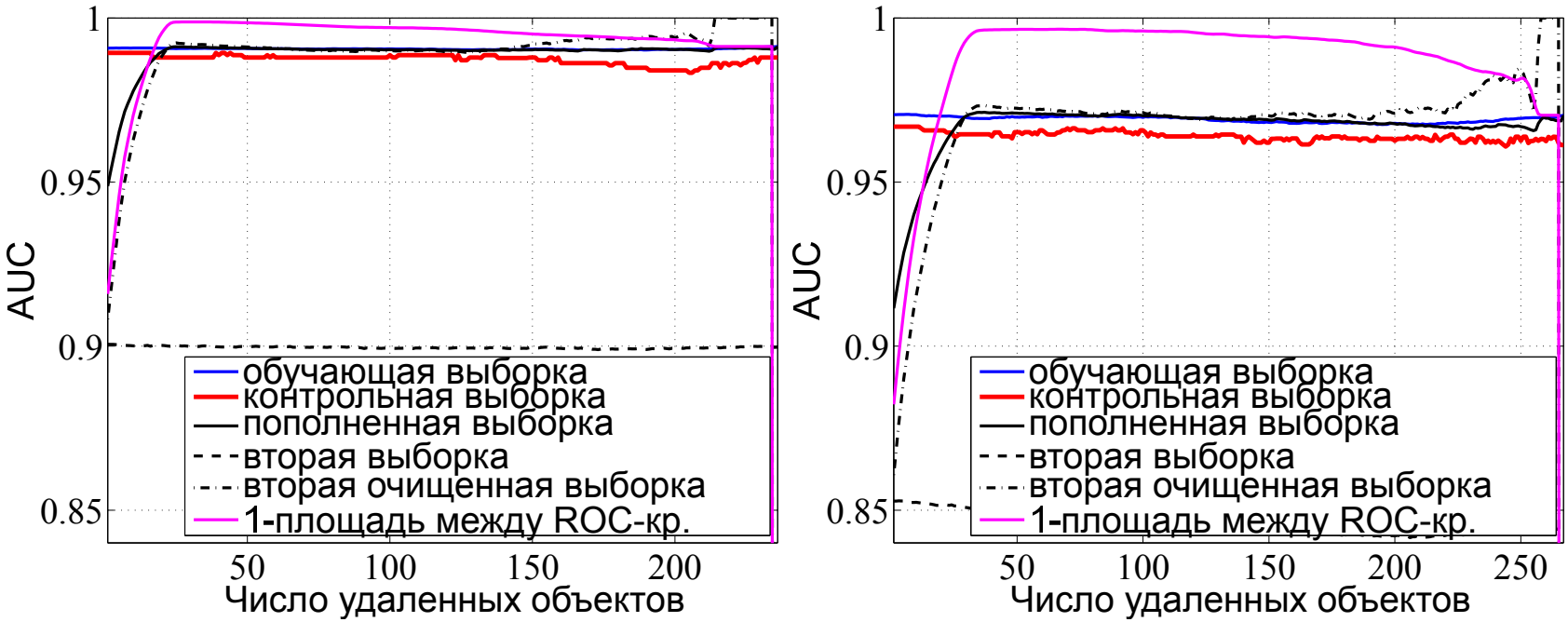


(k) Болезнь Б,
$|P_{train}| = 239, |P_{test}| = 28, |U| = 266$

Вывод: Часть выбросов не идентифицируются.

Метод сближения ROC-кривых

Зависимость значения AUC на различных выборках от числа удаленных объектов из выборки U.



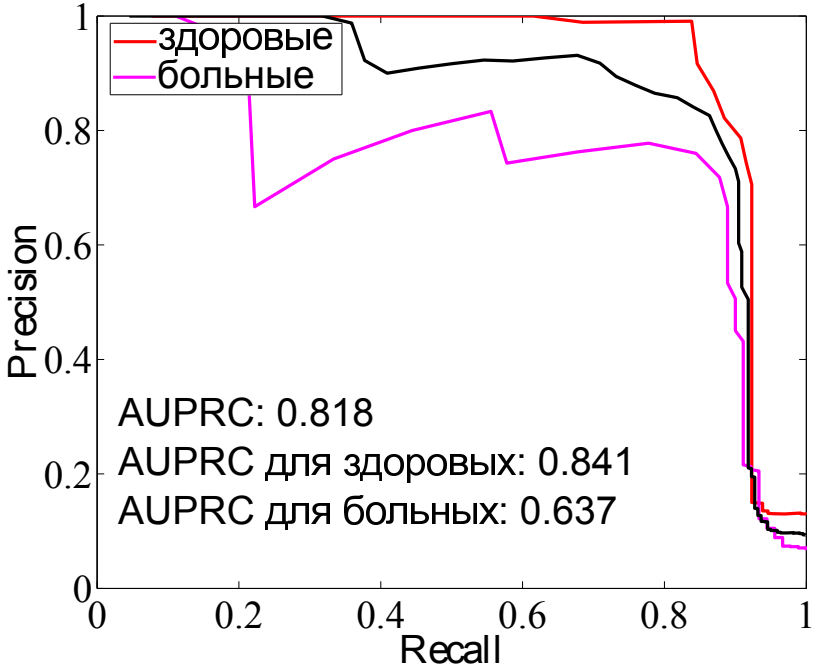
(l) Болезнь А, $|P_{train}| = 211, |P_{test}| = 25, |U| = 235$

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

Вывод: Качество классификации на контрольной выборке не изменилось ввиду достаточной длины обучающей выборки.

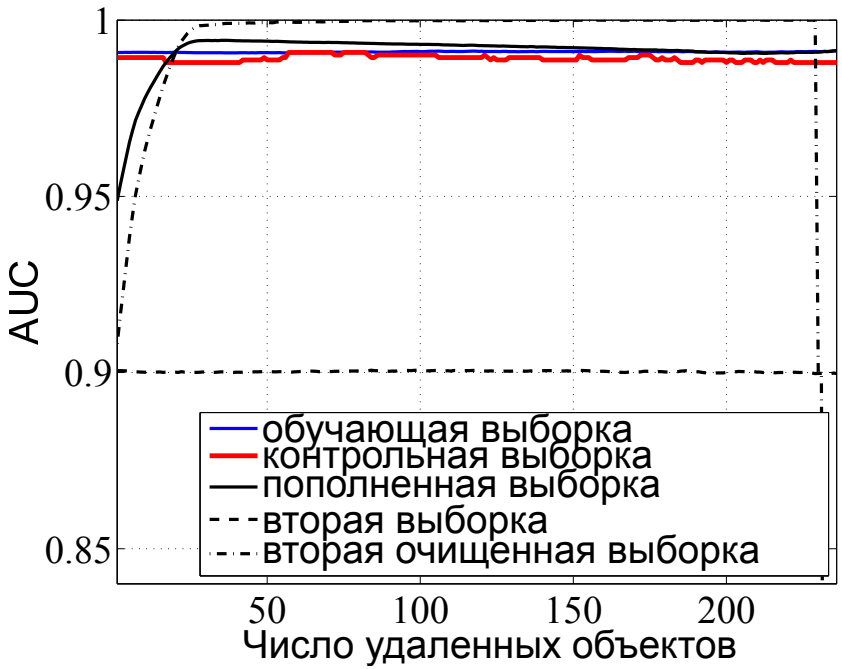
Метод выделения объектов, влияющих на переобучение

Зависимость значения AUC на различных выборках от числа удаленных объектов из U и кривые Precision-Recall для бол. А.



(n)

$|P_{train}| = 211, |P_{test}| = 21, |U| = 235$



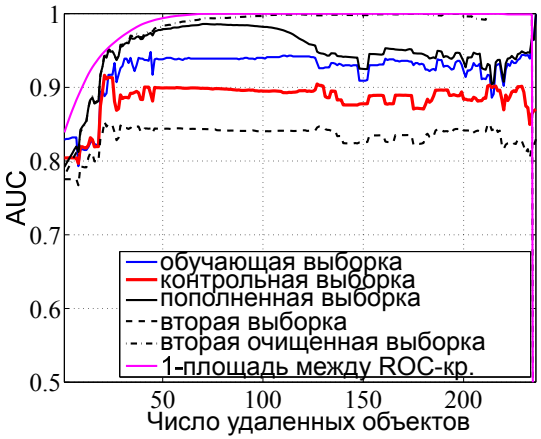
(o)

$|P_{train}| = 211, |P_{test}| = 25, |U| = 235$

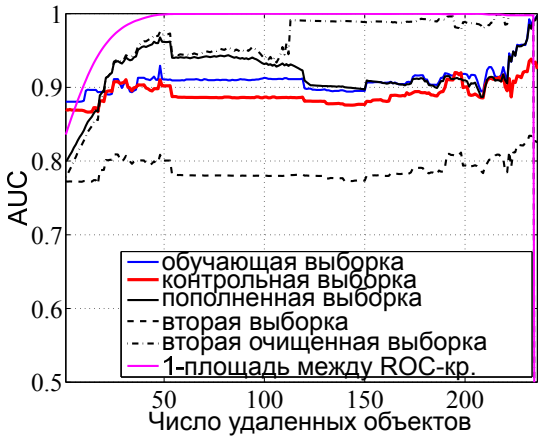
Вывод: Часть выбросов не идентифицируются.

Метод сближения ROC-кривых

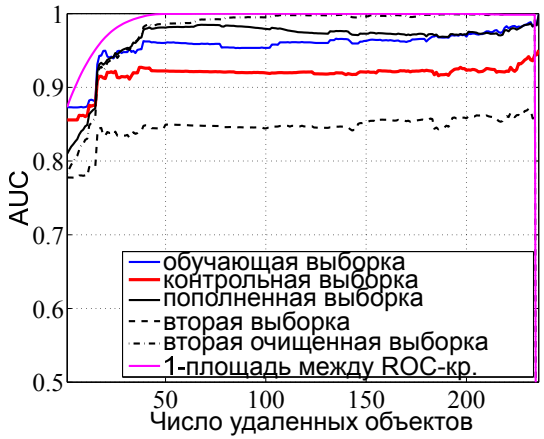
Зависимость значения AUC на различных выборках от числа удаленных объектов из U при различной длине обучающей выборки для бол. А.



(p) $|P_{train}| = 20$, $|P_{test}| = 118$, $U = |235|$



(q) $|P_{train}| = 30$, $|P_{test}| = 118$, $U = |235|$



(r) $|P_{train}| = 40$, $|P_{test}| = 118$, $U = |235|$

Вывод: Нет значимых улучшений в классификации независимой контрольной выборки.

Анализ точности 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**

Заключение

Для решения задачи пополнения обучающей выборки предложены:

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

Разработанные методы предполагается использовать в рамках технологии информационного анализа электрокардиосигналов для повышения качества диагностики заболеваний.

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