Многоклассовый прогноз вероятности наступления инфаркта*

А. П. Мотренко
pastt.petrovna@gmail.com
Московский физико-технический институт, ФУПМ, кафедра "Интеллектуальные системы"

В работе описан алгоритм, позволяющий классифицировать четыре группы пациентов: перенесших инфаркт; больных, имеющих предрасположенность к инфаркту и здоровых пациентов двух групп. Признаками для определения состояния пациента служат измерения концентрации белков в крови. Одной из задач работы является выбор набора маркеров, оптимального для разделения между собой соответствующих групп. Классификация осуществляется по принципу «каждый против каждого», то есть решаются задачи классификации всевозможных пар групп. В силу высокой стоимости анализа крови, объемы данных невелики, поэтому одним из результатов исследования является оценка необходимого объема выборки пациентов.

Ключевые слова: *логистическая регрессия, многоклассовая классификация, выбор признаков, оценка необходимого объема выборки, расстояние Кульбака-Лейблера.*

Введение

При выборе линейных моделей, включающих относительно небольшое количество признаков, решается задача оценки совместной сложности модели и оценки необходимого числа параметров объектов. Особенность исследуемой задачи в том, что стоимость получения выборки высока. Поэтому часть работы посвящена развитию методов оценки необходимого объема выборки по измеренным данным, сопряженному с задачей выбора моделей.

Заболевания сердечно-сосудистой системы могут протекать, не проявляясь клинически. Тем не менее, обнаружение нарушений, связанных с работой сердца, по косвенным признакам, или биомаркерам, вполне возможно [3]. Традиционными маркерами являются возраст, давление крови и уровень холестерина; существуют и другие показатели, например, определенные группы генов [7]. В данной работе предлагается использовать в качестве маркеров концентрации белков в клетках крови. Разделение пациентов по состоянию здоровья на четыре группы: больные, перенесшие инфаркт; больные, имеющие расположенность к инфаркту и здоровые двух типов приводит к задаче многоклассового прогнозирования. Такую задачу можно свести к задаче двуклассовой классификации, используя один из следующих подходов:

1) один против всех. Данный подход заключается в следующем: выделяем один группу пациентов как отдельный класс, а все остальные группы объединяем во второй класс и решаем, таким образом, задачу выделения определенной группы.
2) каждый против каждого. В этом случае перебираются все возможные пары групп.

Другим словами, в первом случае ставится вопрос «относится ли пациент к данной группе?», во втором — «к какому из двух данных групп он принадлежит с большей вероятностью?». Комбинация этих подходов приводит к еще одному способу: разделив каким-либо образом все множество групп на два непересекающихся подмножества, образовать два класса. Например, отличать две имеющиеся группы больных от двух групп здоровых

Научный руководитель В. В. Стрижов

пациентов. Выбор каждой стратегии зависит от конкретных задач; в работе приводится решение задачи с использованием второй стратегии, так как она дает наиболее подробные результаты. Обратим внимание на различие понятий «класс» и «группа». Под группой везде далее будем понимать группу, определяющую состояние здоровья пациента. В свою очередь, класс — понятие, связанное с задачей классификации, он может состоять как из одной, так и из нескольких групп пациентов.

В работе решаются задачи многоклассовой классификации с выбором признаков и оценки минимального объема выборки, достаточного для проведения классификации. Первая часть работы посвящена отбору наиболее информативных признаков, т.е выбору набора признаков, наилучшим образом разделяющего классы. На практике снижение количества измеряемых признаков диагностируемых пациентов приводит к уменьшению финансовых затрат на получение признаков и позволяет увеличить количество исследуемых пациентов, то есть объем выборки.

В работе рассматривается задача разделения пар классов. Предполагается, что число измеряемых признаков избыточно. Задача состоит в отыскании оптимального набора признаков, эффективно разделяющего между собой классы пациентов.

Для каждой пары групп решается задача логистической регрессии [4]. В ее основе лежит предположение о биномиальном распределении независимой переменной с оценкой параметров функции регрессии по методу Ньютона-Рафсона. Выбор признаков в логистической регрессии производится с помощью шаговой регрессии [6] или полного перебора. В данной работе используется полный перебор, т.к он дает экспертам уверенность в том, что рассмотрены все возможные сочетания признаков при выборе модели, а в как функционал качества используется площадь под графиком ROC-кривой [10].

Во второй части работы оценивается минимальный объем выборки, необходимый для проведения вычислительного эксперимента. Для оценки применяются следующие методы: метод доверительных интервалов [9], метод скользящего контроля [8], а также используется расстояние Кульбака-Лейблера [2] для сравнения предполагаемых распределений на различных подвыборках.

Задача классификации

Дана выборка $D = \mathbf{x}_i, y_i, i = 1, \ldots, m$ объектов (пациентов), каждый из которых описывается n признаками (биомаркерами) и принадлежит одному из двух классов $y_i \in \{0, 1\}$. Объединим признаки в столбцы $\boldsymbol{\chi}_j \in \mathbb{R}^m, j = 1, \ldots, n$ и составим из них матрицу $X = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n] = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T\right]^T$. Рассмотрим задачу логистической регрессии. Пусть случайная величина y имеет распределение Бернулли с параметром p, $y \sim B(p, 1-p)$, тогда

$$y = \begin{cases} 1, & p; \\ 0, & 1-p. \end{cases} \tag{1}$$

Функция плотности $p(y|p)$ в таком случае имеет вид

$$p(y|p) = p^y(1-p)^{1-y}. \tag{2}$$

В логистической регрессии предполагается, что вектор ответов $\mathbf{y} = [y_1, \ldots, y_m]$ — бернуллиевский случайный вектор с независимыми компонентами $y_i \sim \mathcal{B}(p_i, 1-p_i)$ и функцией плотности

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}. \tag{3}$$

Определим функцию ошибки следующим образом:

$$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln p_i + (1-y_i)\ln(1-p_i). \tag{4}$$

Другими словами, функция штрафа есть логарифм плотности, или функции правдоподобия, со знаком минус. В случае логистической регрессии

$$p_i = \frac{1}{1+exp(-\mathbf{x}_i^T\mathbf{w})} = \sigma(\mathbf{x}_i^T\mathbf{w}) \equiv \sigma_i. \tag{5}$$

Воспользовавшись тождеством $\frac{d\sigma(\theta)}{d\theta} = \sigma(1-\sigma)$, вычислим градиент функции $E(\mathbf{w})$

$$\nabla E(\mathbf{w}) = -\sum_{i=1}^{m} (y_i(1-\sigma_i) - (1-y_i)\sigma_i)\,\mathbf{x}_i = \sum_{i=1}^{m}(\sigma_i - y_i)\mathbf{x}_i = X^T(\sigma - \mathbf{y}),$$

где $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^T$. Оценка параметров осуществляется по схеме Ньютона-Рафсона. Введем обозначение Σ — диагональная матрица с элементами $\Sigma_{ii} = \sigma_i(1-\sigma_i), i = 1, \ldots, m$. Пусть параметры оцениваются на множестве $\mathcal{W}$. В качестве начального значения $\mathbf{w}_0$ возьмем

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}\in\mathcal{W}} E(\mathbf{w}).$$

Тогда итеративная оценка параметров логистической регрессии (5) имеет вид

$$\mathbf{w}^{k+1} = \mathbf{w}^k - (X^T\Sigma X)^{-1}X^T(\boldsymbol{\sigma} - \mathbf{y}) = (X^T\Sigma X)^{-1}X^T\Sigma(X\mathbf{w}^k - \Sigma^{-1}(\boldsymbol{\sigma} - \mathbf{y})). \tag{6}$$

Процедура оценки параметров повторяется, пока норма разности $\left\|\mathbf{w}^{k+1} - \mathbf{w}^k\right\|$ не станет достаточно мала.

Алгоритм классификации имеет вид:

$$a(\mathbf{x}) = \operatorname{sign}(\sigma(\mathbf{x}, \mathbf{w}) - \sigma_0), \tag{7}$$

где σ_0 — задаваемое пороговое значение функции регрессии, о выборе σ_0 будет рассказано позже. Для контроля за качеством классификации можно использовать следующий функционал:

$$Q = (1 - TPR)^2 + FPR^2,$$

где

$$TPR = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) = 1][y_i = 1]$$

(*true positive rate*) — доля элементов выборки, правильно классифицированных в пользу заданного класса;

$$FPR = \frac{1}{m}\sum_{i=1}^{m}[a(\mathbf{x}_i) = 1][y_i = 0]$$

(*false positive rate*) — доля ошибочно классифицированных в пользу данного класса элементов выборки. Здесь используется обозначение индикаторной функции:

$$[y = 1] = \begin{cases} 1, & y = 1; \\ 0, & y \neq 1. \end{cases} \tag{8}$$

Таким образом, алгоритм тем лучше разделяет классы, чем меньше значение функционала Q. В данной работе используется альтернативный критерий. Отложив на графике по оси абсцисс значения FPR, а по оси ординат — TPR, получим так называемую ROC-кривую, каждая точка которой соответствует некоторому значению σ_0. В алгоритме (7) используется значение σ_0, отвечающее наибольшему расстоянию точки ROC-кривой до линии $TPR = FPR$. В качестве максимизируемого функционала будем использовать площадь под кривой, AUC (area under curve). Выбрав для каждого $\mathbf{x}$ пороговое значение σ_0 равным $\sigma(\mathbf{x}, \mathbf{w})$, вычислим TPR и FPR и построим по ним ROC-кривую. Выбирая различные наборы маркеров (то есть меняя координаты элементов выборки в признаковом пространстве) будем получать различные кривые. Оптимальному случаю соответствует кривая с наибольшей площадью под графиком (AUC).

Пусть $\mathcal{A}$ — некоторое подмножество индексов маркеров, $\mathcal{A} \subseteq \mathcal{J} = \{1, \ldots, n\}$, $\hat{\mathcal{A}}$ — оптимальный набор индексов. Тогда задачу можно сформулировать как задачу максимизации:

$$\hat{\mathcal{A}} = \arg\max_{\mathcal{A} \subseteq \mathcal{I}} \mathrm{AUC}(\sigma(X_{\mathcal{A}}, \mathbf{w}_{\mathcal{A}}), \mathbf{y}), \tag{9}$$

где $X_{\mathcal{A}}$ — состоит из столбцов $\mathbf{x}_j$, $j \in \mathcal{A}$, $\mathbf{w}_{\mathcal{A}}$ — вектор параметров, рассчитанный по формуле (6). При этом разбиение выборки на обучающую и контрольную не производится, т.к. при постановке задачи экспертами наложено ограничение на сложность модели — число признаков, входящих в модель, не должно превышать четырех.
Наборы признаков, полученные с помощью алгоритма (9), будем называть оптимальным для данной пары классов, а сами признаки — *наиболее информативным.*

Подготовка данных. Особенность используемых данных состоит в наличии большого числа пропущенных значений признаков. Прежде чем приступить к решению задачи классификации, предложим некоторые пути решения этой проблемы.

1) Заполнять пустующие позиции средним в данной группе значением признака.
2) Воспользовавшись предположением о вероятностном распределении случайной величины x_{ij}, реализациями которой являются значения признаков $\boldsymbol{\chi}_j$, заполнять пропуски соответствующими этому распределению случайными величинами в интервале от минимального значения признака в выборке до максимального. В данной работе используется предположение о нормальности случайных величин. Обозначим $\mathcal{I} = \{1, \ldots, m\}$ — множество индексов объектов, и фиксируем $\mathcal{N}(\mu, \sigma^2)$ — некоторую реализацию нормально распределенной с математическим ожиданием μ и дисперсией σ^2 случайной величины. Будем заполнять пропущенные величины по следующему правилу:

$$\begin{cases} \min\limits_{i \in \mathcal{I}} x_{ij}, & \min\limits_{i \in \mathcal{I}} x_{ij} > \mathcal{N}(\mu, \sigma^2); \\ \max\limits_{i \in \mathcal{I}} x_{ij}, & \mathcal{N}(\mu, \sigma^2) > \max\limits_{i \in \mathcal{I}} x_{ij}; \\ \mathcal{N}(\mu, \sigma^2), & \min\limits_{i \in \mathcal{I}} x_{ij} \leqslant \mathcal{N}(\mu, \sigma^2) \leqslant \max\limits_{i \in \mathcal{I}} x_{ij}. \end{cases}$$

3) **Multiple imputation for missing values**. Как и в предыдущем пункте, сделаем предположение о распределении пропущенных величин, но не будем сразу заполнять пропуски реализациями этой случайной величины. ведем обозначение для $k \in \mathcal{B}* \subset \mathcal{I} = \{1, \ldots, m\}$ индексов таких, что x_{kj} — пропущенное значение. В процедурах (6), (9) оценку $\mathbf{w}$ и AUC будем проводить по K реализациям переменных $x_{kj}, k \in \mathcal{B}*$. После этого используем медиану K полученных оценок $\mathbf{w}$, AUC.

Таблица 1. Всевозможные значения параметра α

α_1	α_2	$\ldots$	α_n
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	1	1

.

Таблица 2. Возможный результат классификаций с использованием элемента $\mathbf{x}_{m+1}$

	A1	A3	B1	B2
A1	-	0	0	1
A3	1	-	1	1
B1	1	0	-	0
B2	0	0	1	-

.

Предлагается использовать последний подход. Второй способ несколько проще реализовать, однако его использование приводит к неустойчивости результата — определяемый алгоритмом набор признаков существенно зависит от текущей реализаций случайной величины, используемой для заполнения пробелов.

Отбор признаков. Отбор признаков $\mathbf{x}_j, j \in \mathcal{A}$ осуществляется путем полного перебора. Такой подход возможен благодаря сравнительно небольшому количеству признаков. Полный перебор приводит к выбору набора признаков, лучше всего отвечающего заданному критерию. Запишем выражение для функции регрессии в виде

$$f(\mathbf{x}_i, \mathbf{w}) = \sigma(\mathbf{x}_i^T \mathbf{w}), \mathbf{x}_i^T \mathbf{w} = \alpha_1 x_{i1} \mathrm{w}_1 + \alpha_2 x_{i2} \mathrm{w}_2 + \cdots + \alpha_n x_{in} \mathrm{w}_n,$$

здесь $\alpha_j \in \{0, 1\}$ — структурный параметр. Таким образом, перебор признаков сводится к перебору значений элементов α_j вектора структурных параметров, см. (1).

Выполнение прогноза при многоклассовой классификации. По состоянию здоровья пациентов можно разделить на 5 групп.

1) $\mathbf{A}_1$. Группа пациентов, уже перенесших инфаркт.
2) $\mathbf{A}_2$. К этой группе относят пациентов, анализы которых были получены прямо в во время инфаркта. Это очень редкая группа, поэтому в данной работе она не рассматривается.
3) $\mathbf{A}_3$. Пациенты, имеющие предрасположенность к инфаркту.
4) $\mathbf{B}_1, \mathbf{B}_2$. Здоровые пациенты двух типов.

При появлении в выборке нового объекта $\mathbf{x}_{m+1}$ выполняем следующую процедуру. Решаем задачу классификации для всех пар групп. При этом для каждой пары используется оптимальный для нее набор признаков. Решив задачу логистической регрессии (7), в каждом случае получим вероятность p_{m+1} принадлежности объекта к одному из двух рассматриваемых классов. По этим результатам составим таблицу (**??**):

Здесь каждая строка есть результат сравнения некоторого класса с каждым из остальных. Например, в третьей строке содержится следующая информация: объект $\mathbf{x}_{m+1}$ более похож на объект класса B_1, чем на A_1 (в соответствующей ячейке стоит единица), но менее похож на B_1, чем на A_3 и B_2 (в ячейках нули). Присвоив классам A_1, A_3, B_1, B_2 номера 1, 2, 3, 4 соответственно, переформулируем последнее утверждение:

$$a_{23}(\mathbf{x}_{m+1}) = 0,$$

Таблица 3. Результаты отбора признаков

classes	obj. in both classes	in 1st class	best sets	AUC	Err_1	Err_2
A1 A3	31	14	[2, 11, 19, 20]	0.953	0.262	0
			[2, 13, 19, 20]	0.953		
			[2, 16, 19, 20]	0.962		
			[2, 14, 19, 20]	0.966		
			[2, 17, 19, 20]	0.970		
A1 B1	55	14	[3, 13, 18, 19]	0.829	0.254	0
			[12, 13, 15, 19]	0.829		
			[8, 18, 19, 20]	0.831		
			[13, 15, 18, 19]	0.831		
			[12, 13, 18, 19]	0.850		
A1 B1	55	14	[5, 15, 17, 19]	0.901	0.207	0
			[6, 12, 15, 19]	0.901		
			[3, 12, 15, 19]	0.903		
			[9, 12, 15, 19]	0.903		
			[12, 15, 17, 19]	0.909		
A3 B1	58	17	[5, 6, 11, 17]	0.814	0.293	0
			[2, 7, 9, 13]	0.829		
			[7, 13, 18, 20]	0.834		
			[2, 3, 5, 9]	0.835		
			[2, 5, 6, 9]	0.836		
A3 B2	43	17	[2, 3, 5, 9]	0.954	0.239	0
			[2, 3, 9, 19]	0.957		
			[2, 9, 18, 19]	0.959		
			[2, 3, 9, 17]	0.963		
			[2, 3, 9, 13]	0.970		
B1 B2	67	41	[1, 2, 3, 9]	0.821	0.563	0
			[2, 3, 9, 11]	0.823		
			[1, 2, 18, 19]	0.824		
			[2, 13, 18, 19]	0.827		
			[2, 3, 9, 18]	0.829		

где $a_{lk}(x) = \xi \in \{0, 1\}, l, k = 1, \ldots, 4$ — результат работы алгоритма (7) при выборе между классами l и k. Таким образом, мы относим объект к тому классу, для которого сумма элементов таблицы по строке наибольшая:

$$\text{class}(\mathbf{x}_{m+1}) = \arg\max_{l=1,\ldots,4} \sum_{k=1}^{4} a_{lk}(\mathbf{x}_{m+1}), \text{class}(\mathbf{x}_{m+1}) \in \{1, \ldots, 4\}.$$

Если эта сумма для двух классов совпала, результатом будет решение задачи классификации, полученное для этих двух классов.

Оценка объема выборки

В качестве обоснования достоверности классификации приводится исследование необходимого размера выборки пациентов. Рассмотрим три способа получения этой оценки.

Метод доверительных интервалов. Пусть имеется выборка независимых одинаково распределенных случайных величин $\{x_i\}\, i = 1, \ldots, m$. Подсчитанное по этой выборке

среднее арифметическое $\bar{x}$ в общем случае не совпадает с матожиданием μ рассматриваемой случайной величины. Пусть $E = \bar{x} - \mu$ — разница между максимальным измеренным средним арифметическим $\bar{x}$ и μ. При известном среднеквадратичном отклонении σ случайная величина

$$Z = \frac{\bar{x} - \mu}{\sigma\sqrt{m}} = \frac{E}{\sigma\sqrt{m}}$$

принадлежит стандартному нормальному распределению $Z \sim \mathcal{N}(0, 1)$. Тогда

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{m}},$$

где $z_{\alpha/2}$ таково, что вероятность события $\left\{|Z| \leqslant z_{\alpha/2}\right\}$ равна α. Отсюда получаем формулу для оценки размера выборки

$$m = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2. \tag{10}$$

Если $m \geqslant 30$, можно пользоваться этой формулой, заменив в ней среднеквадратичное отклонение σ на его оценку $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Однако в случае $m \leqslant 30$ для использования этой формулы необходимо, чтобы случайные величины x_i были распределены нормально; кроме того, среднеквадратичное отклонение σ должно быть известно.

Скользящий контроль. Выделим из исходной выборки две непересекающиеся подвыборки одинакового размера и назовем одну из них обучающей X^L, а другую — контрольной X^C. Настроим алгоритм на обучающей выборке: найдем параметры регрессии $\mathbf{w}^L$ с помощью процедуры (6) , выберем набор признаков $\mathcal{A}^L$, используя алгоритм (9).

Затем, используя полученные результаты, решим задачу классификации (7) на обучающей X^L и контрольной выборках X^C. Для каждой из выборок получим ошибки $E^L(\mathbf{w}^L)$ и на контроле $E^C(\mathbf{w}^L)$ и вычислим их отношение $S = \frac{E^L(\mathbf{w}^L)}{E^C(\mathbf{w}^L)}$. Будем добавлять в каждую выборку по элементу и проводить всю процедуру заново. Таким образом, сможем построить график зависимости величины S от объема выборки и, наблюдая за ростом этой величины, определим момент, когда она начинает меняться достаточно плавно. С этого момента считаем, что объем выборки достаточен.

Таблица 4. Число вхождений признаков в K лучших для каждой пары классов.

classes/markers	K	L	K/M	K/N	K/O	L/O	K/P	L/P	K/R
A1 A3	0	5	0	0	0	0	0	0	1
A1 B1	0	0	1	0	0	0	1	0	0
A1 B2	0	0	1	1	1	0	0	1	0
A3 B1	0	3	1	3	2	2	0	3	1
A3 B2	0	5	4	1	0	0	0	5	0
B1 B2	2	5	3	0	0	0	0	3	1

classes/markers	L/R	L/R/SA	L/T/SA	L/T/SO	U/V	U/W	U/X	U/Y	U/Z
A1 A3	0	1	1	0	1	1	0	5	5
A1 B1	2	4	0	2	0	0	4	5	1
A1 B2	4	0	0	5	0	2	0	5	0
A3 B1	0	2	0	0	0	1	1	0	1
A3 B2	0	1	0	0	0	1	1	2	0
B1 B2	0	1	0	0	0	0	3	2	0

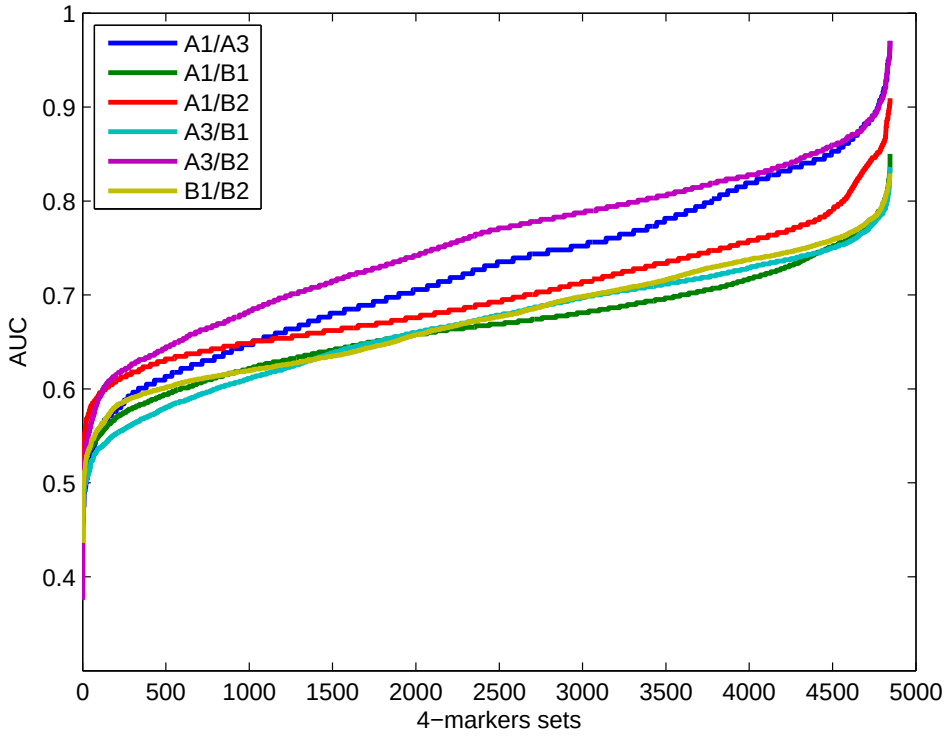

Рис. 1. Для всевозможных наборов из четырех маркеров вычислено значение AUC (для каждой пары класов). На графике отложены эти значения в порядке возрастания. Обратим внимание на то, что значение по оси абсцисс — не более чем порядковый номер значения AUC, не привязанный ни к какому определенному набору маркеров.

Расстояние Кульбака-Лейблера. В задаче восстановления регрессии предполагается, что

$$\mathbf{y} = \mathbf{f}(X, \mathbf{w}),$$

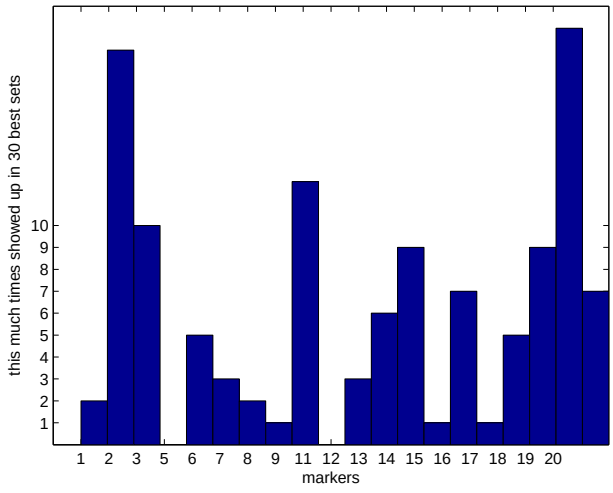

Рис. 2. Количество вхождений каждого из двадцати маркеров в набор «K лучших», $\mathcal{S}$. Например, маркер под номером 2 (L) имеет наибольшую частоту вхождений, следующий за ним — номер 19 (U/Y).

где $\mathbf{f}$ — функция с вектором параметров $\mathbf{w}$, в данной работе это логистическая функция. Покажем, что при биномиальном распределении вектора ответов $\mathbf{y}$ вектор параметров $\mathbf{w}$ распределен нормально. Пусть компоненты $\mathbf{y}$ независимы и $y_i \sim \mathcal{B}(p_i, 1-p_i)$. Рассмотрим некоторую подвыборку исходной выборки. Пусть нам удалось оценить плотность распределения признака на этой подвыборке, назовем ее $p_1(\mathbf{x})$. Удалив из подвыборки один элемент, снова произведем оценку плотности; обозначим полученную функцию $p_2(\mathbf{x})$. Тогда степень «похожести» этих функций будем определять через расстояние Кульбака-Лейблера

$$D(p_1, p_2) = \int_{-\infty}^{+\infty} p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx. \tag{11}$$

Видно, что «расстояние» несимметрично, т.е. $D(p_1, p_2) \neq D(p_2, p_1)$. Если объем выборки достаточен, распределение не должно существенно меняться при малом изменении выборки. Обратное свидетельствует о слишком маленьком объеме выборки. Воспользуемся формулой Байеса для оценки апостериорных вероятностей. Пусть заданы обратные ковариационные матрицы A, B. Тогда при гипотезах о биномиальном распределении вектора ответов $\mathbf{y} \sim \mathcal{B}$ и нормальном распределении вектора параметров $\mathbf{w} \sim \mathcal{N}$, получаем

$$P(\mathbf{w}|D, A, B) = \frac{P(D|\mathbf{w}, B)P(\mathbf{w}|A)}{\int \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)A^T(\mathbf{w}-\mathbf{w}_0)^T}, \tag{12}$$

где D — исследуемые данные, а в $\mathbf{w}_0$ достигается минимум функции ошибки

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)A^T(\mathbf{w}-\mathbf{w}_0)^T + \sum_{i=1}^{m} y_i \ln p_i + (1-y_i)\ln(1-p_i)$$

Заменяя в (11) $p_1(\mathbf{x})$ и $p_2(\mathbf{x})$ на $P(\mathbf{w}|D_1, A, B)$ и $P(\mathbf{w}|D_2, A, B)$ и, учитывая дискретность, распределений получим

$$D(P_1, P_2) = \sum P(\mathbf{w}|D_1, A, B) \ln \frac{P(\mathbf{w}|D_1, A, B)}{P(\mathbf{w}|D_2, A, B)}. \tag{13}$$

Вычислительный эксперимент: классификация и выбор признаков

В этом разделе проводился эксперимент на реальных данных, описанных в разделе «Задача классификации». Пропуски данных были заполнены средними по признакам значениями.

В таблице (3) содержатся следующие результаты: для каждой пары классов указаны K наборов маркеров, давших наибольшие значения максимизируемого критерия AUC и сами значения этого критерия, а также ошибки первого Err_1 (*false negative rate*) и второго Err_2 (*false positive rate*) рода. Здесь число K определяется визуально, по графику 1. На этом графике изображены зависимости значения AUC (в порядке возрастания), полученные при классификации на различных наборах из четырех признаков. Предлагается выбирать такое число K, при котором рост графика меняется еще достаточно сильно. В данном случае, было выбрано значение $K = 5$.

Таким образом, для j-той пары классов найдено некоторое множество оптимальных наборов $\mathcal{S}_j = \bigcup_{i=1}^{K} \{\mathcal{A}_{j_i}\}, j = 1, \ldots, 4$. Объединив признаки из всех наборов из колонки «best

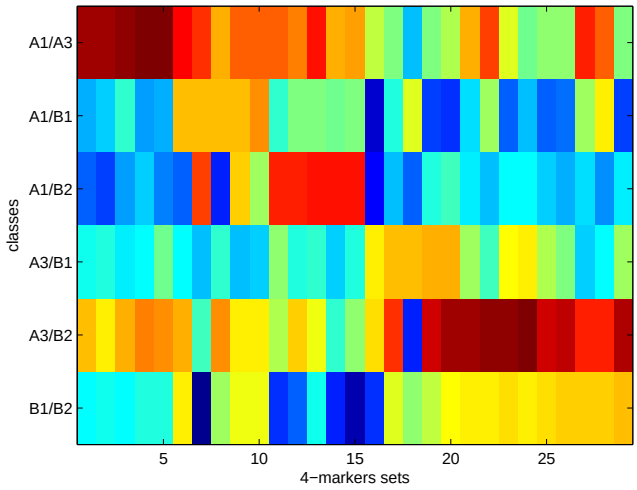

Рис. 3. Для каждого из 30 полученных наборов наиболее информативных признаков приведено его значение AUC (для каждой пары классов).

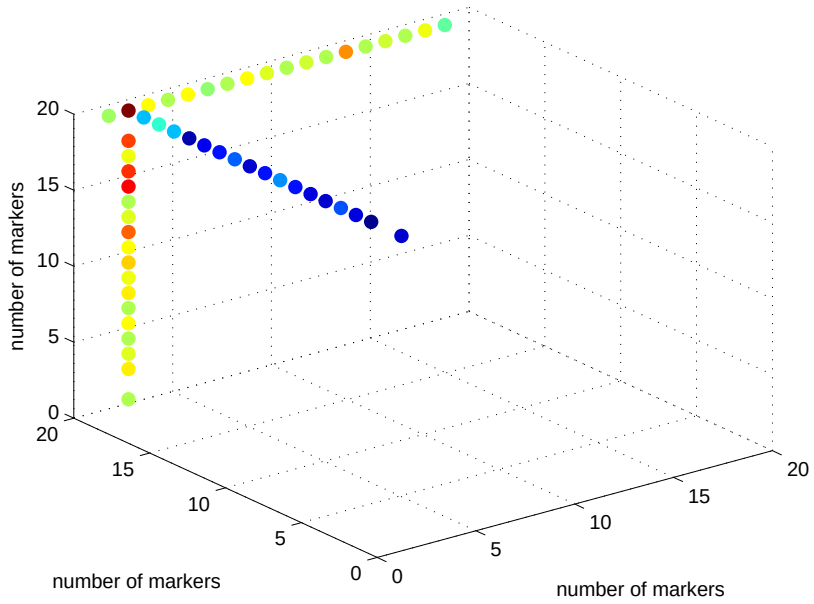

Рис. 4. Оптимальный набор маркеров (2, 19, 20), получаемый при ограничении на сложность модели, равном трем, в окрестности неоптимальных наборов. Чем теплее цвет, тем больше значение оптимизируемого функционала.

sets» таблицы (3), получим некоторое множество наиболее информативных признаков. Для каждого из них можно подсчитать количество его вхождений в наборы из «best sets» и затем построить гистограмму 2, показывающую, насколько часто каждый признак входит в K лучших наборов. Таким образом, гистограмма 2 характеризует степень качества каждого признака по отдельности. Чтобы сравнить между собой наборы признаков, построим таблицу 3. Здесь снова было рассмотрено множество $\mathcal{S}$ всех наборов, вошедших в

K лучших хотя бы для одной пары признаков: $\mathcal{S} = \bigcup_{j=1}^{4} \mathcal{S}_j$. Для всех пар классов проведем классификацию на каждом из этих наборов и сведем полученные значения критерия AUC в таблицу 3. Здесь более теплым тонам соответствуют большие значения AUC, более холодным — меньшие. Таким образом, можно наблюдать ступенчатую структуру таблицы.

Продемонстрируем «оптимальность» найденного набора (на примере A_1 и A_3 и при ограничении на сложность модели, равном трем) с помощью рисунка 4. Каждому набору из дискретной окрестности оптимального (то есть такому набору, у которого от оптимального отличается лишь один элемент) можно сопоставить величину максимизируемого функционала. На рисунке 4 эта величина выражается цветом точки — чем теплее цвет, тем больше значение.

Заключение

В работе проведен поиск наиболее информативных признаков для классификации пациентов на четыре группы с точки зрения наличия нарушений работы сердечно сосудистой системы. При этом задача многоклассовой классификации сводилась к двуклассовой классификации путем рассмотрения всевозможных пар групп. Для каждой из таких пар получен оптимальный набор признаков. Отбор признаков производился на основе полного перебора, преимуществом которого перед другими алгоритмами выбора признаков является его наглядность.

Литература

[1] Bishop C. M. *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] Perez-Cruz F., *Kullback-Leibler Divergence Estimation of Continuous Distributions*, 2008.

[3] Azuaje F., Devaux Y. & Wagner D., *Computational biology for cardiovascular biomarker discovery*, http://bib.oxfordjournals.org/content/10/4/367.abstract

[4] Hosmer D. & Lemeshow S., *Applied logistic regression*, 2000.

[5] MacKay D. J. C., *Information Theory, Inference, and Learning Algorithms*, 2003.

[6] Friedman J., Hastie, Tibshirani R., *Additve logistic regression: a statistical way of boosting*, The Annals of Statistics, 28(2):337-407, 2000.

[7] Breton M. H., Stuart D. Russell, Michelle M. Kittleson, Kenneth L. Baughman, Heidecker J. B., Edward K. Kasper, Ilan S. Wittstein, Hunter C. Champion, Elayne. *Transcriptomic Biomarkers for Individual Risk Assessment in New-Onset Heart Failure.* Circulation, 2008.

[8] Bos S., *How to partition examples between cross-validation set and fraining set?* Labaratory for information representation RIKEN, Hirosawa 2-1, Wako-shi, Saitama, 351-01, Japan.

[9] Реброва О. Ю. *Статистический анализ медицинских данных. Применение прикладного пакета STATISTICA*, 2006.

[10] Fawcet T., *ROC Graphs: Notes and Practical Considerations for Researchers*, HP Laboratories, 2004.