Содержание

1 Описание данных

1.1 Исходные данные

N пациентов, болеющих одной или несколькими из L болезней. (Для $L = 32$ $N = 15183$.) Для каждого пациента известны:

- вектор меток болезней $[d_{ij}]$, где i — номер пациента, $i = 1, \ldots, N$; j —- номер болезни; $j = 1, \ldots, L$;

$$d_{ij} = \begin{cases} 0, \text{ болезни нет или ее наличие не проверялось;} \\ 1, \text{ болезнь есть (диагностирована врачом);} \\ 2, \text{ болезнь есть,эталонный случай (особо достоверный диагноз).} \end{cases}$$

- вектор частот триграмм в его кодограмме $[n_{ik}]$, где i —- номер пациента, $i = 1, \ldots, N$; k — номер триграммы, $k = 1, \ldots, 216$. n_{ik} —- сколько раз триграмма k встретилась в кодограмме i-ого пациента.

1.2 Бинаризация данных

$$y_{ij} = \begin{cases} 0, \text{ если } d_{ij} = 0 \\ 1, \text{ иначе} \end{cases} ; \; x_{ik} = \begin{cases} 1, \; n_{ik} \geq 2 \\ 0, \text{ иначе} \end{cases}$$

2 Описание алгоритма классификации

Для пары классов (кластеров) строится бинарный наивный байесовский классификатор

$$a(\mathbf{x}_i) = sign\left(\sum_{k=1}^{K} x_{ik} w_k + w_0\right) = sign(f(\mathbf{x}_i) + w_0),$$

где $y_i \in \{+1, -1\}$ — метка класса, $x_{ik} \in \{0, 1\}$ — бинарный признак, i — номер объекта, k — номер признака, K — количество значимых признаков,

$$w_k = \ln \frac{N^k_{1+} N^k_{0-}}{N^k_{1-} N^k_{0+}} \text{ — вес признака; } N^k_{xy} = \sum_{i=1}^{l} [y_i = y][x_{ik} = x]$$

Значимыми считаются признаки, имеющие наибольшие по модулю веса. В качестве критерия качества используется AUC. Алгоритм вычисления AUC (для случая пересечения классов):

FPR = 0;

TPR = 0;

AUC = 0;

X = -1*[y,X];

X = -1*sortrows(X,1); //сортировка объектов по убыванию дискриминантной функции $f(\mathbf{x}_i)$

NumRow = size(X,1);

for i = 1:1:NumRow

if (any(data(X(i,2),y1))) //если объект относится к классу "−1"

FPR = FPR + 1/Nminus;

AUC = AUC + TPR/Nminus;

end

if (any(data(X(i,2),y2))) //если объект относится к классу "+1"

TPR = TPR + 1/Nplus;

end

end

Таким образом, если объект относится к обоим классам, то будут выполнены все три действия в следующем порядке:

FPR = FPR + 1/Nminus;
AUC = AUC + TPR/Nminus;
TPR = TPR + 1/Nplus;

3 Кластеризация болезней с использованием AUC в качестве расстояния между кластерами

На обучении K подбирается следующим образом: признаки сортируются в порядке убывания их модулей, в этом же порядке они добавляются в классификатор, и считается AUC на обучающей выборке для каждого K от 1 до 216, выбирается K с наибольшим AUC.

Схема вычисления AUC двух классов (кластеров) — 10-fold CV. Выбираются объекты, относящиеся к определенному классу, полученное множество случайным образом делится на 10 примерно равных частей, то же самое проводится со вторым классом. Получаются две выборки, содержащие по одному классу и состоящие из 10 частей. Далее сливаются части из разных выборок, получается одна выборка, содержащая два класса и состоящая из 10 частей. Далее 9 из 10 частей сливаются в обучающую выборку, оставшаяся часть становится тестовой, на которой вычисляется AUC. Это действие повторяется 10 раз (перебираются все варианты). Из 10 значений AUC берется среднее арифметическое. Точно так же усредняются количество значимых признаков и веса признаков, только количество значимых признаков K еще округляется до ближайшего целого после вычисления среднего арифметического.

После слияния пары болезней с наименьшим AUC в один кластер он считается "агрегированной болезнью"и для каждой пары "новый кластер — болезнь (возможно, "агрегированная")"проводится классификация и подсчет AUC.

3.1 С пересечениями классов

В начале для каждой болезни создается выборка, поделенная на 10 частей. При слиянии болезней сливаются соответствующие части из выборок для этих двух болезней.

3.2 Без пересечений классов

Для каждой пары болезней (даже если они агрегированные) выборки создаются каждый раз заново. То есть из исходных данных выбираются все пациенты, болеющие 1-ой болезнью (1-ым набором болезней), и все

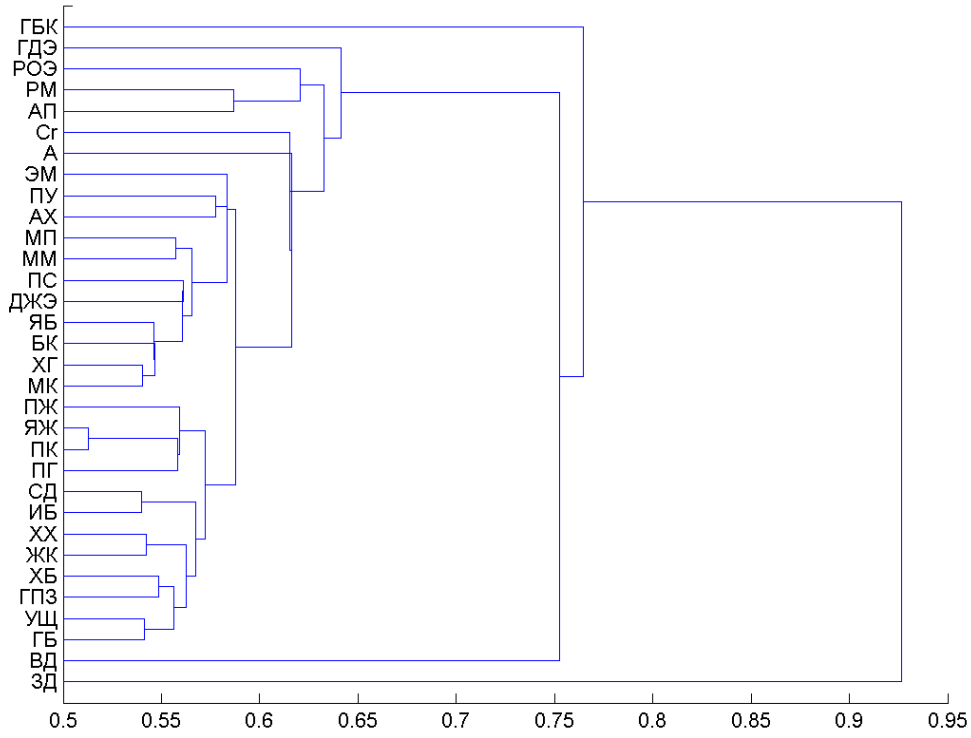


Рис. 1: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (с пересечениями классов). Дендрогамма для 32 болезней

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	50	87	94	95	95	93	93	95	93	91	95	89	95	93	92	98	96	93	94	90	92	95	90	93	93	93	94	92	96	95	91	93
ВД	87	50	73	76	78	69	68	76	71	66	73	63	78	64	67	84	69	67	79	70	68	77	67	77	72	76	70	62	78	74	61	75
ГБ	94	73	50	60	60	57	60	57	54	56	56	65	60	64	58	76	66	59	62	62	56	55	62	60	63	57	59	63	63	56	63	57
ЖК	95	76	60	50	67	63	65	59	58	65	54	71	60	67	62	72	65	62	66	67	63	60	65	63	65	63	63	65	59	59	68	62
ИБ	95	78	60	67	50	63	65	54	58	62	56	71	58	68	63	75	69	63	59	66	61	58	67	60	66	60	64	69	65	58	68	62
МК	93	69	57	63	63	50	57	63	58	54	59	59	63	61	55	80	62	55	68	65	55	60	59	63	62	62	56	61	63	59	60	58
ММ	93	68	60	65	65	57	50	65	58	57	60	59	66	58	56	79	66	59	69	61	58	62	56	69	60	62	59	58	64	61	59	59
СД	95	76	57	59	54	63	65	50	60	64	58	71	56	67	63	77	70	64	59	65	61	59	67	63	64	60	63	69	61	56	68	60
УЩ	93	71	54	58	58	58	58	60	50	57	56	64	61	62	56	75	64	60	66	61	56	60	58	59	57	56	60	60	62	57	63	56
ХГ	91	66	56	65	62	54	57	64	57	50	59	57	64	61	55	80	66	57	67	63	57	60	58	65	59	61	60	57	66	60	58	60
ХХ	95	73	56	54	56	59	60	58	56	59	50	67	60	60	56	73	63	57	63	64	57	57	62	64	64	61	57	65	63	56	63	62
А	89	63	65	71	71	59	59	71	64	57	67	50	72	62	61	84	69	62	72	67	60	67	60	69	63	66	65	60	72	68	59	65
АП	95	78	60	60	58	63	66	56	61	64	60	72	50	71	63	78	69	65	61	66	64	57	67	61	66	55	63	68	59	56	71	59
АХ	93	64	64	67	68	61	58	67	62	61	60	62	71	50	60	76	68	61	72	66	59	65	62	73	63	69	62	58	68	63	61	67
ЯБ	92	67	58	62	63	55	56	63	56	55	56	61	63	60	50	79	64	56	69	65	56	61	57	65	59	59	56	57	63	58	59	62
ГБК	98	84	76	72	75	80	79	77	75	80	73	84	78	76	79	50	74	76	78	79	79	78	79	82	81	83	78	77	79	77	79	79
ГДЭ	96	69	66	65	69	62	66	70	64	66	63	69	69	68	64	74	50	61	72	74	64	69	69	71	70	72	64	61	69	66	63	68
ДЖЭ	93	67	59	62	63	55	59	64	60	57	57	62	65	61	56	76	61	50	69	69	60	59	61	68	65	65	56	61	65	61	60	63
РОЭ	94	79	62	66	59	68	69	59	66	67	63	72	61	72	69	78	72	69	50	66	64	61	72	66	70	65	68	71	65	62	71	65
Cr	90	70	62	67	66	65	61	65	61	63	64	67	66	66	65	79	74	69	66	50	60	66	65	68	58	59	67	63	68	62	68	59
БК	92	68	56	63	61	55	58	61	56	57	57	60	64	59	56	79	64	60	64	60	50	58	59	65	60	61	60	59	64	58	59	61
ГПЗ	95	77	55	60	58	60	62	59	60	60	57	67	57	65	61	78	69	59	61	66	58	50	65	59	63	57	59	68	63	55	67	60
МП	90	67	62	65	67	59	56	67	58	58	62	60	67	62	57	79	69	61	72	65	59	65	50	69	57	61	62	62	66	63	61	62
ПГ	93	77	60	63	60	63	69	63	59	65	64	69	61	73	65	82	71	68	66	68	65	59	69	50	63	55	68	67	67	64	72	58
ПЖ	93	72	63	65	66	62	60	64	57	59	64	63	66	63	59	81	70	65	70	58	60	63	57	63	50	55	65	61	66	64	65	57
ПК	93	76	57	63	60	62	62	60	56	61	61	66	55	69	59	83	72	65	65	59	61	57	61	55	55	50	62	64	62	59	69	51
ПС	94	70	59	63	64	56	59	63	60	60	57	65	63	62	56	78	64	56	68	67	60	59	62	68	65	62	50	61	63	57	63	62
ПУ	92	62	63	65	69	61	58	69	60	57	65	60	68	58	57	77	61	61	71	63	59	68	62	67	61	64	61	50	70	65	58	62
РМ	96	78	63	59	65	63	64	61	62	66	63	72	59	68	63	79	69	65	65	68	64	63	66	67	66	62	63	70	50	59	69	59
ХБ	95	74	56	59	58	59	61	56	57	60	56	68	56	63	58	77	66	61	62	62	58	55	63	64	64	59	57	65	59	50	66	57
ЭМ	91	61	63	68	68	60	59	68	63	58	63	59	71	61	59	79	63	60	71	68	59	67	61	72	65	69	63	58	69	66	50	66
ЯЖ	93	75	57	62	62	58	59	60	56	60	62	65	59	67	62	79	68	63	65	59	61	60	62	58	57	51	62	62	59	57	66	50

Рис. 2: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (с пересечениями классов). Матрица значений 100×AUC для 32 болезней (красный — 50, синий — 100)

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	0	74	88	89	91	85	85	90	86	81	90	77	91	85	85	97	92	86	88	80	84	90	80	86	86	86	89	83	92	89	82	87
ВД	74	0	47	53	56	38	37	53	42	32	45	26	56	28	33	69	39	35	58	40	37	53	35	53	43	51	41	23	56	49	22	50
ГБ	88	47	0	21	20	15	20	14	8	12	12	29	19	27	16	51	32	18	25	23	11	9	24	21	25	14	19	25	25	12	27	13
ЖК	89	53	21	0	34	27	29	17	16	29	8	43	20	34	24	45	29	23	32	33	26	21	30	25	31	25	25	30	18	19	36	25
ИБ	91	56	20	34	0	25	30	8	17	24	13	42	16	36	27	51	38	26	17	32	22	15	35	20	31	20	27	37	29	15	37	23
МК	85	38	15	27	25	0	14	25	17	8	17	19	26	22	10	59	25	11	35	30	10	19	18	27	24	23	13	22	26	18	21	16
ММ	85	37	20	29	30	14	0	30	16	13	20	19	32	16	12	57	31	18	38	23	17	23	11	37	20	23	19	16	27	21	19	19
СД	90	53	14	17	8	25	30	0	20	27	15	41	12	33	25	54	39	29	18	31	21	17	34	26	27	20	26	38	22	12	36	20
УЩ	86	42	8	16	17	17	16	20	0	13	12	27	22	25	12	51	28	19	32	23	12	20	15	19	14	11	20	20	25	14	25	11
ХГ	81	32	12	29	24	8	13	27	13	0	17	13	29	22	11	60	33	14	34	26	14	19	16	31	19	21	20	14	31	20	17	21
ХХ	90	45	12	8	13	17	20	15	12	17	0	34	19	21	13	45	27	13	26	29	14	13	25	27	27	22	14	29	25	12	26	24
А	77	26	29	43	42	19	19	41	27	13	34	0	43	24	23	68	39	24	45	35	19	35	20	37	26	32	30	20	44	36	19	29
АП	91	56	19	20	16	26	32	12	22	29	19	43	0	41	26	56	37	29	23	32	29	14	35	23	32	10	26	37	17	12	41	18
АХ	85	28	27	34	36	22	16	33	25	22	21	24	41	0	20	52	35	23	43	33	19	31	25	46	26	37	23	16	36	25	22	34
ЯБ	85	33	16	24	27	10	12	25	12	11	13	23	26	20	0	58	29	12	38	29	12	22	14	30	18	19	11	13	25	16	18	23
ГБК	97	69	51	45	51	59	57	54	51	60	45	68	56	52	58	0	48	52	56	58	58	56	58	65	61	66	55	55	58	53	59	58
ГДЭ	92	39	32	29	38	25	31	39	28	33	27	39	37	35	29	48	0	22	44	49	27	38	37	42	41	43	27	23	37	32	26	36
ДЖЭ	86	35	18	23	26	11	18	29	19	14	13	24	29	23	12	52	22	0	38	39	19	19	21	35	31	30	12	22	29	23	21	27
РОЭ	88	58	25	32	17	35	38	18	32	34	26	45	23	43	38	56	44	38	0	31	27	22	44	33	40	31	37	41	30	25	43	29
Cr	80	40	23	33	32	30	23	31	23	26	29	35	32	33	29	58	49	39	31	0	21	33	29	36	15	18	34	25	35	25	36	17
БК	84	37	11	26	22	10	17	21	12	14	14	19	29	19	12	58	27	19	27	21	0	15	18	30	19	23	19	18	28	16	18	21
ГПЗ	90	53	9	21	15	19	23	17	20	19	13	35	14	31	22	56	38	19	22	33	15	0	31	19	27	13	19	35	26	10	33	19
МП	80	35	24	30	35	18	11	34	15	16	25	20	35	25	14	58	37	21	44	29	18	31	0	38	14	22	23	23	31	26	22	23
ПГ	86	53	21	25	20	27	37	26	19	31	27	37	23	46	30	65	42	35	33	36	30	19	38	0	26	11	37	33	34	28	43	16
ПЖ	86	43	25	31	31	24	20	27	14	19	27	26	32	26	18	61	41	31	40	15	19	27	14	26	0	10	31	22	32	28	30	15
ПК	86	51	14	25	20	23	23	20	11	21	22	32	10	37	19	66	43	30	31	18	23	13	22	11	10	0	24	28	25	18	38	3
ПС	89	41	19	25	27	13	19	26	20	20	14	30	26	23	11	55	27	12	37	34	19	19	23	37	31	24	0	23	26	15	26	24
ПУ	83	23	25	30	37	22	16	38	20	14	29	20	37	16	13	55	23	22	41	25	18	35	23	33	22	28	23	0	40	30	16	24
РМ	92	56	25	18	29	26	27	22	25	31	25	44	17	36	25	58	37	29	30	35	28	26	31	34	32	25	26	40	0	19	38	18
ХБ	89	49	12	19	15	18	21	12	14	20	12	36	12	25	16	53	32	23	25	25	16	10	26	28	28	18	15	30	19	0	32	13
ЭМ	82	22	27	36	37	21	19	36	25	17	26	19	41	22	18	59	26	21	43	36	18	33	22	43	30	38	26	16	38	32	0	32
ЯЖ	87	50	13	25	23	16	19	20	11	21	24	29	18	34	23	58	36	27	29	17	21	19	23	16	15	3	24	24	18	13	32	0

Рис. 3: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (с пересечениями классов). Матрица значений Индекса Джини, умноженного на 100: $100 \times J = 100 \times (2AUC - 1)$ для 32 болезней (красный — 0, синий — 100)

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	216	88	53	30	71	56	62	37	50	47	46	54	35	32	33	28	41	40	22	32	46	57	41	18	20	58	42	54	29	40	26	34
ВД	88	216	99	131	138	97	34	33	61	50	75	22	109	34	60	39	19	32	38	26	42	152	29	48	17	15	99	33	10	29	24	17
ГБ	53	99	216	44	1	54	64	8	28	70	2	59	59	30	81	70	60	36	129	70	19	47	101	109	43	79	84	42	42	95	114	48
ЖК	30	131	44	216	2	113	119	11	178	105	24	59	40	85	93	74	95	34	39	104	93	63	107	91	70	146	47	133	74	19	156	165
ИБ	71	138	1	2	216	121	98	144	28	103	92	82	147	34	75	65	50	113	139	24	74	51	191	25	46	37	102	127	110	93	200	83
МК	56	97	54	113	121	216	38	62	32	23	12	36	48	31	24	62	28	103	99	99	15	46	23	85	107	30	36	27	45	16	11	71
ММ	62	34	64	119	98	38	216	78	14	62	52	45	80	16	152	110	27	86	50	130	36	89	27	40	34	49	157	19	29	35	7	43
СД	37	33	8	11	144	62	78	216	10	71	66	69	80	24	77	68	28	54	78	31	102	12	45	35	118	53	67	40	72	76	25	35
УЩ	50	61	28	178	28	32	14	10	216	18	9	32	79	53	22	75	30	29	78	98	30	60	166	55	49	62	60	24	85	71	33	58
ХГ	47	50	70	105	103	23	62	71	18	216	22	26	40	36	26	87	38	82	102	81	3	54	52	51	71	51	29	26	22	56	19	29
ХХ	46	75	2	24	92	12	52	66	9	22	216	19	104	49	52	99	37	54	142	169	56	27	149	67	53	120	43	33	121	31	174	65
А	54	22	59	59	82	36	45	69	32	26	19	216	44	21	36	47	54	22	35	25	29	148	42	54	58	36	25	32	45	20	17	14
АП	35	109	59	40	147	48	80	80	79	40	104	44	216	36	24	51	37	54	93	25	29	47	59	35	66	48	45	58	58	64	31	55
АХ	32	34	30	85	34	31	16	24	53	36	49	21	36	216	28	117	37	15	28	32	5	104	15	40	26	81	85	55	21	23	10	27
ЯБ	33	60	81	93	75	24	152	77	22	26	52	36	24	28	216	80	26	51	44	65	34	52	55	56	62	54	32	22	42	62	11	95
ГБК	28	39	70	74	65	62	110	68	75	87	99	47	51	117	80	216	43	105	71	82	87	64	172	103	81	88	77	50	72	95	44	137
ГДЭ	41	19	60	95	50	28	27	28	30	38	37	54	37	37	26	43	216	5	51	43	18	51	118	101	23	15	4	164	26	72	135	23
ДЖЭ	40	32	36	34	113	103	86	54	29	82	54	22	54	15	51	105	5	216	69	44	11	46	65	130	124	28	28	21	37	11	20	47
РОЭ	22	38	129	39	139	99	50	78	78	102	142	35	93	28	44	71	51	69	216	21	55	97	112	21	65	45	137	69	99	84	45	63
Cr	32	26	70	104	24	99	130	31	98	81	169	25	25	32	65	82	43	44	21	216	122	26	41	50	29	27	127	23	26	58	32	41
БК	46	42	19	93	74	15	36	102	30	3	56	29	29	5	34	87	18	11	55	122	216	30	34	65	102	77	17	27	53	26	16	64
ГПЗ	57	152	47	63	51	46	89	12	60	54	27	148	47	104	52	64	51	46	97	26	30	216	71	146	105	125	23	54	18	45	68	66
МП	41	29	101	107	191	23	27	45	166	52	149	42	59	15	55	172	118	65	112	41	34	71	216	34	49	73	161	14	29	46	21	35
ПГ	18	48	109	91	25	85	40	35	55	51	67	54	35	40	56	103	101	130	21	50	65	146	34	216	96	46	65	15	68	102	71	47
ПЖ	20	17	43	70	46	107	34	118	49	71	53	58	66	26	62	81	23	124	65	29	102	105	49	96	216	80	98	18	44	124	41	33
ПК	58	15	79	146	37	30	49	53	62	51	120	36	48	81	54	88	15	28	45	27	77	125	73	46	80	216	98	20	88	55	6	39
ПС	42	99	84	47	102	36	157	67	60	29	43	25	45	85	32	77	4	28	137	127	17	23	161	65	98	98	216	29	63	33	21	74
ПУ	54	33	42	133	127	27	19	40	24	26	33	32	58	55	22	50	164	21	69	23	27	54	14	15	18	20	29	216	12	31	86	24
РМ	29	10	42	74	110	45	29	72	85	22	121	45	58	21	42	72	26	37	99	26	53	18	29	68	44	88	63	12	216	52	16	97
ХБ	40	29	95	19	93	16	35	76	71	56	31	20	64	23	62	95	72	11	84	58	26	45	46	102	124	55	33	31	52	216	9	74
ЭМ	26	24	114	156	200	11	7	25	33	19	174	17	31	10	11	44	135	20	45	32	16	68	21	71	41	6	21	86	16	9	216	28
ЯЖ	34	17	48	165	83	71	43	35	58	29	65	14	55	27	95	137	23	47	63	41	64	66	35	47	33	39	74	24	97	74	28	216

Рис. 4: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (с пересечениями классов). Матрица количества значимых признаков для 32 болезней (красный — 200, синий — 1)

пациенты, болеющие 2-ой болезнью (2-ым набором болезней). Затем из каждого из двух множеств удаляются элементы, входящие в их пересечение. Далее каждая из 2-х выборок делится на 10 частей и соответствующие части разных выборок сливаются. Получается одна выборка из 10 частей, содержащая 2 класса.

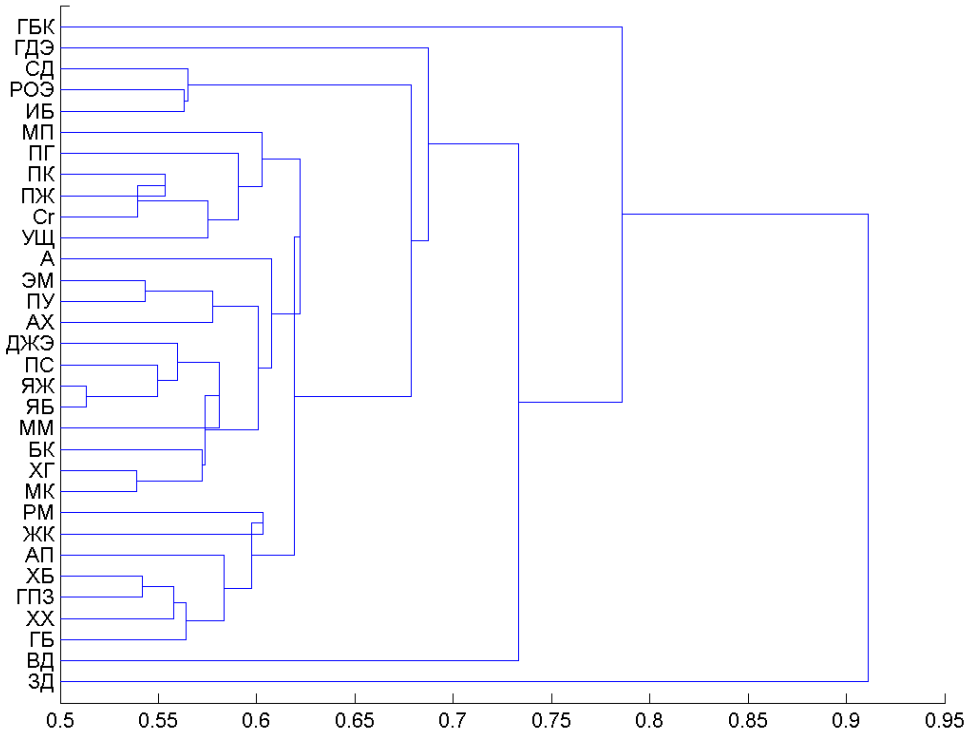


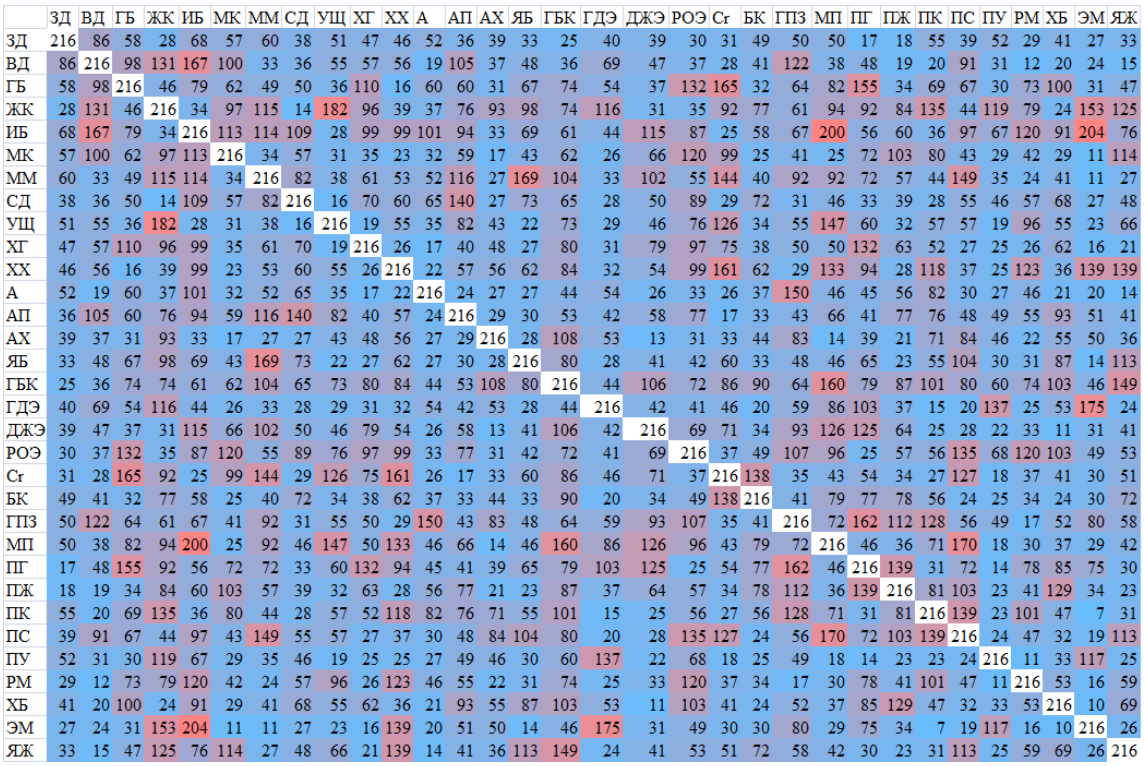
Рис. 5: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без пересечений классов). Дендрогамма для 32 болезней

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	50	87	94	94	95	93	93	95	93	91	95	88	95	92	92	98	96	93	94	91	92	95	90	93	93	92	94	92	96	95	91	93
ВД	87	50	74	77	79	70	70	78	73	67	74	63	79	66	68	85	72	71	79	71	70	77	70	81	72	76	73	62	77	75	63	75
ГБ	94	74	50	61	65	63	64	57	58	62	57	66	59	67	62	75	72	70	59	61	59	59	66	63	64	59	64	65	64	58	66	61
ЖК	94	77	61	50	63	64	67	61	58	66	59	72	61	68	63	73	65	62	66	67	65	61	67	65	67	62	63	64	60	61	69	62
ИБ	95	79	65	63	50	70	69	57	64	69	62	73	60	71	68	75	72	73	56	66	67	63	71	68	67	66	69	71	65	63	73	65
МК	93	70	63	64	70	50	59	66	60	54	62	61	67	63	56	80	66	57	68	66	57	60	61	63	63	63	58	60	64	60	62	58
ММ	93	70	64	67	69	59	50	67	61	58	62	63	66	61	58	79	64	60	70	63	60	63	60	68	62	62	60	58	63	62	63	59
СД	95	78	57	61	57	66	67	50	62	65	60	71	57	70	65	76	70	66	57	66	64	57	69	62	67	60	64	68	61	58	71	59
УЩ	93	73	58	58	64	60	61	62	50	58	56	65	62	63	57	76	66	60	65	63	58	59	60	61	57	55	60	62	60	57	64	57
ХГ	91	67	62	66	69	54	58	65	58	50	63	58	66	63	55	80	66	61	67	61	54	62	59	60	60	61	61	58	66	61	58	65
ХХ	95	74	57	59	62	62	62	60	56	63	50	70	61	64	59	73	63	62	64	64	60	56	66	64	65	64	59	64	62	56	65	65
А	88	63	66	72	73	61	63	71	65	58	70	50	74	64	62	83	69	64	74	68	61	67	61	72	64	67	66	58	72	68	63	66
АП	95	79	59	61	60	67	66	57	62	66	61	74	50	71	65	79	70	68	62	65	65	59	67	63	67	62	64	68	60	58	71	59
АХ	92	66	67	68	71	63	61	70	63	63	64	64	71	50	60	76	68	62	71	66	62	67	65	72	68	69	61	57	69	65	58	68
ЯБ	92	68	62	63	68	56	58	65	57	55	59	62	65	60	50	78	65	56	68	63	57	62	57	67	60	61	55	57	62	59	61	51
ГБК	98	85	75	73	75	80	79	76	76	80	73	83	79	76	78	50	75	76	78	79	79	79	80	81	81	82	78	76	79	77	79	79
ГДЭ	96	72	72	65	72	66	64	70	66	66	63	69	70	68	65	75	50	65	71	74	71	70	69	73	72	72	66	64	69	67	63	67
ДЖЭ	93	71	70	62	73	57	60	66	60	61	62	64	68	62	56	76	65	50	69	68	59	61	62	68	65	67	57	59	65	63	61	63
РОЭ	94	79	59	66	56	68	70	57	65	67	64	74	62	71	68	78	71	69	50	69	64	61	72	66	71	64	69	71	69	64	72	63
Cr	91	71	61	67	66	66	63	66	63	61	64	68	65	66	63	79	74	68	69	50	62	65	65	67	58	60	66	63	67	63	68	60
БК	92	70	59	65	67	57	60	64	58	54	60	61	65	62	57	79	71	59	64	62	50	60	62	65	62	61	61	59	64	59	60	59
ГПЗ	95	77	59	61	63	60	63	57	59	62	56	67	59	67	62	79	70	61	61	65	60	50	66	62	66	58	61	67	63	54	68	62
МП	90	70	66	67	71	61	60	69	60	59	66	61	67	65	57	80	69	62	72	65	62	66	50	66	57	60	62	64	66	65	64	57
ПГ	93	81	63	65	68	63	68	62	61	60	64	72	63	72	67	81	73	68	66	67	65	62	66	50	63	56	68	67	66	64	71	60
ПЖ	93	72	64	67	67	63	62	67	57	60	65	64	67	68	60	81	72	65	71	58	62	66	57	63	50	55	68	62	69	64	65	58
ПК	92	76	59	62	66	63	62	60	55	61	64	67	62	69	61	82	72	67	64	60	61	58	60	56	55	50	64	65	64	59	69	55
ПС	94	73	64	63	69	58	60	64	60	61	59	66	64	61	55	78	66	57	69	66	61	61	62	68	68	64	50	63	62	59	62	61
ПУ	92	62	65	64	71	60	58	68	62	58	64	58	68	57	57	76	64	59	71	63	59	67	64	67	62	65	63	50	68	66	54	61
РМ	96	77	64	60	65	64	63	61	60	66	62	72	60	69	62	79	69	65	69	67	64	63	66	66	69	64	62	68	50	59	70	58
ХБ	95	75	58	61	63	60	62	58	57	61	56	68	58	65	59	77	67	63	64	63	59	54	65	64	64	59	59	66	59	50	67	57
ЭМ	91	63	66	69	73	62	63	71	64	58	65	63	71	58	61	79	63	61	72	68	60	68	64	71	65	69	62	54	70	67	50	66
ЯЖ	93	75	61	62	65	58	59	59	57	65	65	66	59	68	51	79	67	63	63	60	59	62	57	60	58	55	61	61	58	57	66	50

Рис. 6: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без пересечений классов). Матрица значений 100×AUC для 32 болезней (красный — 50, синий — 100)

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	0	74	88	89	91	85	85	89	86	81	90	76	90	84	85	97	92	86	87	81	84	90	81	86	85	84	89	84	91	89	82	86
ВД	74	0	49	53	58	41	39	55	46	35	48	27	57	32	36	70	44	42	57	43	39	53	39	61	44	51	46	24	54	50	25	49
ГБ	88	49	0	23	31	27	28	15	16	25	14	32	18	33	23	51	43	41	18	23	18	18	32	27	28	19	28	31	27	15	31	22
ЖК	89	53	23	0	25	28	33	22	16	32	18	44	23	36	25	46	29	25	32	34	29	23	34	30	34	24	26	28	21	22	38	24
ИБ	91	58	31	25	0	40	38	14	29	39	24	47	20	43	36	49	43	46	13	32	35	25	42	35	34	33	37	42	30	27	45	30
МК	85	41	27	28	40	0	18	32	21	8	23	23	34	26	12	60	33	13	36	32	14	19	22	27	26	26	17	20	29	20	25	15
ММ	85	39	28	33	38	18	0	34	21	17	24	27	32	21	16	58	28	21	40	27	21	26	19	36	24	25	20	16	26	23	25	18
СД	89	55	15	22	14	32	34	0	23	30	20	42	14	40	30	52	40	32	15	31	28	14	37	24	34	20	29	37	22	15	41	18
УЩ	86	46	16	16	29	21	21	23	0	16	12	30	24	26	13	52	32	20	31	25	15	18	20	22	15	11	21	24	21	14	29	14
ХГ	81	35	25	32	39	8	17	30	16	0	27	16	33	26	11	59	33	21	34	22	8	24	17	20	21	22	22	17	32	22	17	29
ХХ	90	48	14	18	24	23	24	20	12	27	0	40	22	28	19	46	27	24	28	29	20	12	32	27	31	27	18	28	25	12	30	29
А	76	27	32	44	47	23	27	42	30	16	40	0	47	27	25	66	39	29	48	37	21	34	22	44	28	33	31	17	44	36	25	32
АП	90	57	18	23	20	34	32	14	24	33	22	47	0	42	30	57	41	35	24	30	30	19	35	26	33	23	27	36	20	17	42	18
АХ	84	32	33	36	43	26	21	40	26	26	28	27	42	0	20	52	35	25	42	32	24	34	30	45	36	38	23	14	38	30	17	35
ЯБ	85	36	23	25	36	12	16	30	13	11	19	25	30	20	0	57	31	13	37	26	14	24	15	34	20	23	11	13	25	18	21	3
ГБК	97	70	51	46	49	60	58	52	52	59	46	66	57	52	57	0	49	52	57	58	57	58	59	63	61	64	56	52	57	53	59	59
ГДЭ	92	44	43	29	43	33	28	40	32	33	27	39	41	35	31	49	0	31	43	48	42	40	38	45	43	44	32	27	38	34	26	35
ДЖЭ	86	42	41	25	46	13	21	32	20	21	24	29	35	25	13	52	31	0	38	36	18	23	24	35	29	35	13	18	29	25	22	26
РОЭ	87	57	18	32	13	36	40	15	31	34	28	48	24	42	37	57	43	38	0	38	28	22	45	32	41	28	38	42	39	28	44	27
Cr	81	43	23	34	32	32	27	31	25	22	29	37	30	32	26	58	48	36	38	0	24	30	31	34	16	20	33	27	33	26	36	19
БК	84	39	18	29	35	14	21	28	15	8	20	21	30	24	14	57	42	18	28	24	0	19	23	31	25	23	23	19	28	19	20	18
ГПЗ	90	53	18	23	25	19	26	14	18	24	12	34	19	34	24	58	40	23	22	30	19	0	33	24	32	15	23	33	26	8	36	24
МП	81	39	32	34	42	22	19	37	20	17	32	22	35	30	15	59	38	24	45	31	23	33	0	32	14	20	25	27	32	30	28	14
ПГ	86	61	27	30	35	27	36	24	22	20	27	44	26	45	34	63	45	35	32	34	31	24	32	0	26	13	36	34	32	28	42	20
ПЖ	85	44	28	34	34	26	24	34	15	21	31	28	33	36	20	61	43	29	41	16	25	32	14	26	0	11	35	25	38	28	31	16
ПК	84	51	19	24	33	26	25	20	11	22	27	33	23	38	23	64	44	35	28	20	23	15	20	13	11	0	28	29	27	18	38	10
ПС	89	46	28	26	37	17	20	29	21	22	18	31	27	23	11	56	32	13	38	33	23	23	25	36	35	28	0	25	24	18	24	23
ПУ	84	24	31	28	42	20	16	37	24	17	28	17	36	14	13	52	27	18	42	27	19	33	27	34	25	29	25	0	37	31	9	22
РМ	91	54	27	21	30	29	26	22	21	32	25	44	20	38	25	57	38	29	39	33	28	26	32	32	38	27	24	37	0	18	41	15
ХБ	89	50	15	22	27	20	23	15	14	22	12	36	17	30	18	53	34	25	28	26	19	8	30	28	28	18	18	31	18	0	35	14
ЭМ	82	25	31	38	45	25	25	41	29	17	30	25	42	17	21	59	26	22	44	36	20	36	28	42	31	38	24	9	41	35	0	32
ЯЖ	86	49	22	24	30	15	18	18	14	29	29	32	18	35	3	59	35	26	27	19	18	24	14	20	16	10	23	22	15	14	32	0

Рис. 7: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без пересечений классов). Матрица значений Индекса Джини, умноженного на 100: $100 \times J = 100 \times (2AUC - 1)$ для 32 болезней (красный — 0, синий — 100)



	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	216	86	58	28	68	57	60	38	51	47	46	52	36	39	33	25	40	39	30	31	49	50	50	17	18	55	39	52	29	41	27	33
ВД	86	216	98	131	167	100	33	36	55	57	56	19	105	37	48	36	69	47	37	28	41	122	38	48	19	20	91	31	12	20	24	15
ГБ	58	98	216	46	79	62	49	50	36	110	16	60	60	31	67	74	54	37	132	165	32	64	82	155	34	69	67	30	73	100	31	47
ЖК	28	131	46	216	34	97	115	14	182	96	39	37	76	93	98	74	116	31	35	92	77	61	94	92	84	135	44	119	79	24	153	125
ИБ	68	167	79	34	216	113	114	109	28	99	99	101	94	33	69	61	44	115	87	25	58	67	200	56	60	36	97	67	120	91	204	76
МК	57	100	62	97	113	216	34	57	31	35	23	32	59	17	43	62	26	66	120	99	25	41	25	72	103	80	43	29	42	29	11	114
ММ	60	33	49	115	114	34	216	82	38	61	53	52	116	27	169	104	33	102	55	144	40	92	92	72	57	44	149	35	24	41	11	27
СД	38	36	50	14	109	57	82	216	16	70	60	65	140	27	73	65	28	50	89	29	72	31	46	33	39	28	55	46	57	68	27	48
УЩ	51	55	36	182	28	31	38	16	216	19	55	35	82	43	22	73	29	46	76	126	34	55	147	60	32	57	57	19	96	55	23	66
ХГ	47	57	110	96	99	35	61	70	19	216	26	17	40	48	27	80	31	79	97	75	38	50	50	132	63	52	27	25	26	62	16	21
ХХ	46	56	16	39	99	23	53	60	55	26	216	22	57	56	62	84	32	54	99	161	62	29	133	94	28	118	37	25	123	36	139	139
А	52	19	60	37	101	32	52	65	35	17	22	216	24	27	27	44	54	26	33	26	37	150	46	45	56	82	30	27	46	21	20	14
АП	36	105	60	76	94	59	116	140	82	40	57	24	216	29	30	53	42	58	77	17	33	43	66	41	77	76	48	49	55	93	51	41
АХ	39	37	31	93	33	17	27	27	43	48	56	27	29	216	28	108	53	13	31	33	44	83	14	39	21	71	84	46	22	55	50	36
ЯБ	33	48	67	98	69	43	169	73	22	27	62	27	30	28	216	80	28	41	42	60	33	48	46	65	23	55	104	30	31	87	14	113
ГБК	25	36	74	74	61	62	104	65	73	80	84	44	53	108	80	216	44	106	72	86	90	64	160	79	87	101	80	60	74	103	46	149
ГДЭ	40	69	54	116	44	26	33	28	29	31	32	54	42	53	28	44	216	42	41	46	20	59	86	103	37	15	20	137	25	53	175	24
ДЖЭ	39	47	37	31	115	66	102	50	46	79	54	26	58	13	41	106	42	216	69	71	34	93	126	125	64	25	28	22	33	11	31	41
РОЭ	30	37	132	35	87	120	55	89	76	97	99	33	77	31	42	72	41	69	216	37	49	107	96	25	57	56	135	68	120	103	49	53
Cr	31	28	165	92	25	99	144	29	126	75	161	26	17	33	60	86	46	71	37	216	138	35	43	54	34	27	127	18	37	41	30	51
БК	49	41	32	77	58	25	40	72	34	38	62	37	33	44	33	90	20	34	49	138	216	41	79	77	78	56	24	25	34	24	30	72
ГПЗ	50	122	64	61	67	41	92	31	55	50	29	150	43	83	48	64	59	93	107	35	41	216	72	162	112	128	56	49	17	52	80	58
МП	50	38	82	94	200	25	92	46	147	50	133	46	66	14	46	160	86	126	96	43	79	72	216	46	36	71	170	18	30	37	29	42
ПГ	17	48	155	92	56	72	72	33	60	132	94	45	41	39	65	79	103	125	25	54	77	162	46	216	139	31	72	14	78	85	75	30
ПЖ	18	19	34	84	60	103	57	39	32	63	28	56	77	21	23	87	37	64	57	34	78	112	36	139	216	81	103	23	41	129	34	23
ПК	55	20	69	135	36	80	44	28	57	52	118	82	76	71	55	101	15	25	56	27	56	128	71	31	81	216	139	23	101	47	7	31
ПС	39	91	67	44	97	43	149	55	57	27	37	30	48	84	104	80	20	28	135	127	24	56	170	72	103	139	216	24	47	32	19	113
ПУ	52	31	30	119	67	29	35	46	19	25	25	27	49	46	30	60	137	22	68	18	25	49	18	14	23	23	24	216	11	33	117	25
РМ	29	12	73	79	120	42	24	57	96	26	123	46	55	22	31	74	25	33	120	37	34	17	30	78	41	101	47	11	216	53	16	59
ХБ	41	20	100	24	91	29	41	68	55	62	36	21	93	55	87	103	53	11	103	41	24	52	37	85	129	47	32	33	53	216	10	69
ЭМ	27	24	31	153	204	11	11	27	23	16	139	20	51	50	14	46	175	31	49	30	30	80	29	75	34	7	19	117	16	10	216	26
ЯЖ	33	15	47	125	76	114	27	48	66	21	139	14	41	36	113	149	24	41	53	51	72	58	42	30	23	31	113	25	59	69	26	216

Рис. 8: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без пересечений классов). Матрица количества значимых признаков для 32 болезней (красный — 200, синий — 1)

4 Кластеризация болезней с использованием расстояния между наборами значимых признаков в качестве расстояния между кластерами

Схема подсчета AUC и усреднения весов признаков и количества значимых признаков точно такая же,как при кластеризации болезней с использованием AUC в качестве расстояния между кластерами.

Рассматриваются только пары "здоровые — больные какой-то болезнью (или набором болезней)". Так как здоровые не болеют ни одной болезнью, то проблема пересечения классов не стоит. После проведения такой классификации у каждой болезни есть значимые признаки и вектор весов признаков. Расстояние между кластерами определяется через веса общих значимых признаков:

$$f(d_1, d_2) = \frac{1}{2n} \sum_{k \in T_1 \bigcap T_2} |\frac{w_{1k}}{w_{1max}} - \frac{w_{2k}}{w_{2max}}|, 0 \leq f(d_1, d_2) \leq 1$$

где w_{jk} — вес k-ого признака, являющегося значимым для обоих классов, для для j-ого класса; $w_{jmax} = \max_{k=\overline{1,216}} |w_{jk}|$ — модуль наибольшего (из всех) по модулю веса признака для j-ого класса; d_j — номер болезни (набор номеров болезней), относящейся к для j-ого классу; $n = |T_1 \bigcap T_2|$ — количество общих значимых признаков; T_j — множество значимых признаков j-ого класса.

5 Весенний семестр 5-ого курса

Задача: уменьшить количество триграмм в эталоне болезни.

В предыдущих экспериментах было построено по 31 бинарному классификатору для каждой болезни, который отличает эту болезнь от фиксированной другой. Идея состояла в том, чтобы из этого 31 эталнона (эталон — множество значимых признаков классификатора) сделать один, который бы отличал болезнь от всех остальных, а потом его редуцировать. Так как пересечение всех эталонов для одной болезни почти у всех пустое, то сделать из 31 эталона один вряд ли получится. Поэтому была проведена кластеризация (построены дендрограммы) 31 эталона для

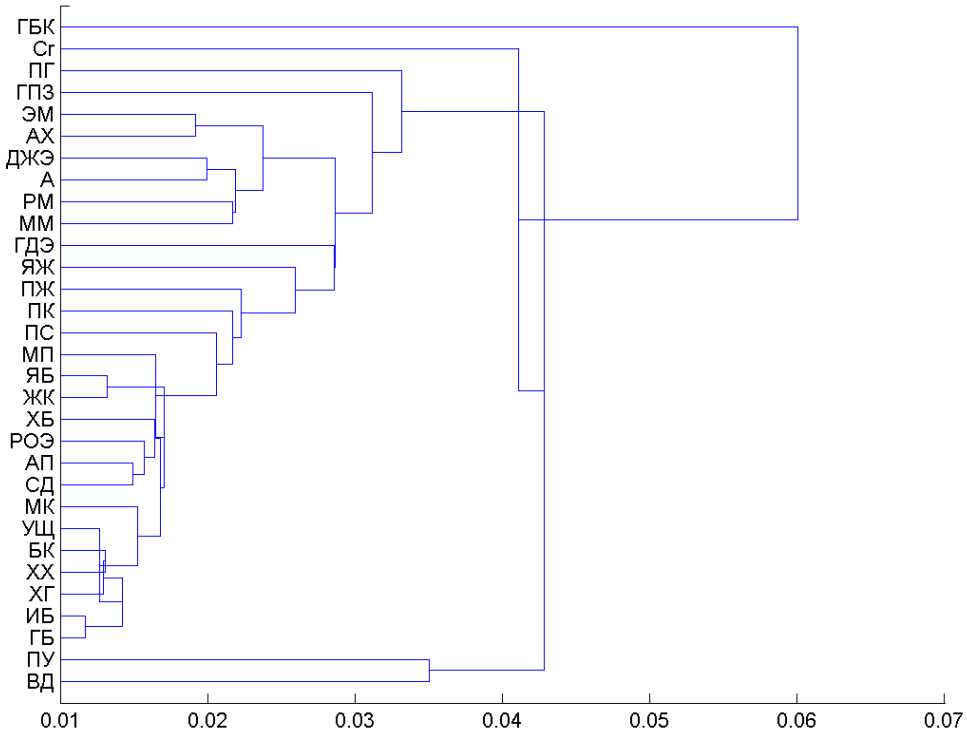


Рис. 9: Кластеризация болезней с использованием расстояния между наборами значимых признаков в качестве расстояния между кластерами. Дендрогамма для 32 болезней

	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ВД	0,0	4,8	4,6	4,8	4,9	5,1	4,0	4,0	4,6	4,3	5,5	4,8	3,6	3,7	7,1	4,6	4,8	4,4	5,4	4,9	5,7	4,3	5,1	4,3	5,5	4,5	3,5	5,8	4,9	2,8	4,6
ГБ	4,8	0,0	2,4	1,2	1,6	3,3	1,9	1,5	1,6	1,8	2,8	2,7	4,2	2,3	6,3	4,1	2,4	2,3	5,1	1,9	1,8	2,9	3,1	2,6	2,4	2,1	4,7	4,8	2,4	3,6	3,0
ЖК	4,6	2,4	0,0	2,8	2,3	3,9	2,5	1,7	2,2	2,3	3,2	3,5	4,9	1,3	5,4	3,5	2,5	3,4	4,7	2,7	3,3	2,2	2,5	3,1	2,8	2,2	3,7	4,7	3,2	4,3	2,8
ИБ	4,8	1,2	2,8	0,0	1,9	2,9	1,4	1,9	1,8	1,9	2,9	2,1	4,1	2,6	7,1	4,3	2,9	1,5	4,6	1,5	2,5	2,9	3,4	2,5	2,8	2,7	4,0	4,3	2,0	3,2	3,1
МК	4,9	1,6	2,3	1,9	0,0	2,9	2,3	2,0	1,5	1,9	2,3	2,7	4,1	2,3	6,2	4,3	2,1	3,0	5,4	1,8	2,0	2,8	2,8	3,5	2,8	1,7	4,8	3,9	1,8	3,4	3,6
ММ	5,1	3,3	3,9	2,9	2,9	0,0	2,6	3,4	2,7	2,3	2,1	2,5	3,2	3,5	7,8	4,4	3,0	3,5	2,9	2,2	3,4	2,9	5,0	2,7	4,4	3,2	3,2	2,2	1,9	2,6	3,8
СД	4,0	1,9	2,5	1,4	2,3	2,6	0,0	2,5	1,7	2,0	2,9	1,5	3,6	2,2	6,4	3,5	2,9	1,7	3,9	1,9	3,1	2,2	3,3	2,7	3,0	2,3	3,0	3,5	1,8	2,7	2,7
УЩ	4,0	1,5	1,7	1,9	2,0	3,4	2,5	0,0	1,7	2,0	2,7	3,2	4,0	1,4	5,7	3,6	2,5	3,0	5,2	2,3	2,8	2,2	3,3	2,3	2,3	2,1	4,2	4,9	2,9	3,7	2,8
ХГ	4,6	1,6	2,2	1,8	1,5	2,7	1,7	1,7	0,0	1,5	2,0	2,2	3,8	2,1	6,3	3,4	2,3	2,4	4,5	1,5	2,6	2,0	3,9	2,8	3,0	1,7	3,8	3,7	1,7	2,8	3,3
ХХ	4,3	1,8	2,3	1,9	1,9	2,3	2,0	2,0	1,5	0,0	2,0	2,6	2,7	2,4	6,3	3,2	1,7	2,5	4,2	1,3	2,9	2,0	3,9	2,9	3,7	1,9	3,5	3,3	1,7	2,2	3,5
А	5,5	2,8	3,2	2,9	2,3	2,1	2,9	2,7	2,0	2,0	0,0	3,2	3,0	3,5	7,0	4,2	2,0	3,5	3,9	2,0	2,9	2,7	4,7	3,3	4,3	2,7	4,2	2,9	2,1	2,7	4,1
АП	4,8	2,7	3,5	2,1	2,7	2,5	1,5	3,2	2,2	2,6	3,2	0,0	3,6	3,0	7,1	3,7	3,2	1,7	3,4	2,4	3,8	2,2	4,5	2,3	3,7	2,5	3,4	2,8	1,5	2,8	3,3
АХ	3,6	4,2	4,9	4,1	4,1	3,2	3,6	4,0	3,8	2,7	3,0	3,6	0,0	4,1	7,0	2,9	2,9	3,2	3,4	3,2	5,0	3,4	5,1	4,0	6,1	4,0	4,0	3,7	3,1	1,9	4,4
ЯБ	3,7	2,3	1,3	2,6	2,3	3,5	2,2	1,4	2,1	2,4	3,5	3,0	4,1	0,0	4,9	3,2	2,8	3,2	4,7	2,8	3,6	2,0	2,9	2,6	2,6	1,9	3,6	5,0	2,9	4,1	2,4
ГБК	7,1	6,3	5,4	7,1	6,2	7,8	6,4	5,7	6,3	6,3	7,0	7,1	7,0	4,9	0,0	6,0	6,1	7,4	9,0	6,9	6,5	6,1	6,3	7,6	6,9	5,8	7,0	8,0	6,7	7,0	6,9
ГДЭ	4,6	4,1	3,5	4,3	4,3	4,4	3,5	3,6	3,4	3,2	4,2	3,7	2,9	3,2	6,0	0,0	3,0	3,4	4,3	3,7	4,8	2,9	3,5	3,8	5,0	3,6	3,9	4,9	3,7	2,5	4,1
ДЖЭ	4,8	2,4	2,5	2,9	2,1	3,0	2,9	2,5	2,3	1,7	2,0	3,2	2,9	2,8	6,1	3,0	0,0	3,1	4,7	2,1	2,8	2,6	3,4	3,9	4,1	2,1	4,1	3,3	2,5	2,2	3,9
РОЭ	4,4	2,3	3,4	1,5	3,0	3,5	1,7	3,0	2,4	2,5	3,5	1,7	3,2	3,2	7,4	3,4	3,1	0,0	3,9	2,4	3,4	2,9	4,1	3,1	4,0	2,9	3,7	3,5	2,2	2,6	3,1
Cr	5,4	5,1	4,7	4,6	5,4	2,9	3,9	5,2	4,5	4,2	3,9	3,4	3,4	4,7	9,0	4,3	4,7	3,9	0,0	4,0	5,5	4,0	6,6	2,6	5,3	5,3	4,1	3,7	3,7	2,9	4,1
БК	4,9	1,9	2,7	1,5	1,8	2,2	1,9	2,3	1,5	1,3	2,0	2,4	3,2	2,8	6,9	3,7	2,1	2,4	4,0	0,0	2,6	2,5	4,1	2,9	3,6	2,5	4,0	3,2	1,5	2,5	3,6
ГПЗ	5,7	1,8	3,3	2,5	2,0	3,4	3,1	2,8	2,6	2,9	2,9	3,8	5,0	3,6	6,5	4,8	2,8	3,4	5,5	2,6	0,0	4,2	3,5	3,4	2,8	2,8	5,7	4,6	3,0	4,2	3,9
МП	4,3	2,9	2,2	2,9	2,8	2,9	2,2	2,2	2,0	2,0	2,7	2,2	3,4	2,0	6,1	2,9	2,6	2,9	4,0	2,5	4,2	0,0	4,0	2,2	3,4	2,4	2,8	3,7	2,6	3,0	2,9
ПГ	5,1	3,1	2,5	3,4	2,8	5,0	3,3	3,3	3,9	3,9	4,7	4,5	5,1	2,9	6,3	3,5	3,4	4,1	6,6	4,1	3,5	4,0	0,0	4,4	2,8	3,2	4,6	6,6	4,2	5,3	3,2
ПЖ	4,3	2,6	3,1	2,5	3,5	2,7	2,7	2,3	2,8	2,9	3,3	2,3	4,0	2,6	7,6	3,8	3,9	3,1	2,6	2,9	3,4	2,2	4,4	0,0	3,2	3,7	3,6	3,9	2,7	3,6	2,8
ПК	5,5	2,4	2,8	2,8	2,8	4,4	3,0	2,3	3,0	3,7	4,3	3,7	6,1	2,6	6,9	5,0	4,1	4,0	5,3	3,6	2,8	3,4	2,8	3,2	0,0	3,3	5,2	6,2	3,7	5,9	2,3
ПС	4,5	2,1	2,2	2,7	1,7	3,2	2,3	2,1	1,7	1,9	2,7	2,5	4,0	1,9	5,8	3,6	2,1	2,9	5,3	2,5	2,8	2,4	3,2	3,7	3,3	0,0	3,7	3,8	2,2	3,4	3,2
ПУ	3,5	4,7	3,7	4,0	4,8	3,2	3,0	4,2	3,8	3,5	4,2	3,4	4,0	3,6	7,0	3,9	4,1	3,7	4,1	4,0	5,7	2,8	4,6	3,6	5,2	3,7	0,0	4,2	4,0	3,0	3,6
РМ	5,8	4,8	4,7	4,3	3,9	2,2	3,5	4,9	3,7	3,3	2,9	2,8	3,7	5,0	8,0	4,9	3,3	3,5	3,7	3,2	4,6	3,7	6,6	3,9	6,2	3,8	4,2	0,0	2,8	2,7	5,2
ХБ	4,9	2,4	3,2	2,0	1,8	1,9	1,8	2,9	1,7	1,7	2,1	1,5	3,1	2,9	6,7	3,7	2,5	2,2	3,7	1,5	3,0	2,6	4,2	2,7	3,7	2,2	4,0	2,8	0,0	2,5	3,6
ЭМ	2,8	3,6	4,3	3,2	3,4	2,6	2,7	3,7	2,8	2,2	2,7	2,8	1,9	4,1	7,0	2,5	2,2	2,6	2,9	2,5	4,2	3,0	5,3	3,6	5,9	3,4	3,0	2,7	2,5	0,0	4,6
ЯЖ	4,6	3,0	2,8	3,1	3,6	3,8	2,7	2,8	3,3	3,5	4,1	3,3	4,4	2,4	6,9	4,1	3,9	3,1	4,1	3,6	3,9	2,9	3,2	2,8	2,3	3,2	3,6	5,2	3,6	4,6	0,0

Рис. 10: Кластеризация болезней с использованием расстояния между наборами значимых признаков в качестве расстояния между кластерами. Матрица значений функции расстояния между болезнями, умноженной на 100: $100 \times f(d_1, d_2)$ (красный — 0, синий — 10)

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
size of union	105	190	216	216	216	216	215	216	216	210	216	201	211	203	216	212	214	214	210	210	214	216	215	214	209	215	216	213	210	215	215	214
size of intersection	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Рис. 11: Мощности пересечений и объединений всех эталонов (31 эталона) для каждой болезни

каждой болезни. В качестве расстояния между кластерами были использованы 2 функции:

$$f_1(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}; f_2(x,y) = 1 - \max\left(\frac{|x \cap y|}{|x|}, \frac{|x \cap y|}{|y|}\right). \quad (1)$$

Результаты для 1-ой функции – в папке "1 для 2-ой – в папке "2".

Кроме этого, для каждой пары болезней i и j для каждой триграммы k было вычислено, с какой частотой триграмма k встречается в эталонах, отличающих болезнь I от болезни J, где $I \neq i$, и $J \neq j$. Результат – в файле Эталоны.xlsx, лист matFrequency.

Применение классификатора "здоровье против болезни i"к случаю "болезнь j против болезни i"показало, что эталон, отличающий болезнь от здоровья, не отличает болезнь от болезни.

6 6-ой курс

X — матрица $N \times M$ объект-признак. Объекты — пациенты, характеризующиеся кодограммой кардиограммы. Признаки бинарные: $x_{ik} = 1$ — в кодограмме i-ого пациента есть триграмма k (повторяется 2 и более раз), $x_{ik} = 0$ — нет. Классы — болезни (включая абсолютное здоровье). Y — матрица $N \times L$ принадлежности к классу, $y_{ij} = 1$, если объект i принадлежит к классу j, и $y_{ij} = 0$, если нет. Пусть $i = \overline{1,N}$ — идентификатор объекта (пациента), $j = \overline{1,L}$ — идентификатор класса (болезни) и $k = \overline{1,M}, M = 216$ — идентификатор признака (триграммы).

Применяется схема кросс-валидации 10-fold CV. AUC и количество значимых признаков усредняется по итерациям кросс-валидации путем нахождения среднего арифметического (с последующим округлением до ближайшего целого для количества значимых признаков).

6.1 Другие методы сортировки признаков

Ранее признаки сортировались по модулям весов (чем больше модуль веса, тем более информативен признак). Были применены другие методы сортировки признаков (обозначения из файла 'Таблицы по данным.xlsx'):

- 'Feature Selection (min(abs(diff(frequency))))' = 'Feature Selection (min)',

	остальные болезни																														
ВД	33	30	28	38	37	29	35	41	30	47	28	38	37	23	34	36	29	42	37	32	41	38	39	36	33	41	29	31	41	36
ГБ	71	47	44	56	56	45	53	58	49	65	44	59	55	41	56	55	43	57	55	49	59	53	56	52	52	62	46	48	62	52
ЖК	74	55	49	61	61	50	56	63	53	69	49	62	59	43	58	58	48	62	60	54	63	58	61	57	56	66	51	53	66	57
ИБ	76	56	52	62	62	50	58	64	54	71	50	65	61	46	62	61	48	62	61	55	64	57	62	58	59	68	52	54	68	58
МК	66	44	40	39	50	39	46	52	42	59	38	53	49	36	49	48	38	52	49	42	52	46	50	45	45	57	39	42	56	47
ММ	66	45	42	39	51	39	47	53	43	60	38	53	50	36	51	50	39	52	50	44	53	48	51	47	47	57	40	42	56	47
СД	75	56	54	50	62	62	59	64	55	70	50	65	61	48	62	61	48	62	61	55	64	58	62	58	59	68	51	54	68	58
УЩ	70	48	44	42	55	54	43	56	47	63	42	57	53	39	53	53	42	55	53	47	56	50	54	49	50	60	43	46	60	50
ХГ	63	42	40	37	48	47	37	44	40	57	36	50	47	34	47	46	36	50	47	41	50	45	48	44	43	53	37	40	53	44
ХХ	72	52	49	46	58	58	47	55	60	67	46	60	57	41	57	56	45	60	57	51	61	56	59	55	54	65	47	50	63	55
А	57	36	33	30	41	41	31	38	44	34	30	44	40	28	40	39	31	44	41	34	44	39	41	37	37	47	31	33	47	39
АП	76	57	54	51	63	63	51	59	65	55	72	66	62	48	62	61	49	63	62	56	65	59	63	59	59	69	52	54	68	58
АХ	63	45	41	39	51	49	40	47	53	42	59	40	49	32	47	48	40	53	49	44	53	50	51	48	46	55	41	42	54	48
ЯБ	66	46	43	40	52	51	40	48	54	44	61	39	54	35	49	49	40	53	51	45	54	50	52	48	47	57	40	43	57	48
ГБК	83	71	66	66	76	75	67	72	77	68	82	69	74	75	69	72	65	77	75	72	77	76	78	77	73	78	68	70	77	73
ГДЭ	70	51	46	45	57	57	46	53	59	48	66	46	57	55	35	53	45	60	56	51	59	56	59	56	51	61	46	49	60	55
ДЖЭ	68	48	44	42	54	53	43	49	56	45	62	43	55	52	37	50	42	56	53	46	56	51	54	50	48	60	43	46	59	52
РОЭ	76	58	56	53	64	64	53	61	65	57	71	53	66	63	50	64	63	64	63	57	66	61	64	61	61	69	55	56	68	60
Cr	66	46	44	41	52	51	41	48	54	45	60	40	54	51	40	53	52	41	51	45	53	47	50	47	49	56	42	44	57	47
БК	66	45	43	39	51	51	40	48	53	44	60	39	53	50	37	51	50	38	52	44	54	48	51	47	47	57	41	43	57	47
ГПЗ	75	52	49	47	59	59	46	55	60	51	67	46	62	58	46	60	57	44	59	57	61	54	58	53	55	66	48	50	65	54
МП	65	44	41	38	50	49	38	45	51	42	58	37	52	48	37	50	48	38	50	49	42	45	48	44	46	54	38	41	55	45
ПГ	74	51	48	45	58	59	46	53	59	50	66	45	62	57	44	59	57	44	57	57	50	60	57	52	54	64	48	50	64	54
ПЖ	67	45	42	39	52	52	41	47	53	43	61	40	53	50	36	51	50	40	53	51	44	54	47	47	47	58	41	44	57	49
ПК	72	50	46	44	56	56	44	52	58	49	65	43	60	55	44	58	55	43	56	55	48	58	51	55	53	63	45	48	63	52
ПС	70	50	46	44	56	55	45	52	58	48	65	44	57	54	40	53	53	44	58	55	48	58	54	57	52	62	45	47	61	53
ПУ	60	41	38	35	47	45	35	43	49	38	56	34	48	45	31	44	44	35	49	46	40	49	45	47	43	42	35	38	51	42
РМ	75	56	53	50	62	62	50	58	64	54	71	50	64	61	46	60	60	49	63	61	55	64	59	62	59	58	67	54	67	58
ХБ	72	53	50	47	59	58	47	55	61	51	67	47	61	57	44	58	57	46	59	57	52	61	56	59	55	54	64	48	64	55
ЭМ	60	41	38	35	47	46	36	43	49	38	56	36	47	45	28	43	44	36	50	46	40	50	46	47	45	42	51	37	39	45
ЯЖ	70	50	48	44	56	55	44	52	58	49	64	43	58	55	42	56	55	43	56	55	48	58	51	55	51	52	61	45	47	61

Рис. 12: $100 \times AUC$. Результат применения классификатора "здоровье – болезнь i"к случаю "болезнь j – болезнь i".

every&all	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC	0,91	0,74	0,81	0,85	0,81	0,83	0,82	0,82	0,79	0,82	0,85	0,83	0,83	0,89	0,80	0,79	0,87	0,91	0,79	0,66	0,83	0,90	0,82	0,87	0,85	0,82	0,83	0,84	0,75	0,85	0,81	0,83
K	38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	74	1	1	1	29	1	1	1	1	1	1	1	1	1	1	1	1
size of intersection	0																															
size of union	118																															

Рис. 13: Результат работы классификатора "болезнь – остальные болезни"@

- 'Feature Selection (average(abs(diff(frequency))))' = 'Feature Selection (average)',
- 'Feature Selection (min(abs(diff(ln(frequency)))))'.

Для выполнения такой сортировки признаков нужно:

1. Вычислить по всей выборке $frequency = F$ — матрица $L \times M$,

$$F_{jk} = \frac{\sum_{i=1}^{N}[y_{ij}] \cdot [x_{ik}]}{\sum_{i=1}^{N}[y_{ij}]}$$

   — отношение количества пациентов, страдающих j-ой болезнью и имеющих в кодограмме k-ую триграмму, к общему количеству пациентов, страдающих j-ой болезнью. Назовем эту величину частотой триграммы k у болезни j.

   Для 'Feature Selection (min(abs(diff(ln(frequency)))))' нужно еще поэлементно вычислить натуральный логарифм от F: $\ln(F) = \| \ln F_{jk} \|$.

2. Выполнить операцию 'diff':

$$diff(F) = D = \|D_{j_1 j_2 k}\|, D_{j_1 j_2 k} = F_{j_1 k} - F_{j_2 k},$$

   то есть найти все попарные разности частоты каждой триграммы у разных болезней.

3. Найти поэлементный модуль трехмерной матрицы D: $abs(D) = \| |D_{j_1 j_2 k}| \|$.

4. 'Feature Selection (min)' — для j-ой болезни найти для каждой триграммы минимум $abs(D)$ по остальным болезням:

$$\mathbf{R}_j = \|R_{jk}\|, R_{jk} = \min_{j' \neq j} |D_{jj'k}|;$$

   'Feature Selection (average)' — для j-ой болезни найти для каждой триграммы среднее арифметическое $abs(D)$ по остальным болезням (кроме абсолютного здоровья):

$$\mathbf{R}_j = \|R_{jk}\|, R_{jk} = \frac{1}{L-2} \sum_{j' \neq j, j' \neq 1} |D_{jj'k}|.$$

Таким образом вычисляется матрица $R_{L\times M}$, каждая ее строка — вектор значений, характеризующий значимость признака для данной болезни. Чтобы применить такую сортировку в двухклассовом байесовском классификаторе, нужно получить характеристику значимости признаков для пары болезней: $R_{j_1j_2k} = \max\{R_{j_1k}, R_{j_2k}\}$. Признаки сортируются по убыванию $R_{j_1j_2k}$

При вычислении весов признаков во время кросс-валидации, то есть по той выборке, по которой происходило обучение на данной итерации, такая сортировка ('Feature Selection (min)' и 'Feature Selection (average)') весов признаков увеличивает AUC по сравнению с сортировкой по модулям весов в среднем по парам болезней (без удаления пересечений классов) на $0,015$, 'Feature Selection (min(abs(diff(ln(frequency)))))' — на $0,013$. Однако, вычисление весов по всей выборке до кросс-валидации повышает AUC в среднем по парам болезней (без удаления пересечений классов) на $0,027$, то есть увеличение почти в 2 раза больше.

Выше указывались результаты для схемы обучения и тестирования 'каждый против каждого'. Сортировка 'Feature Selection (min)' использовалась в обучении 'каждый против всех' (результат на рис. 14) с последующим тестом полученных классификаторов по схеме 'каждый против каждого' (результат на рис. 15).

Feature Selection (min(abs(diff(frequency)))), train: every & all, test: every & every (тест на схеме every & all), веса признаков посчитаны для каждой части выборки индивидуально

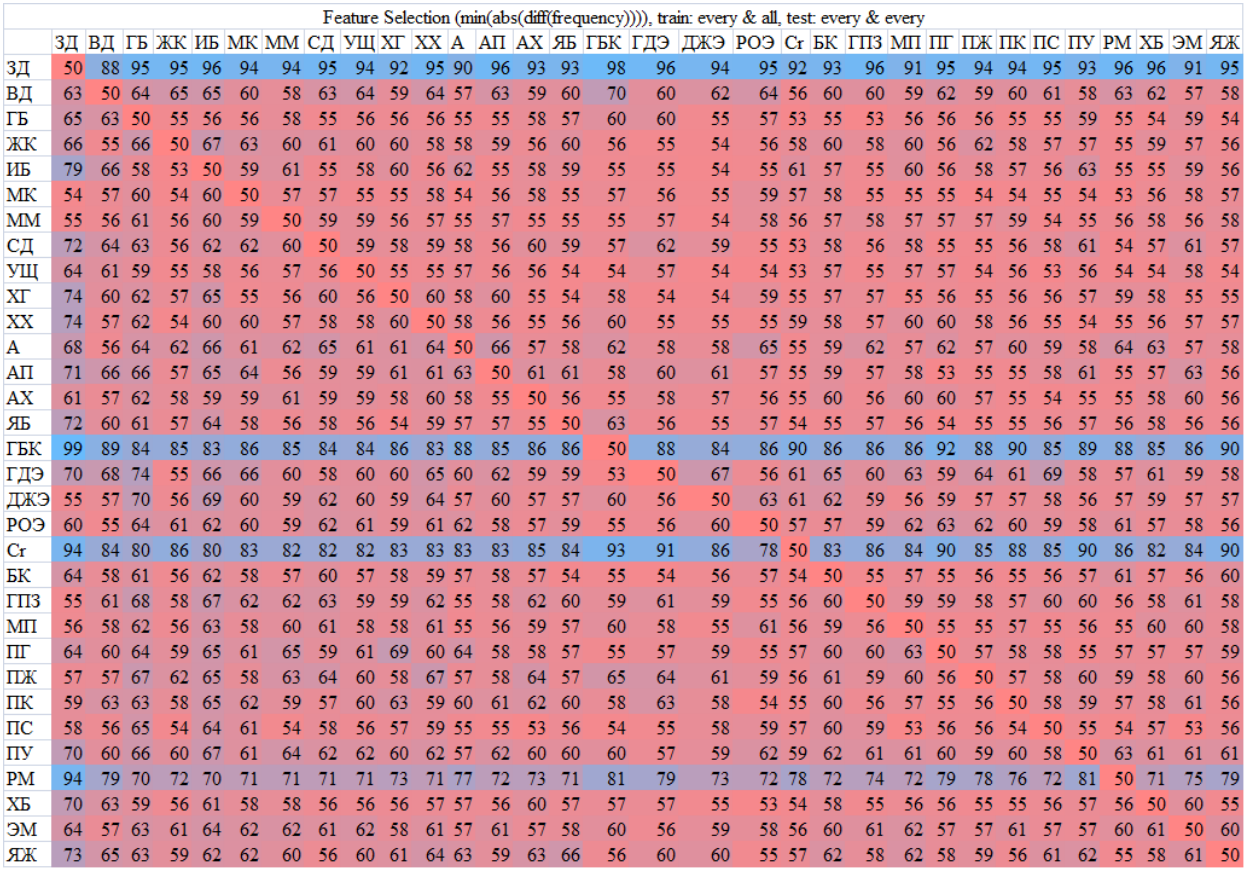
every&all	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC	0,90	0,71	0,69	0,74	0,69	0,71	0,69	0,74	0,69	0,69	0,72	0,73	0,75	0,76	0,70	0,75	0,76	0,75	0,68	0,61	0,72	0,76	0,70	0,74	0,70	0,72	0,72	0,71	0,64	0,73	0,75	0,74
K	49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	97	1	1	1	213	1	1	1	1	1	1	1	1	28	1	1	1

Рис. 14: Результаты на тесте после обучения по схеме 'каждый против всех'. Метод сортировки весов — 'Feature Selection (min)', веса признаков вычислены отдельно на каждой итерации кросс-валидации по обучающей выборке.

Как видно из результатов, признаки отбираются не адекватно (у большинства болезней отобрался 1 признак). Поэтому была применена еще одна схема обучения — 'average (every vs every)' (обозначения из файла 'Таблицы по данным.xlsx'). Она так же, как и 'каждый против всех', дает список значимых признаков для каждой болезни, только он получается как результат максимизации среднего арифмитического значений AUC по парам, составленным с каждой из других болезней:

Для каждой болезни $j = 1, 2, ..., L$:

Для каждого количества признаков $k = 1, 2, ..., M$:



	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	50	88	95	95	96	94	94	95	94	92	95	90	96	93	93	98	96	94	95	92	93	96	91	95	94	94	95	93	96	96	91	95
ВД	63	50	64	65	65	60	58	63	64	59	64	57	63	59	60	70	60	62	64	56	60	60	59	62	59	60	61	58	63	62	57	58
ГБ	65	63	50	55	56	56	58	55	56	56	56	55	55	58	57	60	60	55	57	53	55	53	56	56	56	55	55	59	55	54	59	54
ЖК	66	55	66	50	67	63	60	61	60	60	58	58	59	56	60	56	55	54	56	58	60	58	60	56	62	58	57	57	55	59	57	56
ИБ	79	66	58	53	50	59	61	55	58	60	56	62	55	58	59	55	55	54	55	61	57	55	60	56	58	57	56	63	55	55	59	56
МК	54	57	60	54	60	50	57	57	55	55	58	54	56	58	55	57	56	55	59	57	58	55	55	55	54	54	55	54	53	56	58	57
ММ	55	56	61	56	60	59	50	59	59	56	57	55	57	55	55	55	57	54	58	56	57	58	57	57	57	59	54	55	56	58	56	58
СД	72	64	63	56	62	62	60	50	59	58	59	58	56	60	59	57	62	59	55	53	58	56	58	55	55	56	58	61	54	57	61	57
УЩ	64	61	59	55	58	56	57	56	50	55	55	57	56	56	54	54	57	54	54	53	57	55	57	57	54	56	53	56	54	54	58	54
ХГ	74	60	62	57	65	55	56	60	56	50	60	58	60	55	54	58	54	54	59	55	57	57	55	56	55	56	56	57	59	58	55	55
ХХ	74	57	62	54	60	60	57	58	58	60	50	58	56	55	56	60	55	55	55	59	58	57	60	60	58	56	55	54	55	56	57	57
А	68	56	64	62	66	61	62	65	61	61	64	50	66	57	58	62	58	58	65	55	59	62	57	62	57	60	59	58	64	63	57	58
АП	71	66	66	57	65	64	56	59	59	61	61	63	50	61	61	58	60	61	57	55	59	57	58	53	55	55	58	61	55	57	63	56
АХ	61	57	62	58	59	59	61	59	59	58	60	58	55	50	56	55	58	57	56	55	60	56	60	60	57	55	54	55	55	58	60	56
ЯБ	72	60	61	57	64	58	56	58	56	54	59	57	57	55	50	63	56	55	57	54	55	57	56	54	55	55	56	57	56	58	56	56
ГБК	99	89	84	85	83	86	85	84	84	86	83	88	85	86	86	50	88	84	86	90	86	86	86	92	88	90	85	89	88	85	86	90
ГДЭ	70	68	74	55	66	66	60	58	60	60	65	60	62	59	59	53	50	67	56	61	65	60	63	59	64	61	69	58	57	61	59	58
ДЖЭ	55	57	70	56	69	60	59	62	60	59	64	57	60	57	57	60	56	50	63	61	62	59	56	59	57	57	58	56	57	59	57	57
РОЭ	60	55	64	61	62	60	59	62	61	59	61	62	58	57	59	55	56	60	50	57	57	59	62	63	62	60	59	58	61	57	58	56
Cr	94	84	80	86	80	83	82	82	82	83	83	83	83	85	84	93	91	86	78	50	83	86	84	90	85	88	85	90	86	82	84	90
БК	64	58	61	56	62	58	57	60	57	58	59	57	58	57	54	55	54	56	57	54	50	55	57	55	56	55	56	57	61	57	56	60
ГПЗ	55	61	68	58	67	62	62	63	59	59	62	55	58	62	60	59	61	59	55	56	60	50	59	59	58	57	60	60	56	58	61	58
МП	56	58	62	56	63	58	60	61	58	58	61	55	56	59	57	60	58	55	61	56	59	56	50	55	55	57	55	56	55	60	60	58
ПГ	64	60	64	59	65	61	65	59	61	69	60	64	58	58	57	55	57	59	55	57	60	60	63	50	57	58	58	55	57	57	57	59
ПЖ	57	57	67	62	65	58	63	64	60	58	67	57	58	64	57	65	64	61	59	56	61	59	60	56	50	57	58	60	59	58	60	56
ПК	59	63	63	58	65	62	59	57	60	63	59	60	61	62	60	58	63	58	54	55	60	56	57	55	56	50	58	59	57	58	61	56
ПС	58	56	65	54	64	61	54	58	56	57	59	55	55	53	56	54	55	58	59	57	60	59	53	56	56	54	50	55	54	57	53	56
ПУ	70	60	66	60	67	61	64	62	62	60	62	57	62	60	60	60	57	59	62	59	62	61	61	60	59	60	58	50	63	61	61	61
РМ	94	79	70	72	70	71	71	71	71	73	71	77	72	73	71	81	79	73	72	78	72	74	72	79	78	76	72	81	50	71	75	79
ХБ	70	63	59	56	61	58	58	56	56	56	57	57	56	60	57	57	57	55	53	54	58	55	56	56	55	55	56	57	56	50	60	55
ЭМ	64	57	63	61	64	62	62	61	62	58	61	57	61	57	58	60	56	59	58	56	60	61	62	57	57	61	57	57	60	61	50	60
ЯЖ	73	65	63	59	62	62	60	56	60	61	64	63	59	63	66	56	60	60	55	57	62	58	62	58	59	56	61	62	55	58	61	50

Рис. 15: $100 \times AUC$. Результаты на тесте по схеме 'каждый против каждого' после обучения по схеме 'каждый против всех'. Метод сортировки весов — 'Feature Selection (min)', веса признаков вычислены отдельно на каждой итерации кросс-валидации по тестовой выборке.

Для каждой болезни $j' = 1, 2, ..., L, j' \neq j$:

вычислить усредненный по кросс-валидации $AUC_{jj'k}$

усреднить $AUC_{jj'k}$ по болезням: $AUC_{jk} = \frac{\sum_{j'} AUC_{jj'k}}{(L-1)}$

найти максимум AUC_{jk} по количесву признаков: $AUC_j = \max_k AUC_{jk}$, $K_j = \arg\max_k AUC_{jk}$

Feature Selection (min(abs(diff(frequency)))), train: average (every & every), test: every & every, веса признаков посчитаны для каждой части выборки индивидуально																															
average (every & every)	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC	0,93	0,73	0,64	0,68	0,67	0,65	0,66	0,67	0,64	0,65	0,65	0,68	0,68	0,68	0,65	0,81	0,72	0,67	0,71	0,70	0,65	0,66	0,67	0,72	0,69	0,68	0,66	0,68	0,69	0,65	0,67	0,70
K	49	215	216	202	212	214	214	216	215	216	215	216	107	201	210	154	216	216	208	213	209	210	216	123	194	158	202	212	114	215	206	69

Рис. 16: Результаты на обучении по схеме 'average(every vs every)'. Метод сортировки весов — 'Feature Selection (min)', веса признаков вычислены отдельно на каждой итерации кросс-валидации по обучающей выборке.

Как видно из результатов (см. рис. 17 и 16), эта схема обучения наоборот, отбирает большую часть признаков, чего хотелось бы избежать.

6.2 Веса вычисляются по всей выборке и для обучения, и для контроля

Байесовский классификатор, веса признаков считаются по всей выборке. До этого во время кросс-валидации веса считались по той выборке, по которой происходило обучение на данной итерации. Было замечено, что вычисление весов признаков по всей выборке до кросс-валидации повышает качество классификации (AUC), но не существенно (в среднем по парам болезней (без удаления пересечений классов) — на $0,027$). Однако, такой метод вычисления весов дает почти в 2 раза больший прирост AUC, чем рассмотренные выше альтернативные методы сортировки признаков.

Эксперимент с сортировкой признаков 'Feature Selection (min)' и схемой обучения 'average (every vs every)' был повторен с вычислением весов признаков по всей выборке. Этот эксперимент показал уменьшение AUC в среднем по парам болезней на $0,008$ по сравнению с сортировкой признаков по модулям весов и схемой обучения 'каждый против каждого' с вычислением весов также по всей выборке.

В качестве порога бинаризации признаков было применено среднее арифметическое значение признака по всей выборке. Этот эксперимент

Feature Selection (min(abs(diff(frequency)))), train: average (every & every), test: every & every, веса признаков посчитаны для каждой части выборки индивидуально																																
	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	50	88	95	95	96	94	94	95	94	92	96	90	96	93	93	97	96	94	95	92	94	96	92	94	94	94	95	93	96	96	92	95
ВД	89	50	76	82	80	73	72	80	75	71	76	70	82	72	72	88	80	73	83	83	73	81	71	88	81	83	74	82	82	78	69	85
ГБ	95	76	50	72	58	62	65	64	63	65	63	71	68	70	65	86	75	66	74	80	65	67	67	83	77	78	68	80	76	68	69	83
ЖК	95	82	72	50	70	74	75	71	74	76	72	81	76	78	75	88	84	81	79	86	75	76	76	87	85	83	77	85	79	75	78	86
ИБ	96	80	58	70	50	66	70	64	65	68	64	75	67	73	68	85	79	72	74	79	67	69	71	83	78	78	71	82	76	69	72	83
МК	93	73	62	74	66	50	66	69	67	68	66	70	71	71	65	88	79	71	78	82	65	70	68	84	79	78	68	80	76	67	67	84
ММ	94	72	65	75	70	66	50	72	66	68	69	68	76	68	70	88	82	74	81	83	68	72	64	85	77	79	74	81	77	69	64	84
СД	95	80	64	71	64	69	72	50	68	72	68	77	71	73	70	87	83	77	74	81	69	71	73	86	79	79	73	83	78	71	73	85
УЩ	94	75	63	74	65	67	66	68	50	68	66	72	71	71	69	87	81	75	77	82	68	74	67	85	79	77	73	81	76	69	70	84
ХГ	92	71	65	76	68	68	68	72	68	50	66	70	74	72	70	89	83	72	78	83	65	72	69	81	80	78	71	82	78	70	69	86
ХХ	95	76	63	73	64	66	69	68	66	66	50	73	71	69	67	86	77	68	76	82	66	70	69	85	77	79	69	82	77	70	69	83
А	90	70	71	81	75	70	68	77	72	70	73	50	79	71	73	90	84	75	81	84	70	77	71	85	80	82	75	84	80	75	69	85
АП	96	82	66	74	65	69	73	69	69	73	68	78	50	78	71	88	80	73	76	81	73	75	75	85	80	78	74	84	78	72	78	83
АХ	94	72	70	77	73	70	68	73	71	72	69	71	78	50	74	87	83	73	82	85	70	75	70	90	77	81	74	84	78	71	67	85
ЯБ	93	72	65	75	68	65	70	70	68	70	67	73	72	74	50	89	83	73	78	85	71	72	70	87	80	79	70	82	77	67	69	82
ГБК	98	88	85	86	84	87	87	86	86	88	84	89	87	86	87	50	90	87	87	91	87	88	87	94	90	91	86	92	89	86	87	93
ГДЭ	96	80	75	84	79	79	82	83	81	83	77	84	81	83	83	91	50	74	86	91	78	83	83	92	88	88	76	87	85	80	83	90
ДЖЭ	94	73	66	81	72	71	74	77	75	72	68	75	76	73	73	88	74	50	81	85	70	75	74	89	85	83	73	83	82	74	72	88
РОЭ	95	83	74	79	74	78	80	74	77	78	76	81	78	82	78	87	86	81	50	77	78	80	81	87	82	85	80	86	78	78	80	87
Cr	94	83	80	86	79	82	82	81	82	83	82	84	82	86	85	93	91	85	77	50	83	83	83	91	86	86	86	90	85	81	85	90
БК	93	73	65	75	67	65	68	69	68	65	66	70	74	70	71	88	78	70	78	83	50	71	68	84	77	79	69	81	78	68	67	84
ГПЗ	96	81	67	76	69	70	72	71	74	72	70	77	76	76	72	89	83	75	80	83	71	50	75	87	79	82	73	84	81	74	74	86
МП	92	71	67	75	71	68	64	73	67	69	69	71	76	71	70	88	83	74	81	83	69	75	50	87	77	79	73	81	77	69	68	84
ПГ	96	88	81	85	80	82	84	84	83	81	83	84	83	88	86	93	91	87	85	91	83	85	85	50	86	84	86	90	86	85	87	89
ПЖ	95	81	77	85	78	79	77	79	79	80	77	80	82	78	80	92	88	84	83	86	77	79	77	87	50	85	83	85	85	80	80	87
ПК	94	84	78	84	77	78	79	78	77	78	78	81	78	82	79	92	88	83	84	85	79	81	79	85	84	50	81	88	84	81	81	87
ПС	95	74	68	78	71	68	74	73	73	71	69	75	76	74	70	88	76	72	80	85	69	73	73	87	83	82	50	84	81	73	72	86
ПУ	95	82	80	85	82	80	81	83	81	82	82	84	84	84	82	93	88	83	86	89	81	84	80	91	85	88	84	50	86	83	82	90
РМ	96	82	74	78	75	74	75	76	74	76	75	80	78	78	76	89	84	79	77	84	77	80	76	86	83	84	79	85	50	76	77	87
ХБ	95	78	68	75	69	67	69	71	69	70	70	75	74	71	67	87	80	74	79	81	68	74	69	87	80	83	73	83	78	50	71	86
ЭМ	92	69	69	77	72	67	64	73	69	69	69	69	78	68	69	88	83	72	80	85	67	75	68	88	80	81	72	82	78	71	50	85
ЯЖ	95	84	81	84	80	82	82	82	83	82	82	82	81	85	81	92	90	85	85	88	82	84	82	89	85	86	84	89	85	84	84	50

Рис. 17: $100 \times AUC$. Результаты на тесте по схеме 'каждый против каждого' после обучения по схеме 'average(every vs every)'. Метод сортировки весов — 'Feature Selection (min)', веса признаков вычислены отдельно на каждой итерации кросс-валидации по тестовой выборке.

проводился без учета 'абсолютного здоровья', то есть объекты, относящиеся к этому классу, были удалены из выборки, так как этот класс сильно отличается от всех остальных. Итак:

1. $ECG = \|ecg_{ik}\|_{N \times M}$ — матрица объект-признак с небинаризованными признаками, где ecg_{ik} — сколько раз k-ая триграмма встретилась в i-ой кодограмме. Вычислить $\delta = \|\delta_k\|$, $\delta_k = \frac{\sum_i^N ecg_{ik}}{N}$ — порог бинаризации для k-ого признака.

2. Бинаризовать признаки:

$$x_{ik} = \begin{cases} 1, & ecg_{ik} \geq \delta_k; \\ 0, & \text{иначе.} \end{cases}$$

Сравнение матрицы AUC для всех пар классов (кроме абсолютного здоровья) со схемой обучения 'каждый против каждого', сортировкой признаков по модулям весов и вычислением весов по всей выборке при бинаризации признаков по порогу 2 и по порогу — среднему значению признака показывает, что в среднем по парам болезней AUC больше на $0,003$ в случае, когда порог бинаризации равен 2.

6.3 Кластеризация внутри класса

Как видно из предыдущих экспериментов, наивный байесовский классификатор в этой задаче отбирает около трети всех признаков, и при этом качество классификации в среднем по парам болезней оставляет желать лучшего. Если рассматривать зависимость AUC от количества учитываемых признаков c, то характерным для большинства пар классов будет следующее: AUC почти не изменятеся на промежутке примерно от K до 216 признаков, где $K = \arg\max_c AUC(c)$. Возможно, что плохое качество классификации является результатом избыточного набора значимых признаков.

Отсюда следует вывод, что нужно найти более эффективный способ отбора признаков, и возможно, качество классификации улучшится. Неэффективный отбор признаков может быть связан с тем, что классы внутри себя не однородны, и внутри них есть кластера, характеризующиеся разными наборами признаков. Чтобы это проверить, была проведена кластеризация внутри класса. Первой идеей было использовать LVQ, в котором сочетаются кластеризация и классификация.

В LVQ кластеры привязываются к классу, то есть объект, отнесенный к какому-либо кластеру, автоматически признается принадлежащим к тому же классу, что и этот кластер. В таком случае AUC вычислить нельзя, потому что неизвестно значение дискриминантной функции, но можно вычислить TPR и FPR. По этим показателям видно, что сохраняется общая тенденция: абсолютное здоровье и болезнь различаются хорошо ($TPR > 0.9$, $FPR < 0.2$), болезнь от болезни — плохо (тенденция относить все к одному классу, то есть TPR и FPR оба близки либо к 0, либо к 1). Эксперимент был поставлен на выборках с небинаризованными признаками без пересечений классов со следующими парами классов: МКЭ vs ХХЭ, ММЭ vs ХХЭ, ММЭ vs ЯБЭ, ХХЭ vs ЯБЭ, ЖКЭ vs ИБЭ и ИБЭ vs СДЭ с 4 кластрами в классе, АЗ vs ЖКЭ с перебором количества кластеров от 1 до 4 в классе (от изменения количества кластеров картина почти не менялась). Использовалась реализация LVQ в Matlab: Neural Network Toolbox/Pattern Recognition and Classification/Learning Vector Quantization. В качестве признаков использовались биграммы, а не триграммы, для уменьшения размерности признакового пространства (биграмм всего 36). Использовалась евклидова метрика.

Чтобы можно было понять, улучшает ли кластеризация качество классификации, был использован следующий способ вычисления AUC для LVQ:

- обучение LVQ на двухклассовой выборке без пересечений классов;
- для каждого объекта выборки вычисляется значение дискриминантной функции:
  - объект кластеризуется и классифицируется в соответствии с обученным LVQ;
  - находится ближайший к объекту кластер из другого класса (не из того класса, к которому объект был отнесен на предыдущем шаге);
  - на кластере, к которому относится объект, и на ближайшем к объекту кластере из другого класса как на двухклассовой выборке (кластреы — это классы) строится наивный байесовский классификатор с сортировкой признаков по модулям весов и порогом бинаризации признаков равным 2 (для биграмм — 8);

  - значение дискриминантной функции на объекте в соответствии с LVQ и в соответствии с построенным на кластерах наивном байесовском классификаторе принимаются равными;

- вычисляется AUC.

Такая схема была применена к одной паре классов, и выяснилось, что LVQ находит в классе не больше 2 кластеров. Точнее, центры остальных кластеров мало отличаются от найденных двух, и при кластеризации после обучения вся обучающая выборка относится только к двум кластерам. Поэтому было решено использовать более быстро обучающийся алгоритм — k-means. Использовалась его реализация в Matlab: Statistics Toolbox/Exploratory Data Analysis/Cluster Analysis/k-Means Clustering. Так как k-means не классифицирует, то схема эксперимента немного изменилась: каждый класс в паре кластеризовался отдельно, полученные кластеры привязывались к классу. Затем, также, как в LVQ, после кластеризации по кластеру определялся класс. k-means стабильно находил от 1 до 3 кластеров в классе, по 4 кластера в классе находил с перебоями (на некоторых итерациях кросс-валидации к одному из найденных кластеров не относился ни один объект из обучающей выборки), больше 4 кластеров найти не мог (остальные были пустыми). Эксперимент был проведен на одной паре 'здоровье — болезнь' и на одной паре 'болезнь — болезнь' без пересечений классов, признаки — небинаризованные биграммы, метрика евклидова. По сравнению с НБ классификатором значимых улучшений по значению AUC нет. Эксперимент был повторен для триграмм с евклидовой и манхэттенской метриками, в обоих случаях улучшения также нет.

6.4 EM-алгоритм

Для кластеризации внутри класса реализован на Python (но не отлажен) EM-алгоритм с пуассоновсим распределением (признаки считаются независимыми), показанный на рис. 18 и 19.

Далее нужно его модифицировать так, чтобы он отбирал признаки. Пока предполагается использовать метод Feature Saliency из статьи M. Law, M. Figueiredo, and A. Jain. Feature Saliency in Unsupervised Learning 2002

Алгоритм 2.2. EM-алгоритм с фиксированным числом компонент

Вход:

выборка $X^m = \{x_1, \ldots, x_m\}$;
k — число компонент смеси;
$\Theta = (w_j, \theta_j)_{j=1}^k$ — начальное приближение параметров смеси;
δ — параметр критерия останова;

Выход:

$\Theta = (w_j, \theta_j)_{j=1}^k$ — оптимизированный вектор параметров смеси;

ПРОЦЕДУРА EM (X^m, k, Θ, δ);
повторять
  E-шаг (expectation):
  для всех $i = 1, \ldots, m, \;\; j = 1, \ldots, k$

$$g_{ij}^0 := g_{ij}; \quad g_{ij} := \frac{w_j \varphi(x_i; \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i; \theta_s)};$$

  M-шаг (maximization):
  для всех $j = 1, \ldots, k$

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i; \theta); \quad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

пока $\max_{i,j} |g_{ij} - g_{ij}^0| > \delta$;
вернуть $(w_j, \theta_j)_{j=1}^k$;

Рис. 18: Процедура EM

Алгоритм 2.3. EM-алгоритм с последовательным добавлением компонент

Вход:

выборка $X^m = \{x_1, \ldots, x_m\}$;
R — максимальный допустимый разброс правдоподобия объектов;
m_0 — минимальная длина выборки, по которой можно восстановить плотность;
δ — параметр критерия останова;

Выход:

k — число компонент смеси;
$\Theta = (w_j, \theta_j)_{j=1}^k$ — веса и параметры компонент;

1: начальное приближение — одна компонента:
$\theta_1 := \arg\max_\theta \sum_{i=1}^m \ln \varphi(x_i; \theta)$; $\quad w_1 := 1$; $\quad k := 1$;
2: **для всех** $k := 2, 3, \ldots$
3: выделить объекты с низким правдоподобием:
$U := \{x_i \in X^m : p(x_i) < \max_j p(x_j)/R\}$;
4: **если** $|U| < m_0$ **то**
5: **выход** из цикла по k;
6: начальное приближение для k-й компоненты:
$\theta_k := \arg\max_\theta \sum_{x_i \in U} \ln \varphi(x_i; \theta)$; $\quad w_k := \frac{1}{m}|U|$;
$w_j := w_j(1 - w_k)$, $\; j = 1, \ldots, k-1$;
7: EM (X^m, k, Θ, δ);

Рис. 19: EM алгоритм с последовательным добавлением компонент