УДК 681.513.8

Н.В. Кондрашова, В.А. Павлов, А.В. Павлов

Решение задачи медицинской диагностики с применением линейного дискриминантного анализа и МГУА

Рассмотрена постановка задачи классификации МГУА и линейного (канонического) дискриминантного анализа для медицинской дифференциальной диагностики. Представлены результаты их сравнительного анализа. Описан новый алгоритм классификации на основе быстродействующего рекуррентного алгоритма МГУА, минимизирующий критерий уравновешенных ошибок первого и второго рода.

The classifying problem statement of the GMDH and a the linear discriminant analysis (LDA) for the medical differential diagnosis is considered. A new classification algorithm based on the fast-acting recurrent GMDH algorithm and the minimization of criterion balanced errors of the first and second kind is described. A comparative analysis of the GMDH and LDA is shown.

Розглянуто постановку задачі класифікації МГУА та лінійного (канонічного) дискримінантного аналізу для медичної диференційної діагностики. Наведено результати їх порівняльного аналізу. Описано новий алгоритм класифікації на основі швидкодіючого рекурентного алгоритму МГУА, що мінімізує критерій збалансованих помилок першого і другого роду.

Введение. С ростом объемов накопленных медицинских информационных баз необходимы новые методы, алгоритмы и программные средства интеллектуального анализа данных для построения моделей сложных процессов и систем. Создание простого и удобного в использовании компьютерного советчика для врача значительно облегчит его работу. В настоящее время нам не известны эффективные технологии, которые бы давали возможность врачу комплексно подходить к решению задач анализа клинических данных и дифференциальной диагностики заболеваний. Поэтому главная цель статьи – повышение точности медико-информационной системы оценивания текущего состояния пациентов для эффективной дифференциальной диагностики, за счет применения известных современных подходов и совершенствования разработанных алгоритмов.

Методы самоорганизации моделей позволяют автоматически находить ранее неизвестные функциональные зависимости, заложенные в выборке данных, и, следовательно, открывать новые знания [1, 2]. В работе [3] была разработана система классификаторов для решения задачи дифференциальной диагностики на основе алгоритма МАКСО метода группового учета аргументов (МГУА). При решении задачи классификации диагнозов результаты, полученные по МГУА, оказались наиболее точным в сравнении с вероятностными методами, использующими идеи Байеса, Борда и Кондорсе [4, 5]. Одним из наиболее известных, развиваемых в последнее время направлений в классификации есть дискриминантный анализ [6].

Реализованный в современных пакетах канонический дискриминантный анализ (КДА) в равной степени можно отнести как к индуктивному, так и к вероятностному подходу, так как он основан на максимизации отношения межгрупповой дисперсии к внутригрупповой, использует шаговые методы регрессионного анализа, критерий внешнего дополнения в виде критерия Фишера для отбора признаков, а также априорные знания о ковариационных матрицах и функциях распределения и др. Для КДА не есть принципиальным, являются модели (дискриминантные функции) линейными или нелинейными по входным переменным, поскольку нелинейные преобразования осуществляются на стадии предобработки данных. Проблема состоит в том, что матрицы даже средних размеров достаточно долго импортируются, не говоря о больших размерностях. Например, матрица из 300 наблюдений и 250 переменных в одном из пакетов импортируется около 30 мин. Этот технический недостаток в доступных пакетах с КДА, вероятно, со временем будет преодолен. В качестве современной пакетной реализации МГУА для построения системы классификаторов используется быстродействующий обобщенный релаксационный итерационный алгоритм (ОРИА) («не зависящий» от числа наблюдений), в котором автоматизированы: пред-

обработка данных, оптимальный отбор признаков, а импорт данных занимает доли секунды [7]. ОРИА ориентирован на работу с выборками большой размерности (более десяти тысяч переменных). Отметим, что в известных пакетах, реализующих КДА, результат классификации рассчитывается по минимальному расстоянию Махаланобиса от точки наблюдения до центра искомого класса. При этом, чем больше признаков, тем больше расстояние Махаланобиса, так как оно обобщает Евклидово расстояние. В работе [8] для определения оптимального состава признаков дискриминантной функции вместо шаговой процедуры оптимизации набора признаков по критерию Фишера применяется принцип МГУА – минимизации критерия внешнего дополнения, основанного на разбиении выборки. Этот специальный критерий сконструирован на основе оценки расстояния Махаланобиса и используется при оценке близости к классу.

В настоящей статье результаты канонического дискриминантного анализа применяются с целью повышения точности классификации индуктивно-вероятностного подхода. Индуктивный подход усилен разработкой нового алгоритма классификации на основе быстродействующего РИА МГУА, описанного в [9]. Особенность предлагаемого алгоритма классификации – использование в системе классификаторов для автоматической минимизации уравновешенных ошибок первого и второго рода критерия, на основе меры Ван Ризбергена, известной также под названием *F*-оценки (*F-measure*).

Статья посвящена повышению точности системы дифференциальной диагностики легких случаев патологий гемостаза или, иначе, коагулопатии и тромбоцитопатии (КиТ), основываясь на данных о возрасте и геморрагических признаках. К легким случаям патологии гемостаза относят: болезнь Виллебранда, коагулопатию, дезагрегационную тромбоцитопатию и комбинированную патологию системы гемостаза [10]. Все они связаны с несворачиваемостью крови человека. Для проведения диагностики врачи руководствуются своим опытом и клиническими признаками. Однако достоверно определить диагноз врачу сложно в связи со схожей, трудно различимой клинической картиной. Кроме того, не всегда наличие геморрагических признаков связано с болезнью [11].

Установлено, что распространенность легких форм КиТ составляет для болезни Виллебранда 1,9 процентов, дезагрегационных тромбоцитопатий – 1,6; коагулопатий – 0,9, комбинированной патологии системы гемостаза– 0,6 [12]. Диагностирование при помощи дорогостоящих химико-биологических тестов доступно не всем нуждающимся в таком обследовании. Поэтому актуальна задача дифференциальной диагностики (классификации) пациентов по заданным признакам с целью снижения рисков неправильного установления диагноза и экономии расходов на лабораторные тесты.

Постановка задачи

По результатам лабораторного тестирования известно разделение на группы больных, объединенных одним диагнозом. В отсутствии лабораторных исследований разделение затруднено из-за большого числа признаков и значительной схожести клинических проявлений у больных, принадлежащих различным группам диагнозов.

Пусть в пространстве клинических признаков x_i, $i = 1, \ldots, m$ нужно распознать четыре диагноза (класса): 1 – болезнь Виллебранда (БВ), 2 – дезагрегационная тромбоцитопатия (ДТ), 3 – коагулопатия (КП), 4 – комбинированная патология системы гемостаза (КПСГ).

В женской группе соответствующего возрастного интервала наблюдаются следующие признаки ($m = 12$): 1 – носовое кровотечение (НК), 2 – кровоточивость десен (КД), 3 – кровотечение после экстракции зубов (КПЕЗ), 4 – интра и послеоперационное кровотечение (ПОК), 5 – послетравматическая гематома (ПТГ), 6 – кровотечение из поверхностных ран (КПР), 7 – продолжительное не заживление ран (ПНЗ), 8 – желудочно-кишечное кровотечение (ЖКК), 9 – послеинъекционная гематома (ПИГ), 10 – послеродовое кровотечение (ПРК), 11 – ювенальное маточное кровотечение (ЮМК). Возраст – двенадцатый признак. Фрагмент данных представлен в табл. 1.

Таблица 1. Геморрагические признаки у женщин в возрасте от 19 до 49 лет

П–т №	Д–з	Клинические признаки											
		НК	КД	КПЭЗ	ПОК	ПТТ	КПР	ПНЗ	ЖКК	ПИГ	ПРК	ЮМК	В–т
70	ДТ	+	+	+	+	–	–	–	–	–	1	+	38
79	ДТ	+	+	+	+	+	+	+	+	–	+	+	43
75	КП	+	+	+	1	+	–	–	–	–	1	+	21
35	ДТ	+	+	+	1	+	–	–	–	–	1	+	21
74	БВ	+	+	+	+	+	+	+	–	–	1	+	49
1	БВ	+	–	–	–	+	+	+	–	–	–	+	19
22	КП	+	+	–	1	+	+	+	–	–	–	+	20
52	ДТ	+	+	+	1	+	–	–	–	–	–	–	49
10	БВ	+	+	+	1	+	–	–	–	–	–	+	31
30	КП	+	–	+	–	+	+	+	–	–	1	+	49
42	ДТ	+	+	+	1	+	+	+	–	–	+	+	28
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Обозначение: П-т – пациент, Д-з – диагноз, В-т – возраст, + означает наличие признака, – означает отсутствие признака, 1 – не было условий для проявления признака.

Примечание. Автор исходных данных д.м.н. Томилин В.В., ГУ «Институт гематологии и трансфузиологии» АМН Украины.

Задача заключается в том, чтобы

- найти «дифференцирующие» признаки;
- по комплексу таких признаков построить «разделяющую» (дискриминантную) функцию, выполнив при этом условия анализа:
  - наличие тестовых выборок групп;
  - наличие контрольных выборок групп

и требования к данным:

– слабая коррелированность признаков,

– сходство ковариационных матриц тестовой и контрольной выборок, совпадающих у обоих методов: МГУА и КДА.

Требование слабой коррелированности – выполнено, так как парная корреляция признаков варьирует в пределах от 0,0 до 0,3; сходство ковариационных матриц может быть выполнено, если предварительно применить квазиоптимальное разбиение выборок [13].

Численным экспериментом установлено, что точность единого классификатора на четыре класса существенно ниже, чем точность системы четырех бинарных классификаторов. Поэтому построим функцию для каждого из классификаторов $\ell = \overline{1,4}$ по принципу отделения больных заданного класса (диагноза) от всех остальных:

$$\psi_\ell(\Theta, \mathbf{X}) = 0, \qquad (1)$$

где $\mathbf{X}$ – матрица признаков; Θ – вектор параметров, ψ_ℓ – функция линейная по параметрам. Функция (1), называемая решающим правилом, должна удовлетворять:

$\psi_\ell(\Theta, \mathbf{X}) \geq 0$ – для пациентов одного из диагнозов и

$\psi_\ell(\Theta, \mathbf{X}) < 0$ – для остальных пациентов.

Постановка задачи классификации по МГУА

Задано множество наблюдений $\mathbf{v} \in W \subset \Re^m$, которые являются вектор–строками матрицы данных $\mathbf{X}$, $\dim \mathbf{X} = n_W \times m$, вектор–столбцы признаков которой – $\mathbf{x}_i$, $i = 1,..., m$; m – количество признаков; n_W – количество наблюдений. Заданы также векторы выхода $\mathbf{y}_\ell$, $\dim \mathbf{y}_\ell = n_W \times 1$, элементы $y_{\ell,i}$, $i = \overline{1, n_W}$ которого задают для различных наблюдений функцию принадлежности к классам $\ell = \overline{1,4}$ значениями y^+, y_-, соответственно: пациент принадлежит или не принадлежит классу. Классификация проводится по принципу отделения каждого диагноза от всех остальных. Поскольку дальнейшие результаты применяются к каждому из получаемых классификаторов, индекс класса ℓ в данном разделе и последующих двух опускается.

Предполагается, что $y_i \in \Re$ – сумма значений неизвестной функции $\breve{y}(\mathbf{v}_i)$ $i = \overline{1, n_W}$ и случайной величины ξ_i с нулевым математическим ожиданием и ограниченной дисперсией. Случайные величины ξ_l, ξ_i есть независимыми, их ковариации $\boldsymbol{\xi}_i^T \boldsymbol{\xi}_l = 0$, $i \neq l \in [1,2,...,n_W]$, ξ_i также не зависят от значений функции $\breve{y}(\mathbf{v})$, $\mathbf{v} \in W$; $\boldsymbol{\xi}_l$, $\boldsymbol{\xi}_l$ – векторы, $\dim \boldsymbol{\xi}_i = \dim \boldsymbol{\xi}_l = t \times 1$.

В общем случае матрица признаков может содержать не только наблюдаемые признаки, но и их нелинейные преобразования. Предполагается, что функции преобразований $f_j(\mathbf{v}) \in \Phi$, $f_j : \Re^m \to \Re$, где $\Phi = \{f_j(\mathbf{v})\}_{j=\overline{1,J}}$ – заданное конечное множество функций.

Рассмотрим модели $\breve{y}(\mathbf{v})$ линейные по переменным, т.е. $f_j(\mathbf{v}) = v_j$, $j = \overline{1,m}$, где $v_j \in \Re$ – элемент вектор–строки $\mathbf{v} \in W$. Функция $\breve{y}(\mathbf{v})$ аддитивна по исходным переменным и линейная по параметрам $\theta_{j,k}$.

$$\breve{y}(\mathbf{v},\Theta_k)=\theta_{0,k}+\sum_{j=1}^{m} v_j\theta_{j,k},\ \mathbf{v}\in\Re^m,$$
$$\Theta_k\in\Re^K,\ k=\overline{1,K}. \qquad (2)$$

Векторы неизвестных параметров Θ_k, $k=\overline{1,K}$, $\dim\Theta_k=(m+1)\times 1)$ отличаются между собой тем, что разные их компоненты принудительно принимают нулевые значения. $\mathbf{d}_k$ – бинарный вектор, определяющий подмножество аргументов модели. s_k – число ненулевых компонент вектора $\mathbf{d}_k$. D – множество всех возможных бинарных векторов размерности $m+1$, мощность множества $|D|=K$, Θ_k – вектор, имеющий ненулевые компоненты вектора Θ, $\Theta_k=diag(\mathbf{d}_k)\cdot\Theta$. $diag(\mathbf{d}_k)$ – диагональная матрица, $\dim[diag(\mathbf{d}_k)]=(m+1)\times(m+1)$, диагональю которой есть вектор $\mathbf{d}_k$; некоторые элементы вектора $\mathbf{d}_{kj}=1$, а остальные компоненты вектора $\mathbf{d}_k$ – нулевые.

Векторный вид моделей (2):

$$\breve{\mathbf{y}}_k(\mathbf{X})=\breve{\mathbf{X}}\,\Theta_k=\boldsymbol{\theta}_{0,k}+\sum_{j=1}^{m}\theta_{j,k}\mathbf{x}_j,\ k=\overline{1,K},$$

где $\breve{\mathbf{y}}(\mathbf{X},\Theta_k)$ – вектор–столбец, $\dim\breve{\mathbf{y}}(\cdot)=n_W\times 1$; матрица $\breve{\mathbf{X}}=(\mathbf{1}\vdots\mathbf{X})$ является расширением слева матрицы $\mathbf{X}$ на вектор–столбец из единиц, $\dim\breve{\mathbf{X}}=n_W\times(m+1)$; $\boldsymbol{\theta}_{0,k}$ – вектор–столбец свободных членов модели, $\dim\boldsymbol{\theta}_{0,k}=n_W\times 1$; $\mathbf{x}_j$ – вектор–столбец матрицы $\mathbf{X}$, $\dim\mathbf{x}_j=n_W\times 1$. Элементы v_i, $i=\overline{1,m}$ вектора $\mathbf{v}$ в функции $y(\mathbf{v},\Theta_k)$ содержат значения признаков, наблюдаемых у пациентов исходной выборки. Элементы $\theta_{i,k}$, $i=\overline{0,m}$ вектора Θ будем называть параметрами модели. Структурой модели назовем функцию $y(\mathbf{v},\Theta_k)$, заданную с точностью до значений вектора параметров Θ_k, $k=\overline{1,K}$. $K=2^m$ при полном переборе вариантов структур модели. Сложность структуры модели s_k определяется количеством компонент вектора Θ_k, принимающих ненулевые значения: чем больше ненулевых компонент, тем сложней модель.

Цель МГУА – найти оптимальную модель $y^*=f^*(\mathbf{v},\widehat{\Theta}^*)$ – результат решения следующих двух задач.

Поиск оценок параметров $\widehat{\Theta}_k$, $k=\overline{1,K}$ $\forall\ \mathbf{d}_k\in D$ при решении K задач непрерывной оптимизации:

$$\widehat{\Theta}_k=\arg\min_{\Theta_k\in\Re^{S_k}} QR(y(\mathbf{v}),\breve{y}(\mathbf{v},\Theta_k)),$$
$$\mathbf{v}\in A\subseteq W,\ k=\overline{1,K}, \qquad (3)$$

где A – обучающая выборка – множество наблюдений, используемое для оценивания параметров Θ_k, $k=\overline{1,K}$.

Поиск оптимальной модели при решении K задач дискретной оптимизации:

$$y^*(\mathbf{v},\widehat{\Theta}^*)=\arg\min_{\mathbf{d}_k\in D,\ k=\overline{1,K}} CR(y(\mathbf{v}),\breve{y}(\mathbf{v},\widehat{\Theta}_k)),$$
$$\mathbf{v}\in B\subseteq W, \qquad (4)$$

где B – проверочная выборка – множество наблюдений, используемое для оценивания качества модели $\breve{y}(\mathbf{v},\widehat{\Theta}_k)$ по критерию CR.

Критерий QR – критерий качества для оценок параметров $\widehat{\Theta}_k$, CR – критерий качества модели. В качестве критерия QR используется остаточная сумма квадратов:

$$RSS_A=\left(\mathbf{y}_A-\mathbf{X}_A\widehat{\Theta}_{A,k}\right)^2,\ k=\overline{1,K}.$$

В качестве внешнего критерия CR в МГУА используется *критерий регулярности*:

$$AR_{A/B}=\left(\mathbf{y}_B-\mathbf{X}_B\widehat{\Theta}_{A,k}\right)^2,\ k=\overline{1,K}.$$

Полностью заданную функцию $y(\mathbf{v},\widehat{\Theta}_k)$ называют решением или моделью, а *функцию принадлежности* $y^*(\mathbf{v},\widehat{\Theta}^*)$ – оптимальным решением задачи (оптимальной моделью).

Задача (3) – (4) называется задачей построения структуры и параметров функции принадлежности классификатора МГУА.

Оценивание результатов дифференциальной диагностики и информационного поиска

Построим систему дифференциальной диагностики, объединяющую несколько классификаторов, каждый из которых выделяет пациентов одного заданного диагноза. Классификаторы разбиения на два класса назовем бинарными. Далее используемые термины и понятия: истинно положительный результат (*True Positive*, *TP*); ложноположительный результат: ошибка

первого рода – «пропуск цели» (*False Positive*, *FP*); ложноотрицательный результат: ошибка второго рода – «ложная тревога» (*False Negative*, *FN*); истинно отрицательный результат (*True Negative*, *TN*); чувствительность (*Sensitivity*, *Se*); специфичность (*Specificity*, *Sp*); положительное прогностическое значение (*Positive Predictive Value*, *PPV*); отрицательное прогностическое значение (*Negative Predictive Value*, *NPV*) обычно представляют табл. 2.

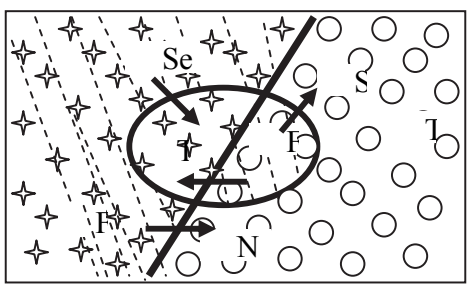
Таблица 2. Результаты теста в зависимости от состояний пациентов, определенные «Золотым стандартом»

		Состояния, определенные «Золотым стандартом»		
		Положительное состояние	Отрицательное состояние	
Результат теста	Положительный результат теста	Истинно положительный результат, *TP*	Ложноположительный результат, *FP* (ошибка I рода)	Положительное прогностическое значение, *PPV*
	Отрицательный результат теста	Ложноотрицательный результат, *FN* (ошиб. II рода)	Истинно отрицательный результат, *TN*	Отрицательное прогностическое значение, *NPV*
		Чувствительность, $Se = \Sigma$ Истинно положительных / Σ положительных состояний	Специфичность, $Sp = \Sigma$ Истинно отрицательных / Σ отрицательных состояний	

Между отсеиванием больных, имеющих заданный диагноз, и поиском полезной информации существует аналогия. Воспользуемся понятием релевантности, принятым в информационном поиске. В информационном поиске релевантность оценивается точностью и полнотой. В диагностике точности (*Precision*, *P*) соответствует *PPV*, а полноте (*Recall*, *R*) соответствует чувствительность:

$$P = PPV = \frac{TP}{TP + FP}, \quad R = Se = \frac{TP}{TP + FN}.$$

Пусть построенный классификатор разделил классы, как показано на рисунке. Здесь релевантные точки (пациенты с заданным диагнозом), обозначенные крестиками, находятся слева от прямой, а точки, найденные классификатором, находятся в овале. Заштрихованные пунктиром области представляют ошибки классификатора. Заштрихованная область слева – это релевантные точки, не найденные классификатором (пропуск цели), заштрихованная область справа – найденные, но нерелевантные точки (обозначены кружочками) (ложная тревога). Нерелевантные точки – это не имеющие отношения к заданному диагнозу пациенты (в дифференциальной диагностике «здоровые» соответствуют прочим диагнозам).



Множества и области принадлежности наблюдений (пациентов) при классификации на два класса

Специфичность – вероятность правильной классификации «здоровых» людей среди всех «здоровых», имеющихся в выборке. В информационном поиске специфичность называется выпадением (*fall-out*) и характеризует вероятность нахождения нерелевантного ресурса:

$$Sp = fall - out = \frac{\grave{O}N}{TN + FP}.$$

На рисунке число, определенных классификатором «действительно здоровых» пациентов – это сумма кружочков, не вошедших в овал, а «здоровые», отнесенные к больным – это кружочки, заключенные в овале.

Точность – это пропорция левой незаштрихованной области по отношению ко всем точкам, заключенным в овале (горизонтальная стрелка), или – это точность положительных (внутри овала) ответов классификатора. Чувствительность (полнота) – это пропорция левой незаштрихованной области к области слева от прямой (диагональная стрелка). *NPV* так же, как *PPV*, является точностью тестирования, но это точность отрицательных ответов классификатора (вне овала):

$$NPV = \frac{TN}{TN + FN}.$$

Критерии для системы дифференциальной диагностики

При скрининге величины *PPV* и *NPV* могут значительно отличаться, так как отличаются в несколько десятков раз ошибки I и II рода. Можно привести пример, в котором при скрининговом тестировании отрицательный результат (*NPV)* очень хорош. Например, только один из 10 тыс. может получить ложный диагноз (значение *NPV* близко к единице). В то же время при относительно большом количестве ложнополо-

жительных (ошибок I рода) результатов и небольшом – истинно положительных, результат положительного скрининг-теста может быть плохим при подтверждении данного диагноза. Например, из 30 больных, имеющих заданный диагноз, только двадцати – установили правильный диагноз, т.е. *PPV*=2/3. Поэтому полезно объединять качество положительных и отрицательных ответов, ошибок первого и второго рода.

В положительном тесте воспользуемся известной в информационном поиске *F*-оценкой:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}, \qquad (5)$$

объединив *PPV* и *Se* в одной усредненной величине. *F*-оценка, являясь средним гармоническим точности *P* (*PPV*) и полноты *R* (*Se*), при большом отличии усредняемых значений приближается к минимальному из них (среднее арифметическое при этом приближается к значению, равному 1/2). Если точности и полноте придается одинаковый вес ($\alpha = 0{,}5$), то имеет место сбалансированная *F*-оценка (F_1-*measure*). Введем другое обозначение для этой величины, так как литерой *F* обычно обозначается критерий Фишера. Следующие величины: *Se* = *R*, *Sp* = *fall-out*, *PPV* = *P*, *NPV*, *F* – относительные частоты (вероятности). Поскольку для классификатора в качестве *советчика врача* важна способность из общего потока людей выделять больных, устанавливая им правильный диагноз, то назовем оценку такой способности – результатом позитивного тестирования – позитивным шансом (*Positive Chance* – *PC*):

$$PC = \frac{2P\cdot R}{P+R} = \frac{2TP}{2TP+FP+FN},\ PC \in [0,1]. \qquad (6)$$

Способность отсеивать «здоровых» будем оценивать результатом тестирования, который назовем альтернативным шансом – *Alternative Chance* (*AC*):

$$AC = \frac{2TN}{2TN+FP+FN},\ AC \in [0,1], \qquad (7)$$

где *TP*, *FP*, *FN*, *TN* – обозначают мощности множеств или общее число точек (пациентов), попавших в соответствующие области (см. рисунок). Поскольку *FP*, *FN* обозначают число ошибок соответственно первого и второго рода, то, если придавать равный вес не только *Se* и *PPV*, но и этим ошибкам (т.е. $FP = FN = \Phi$), формулы (6) и (7) упрощаются:

$$PC_e = \frac{TP}{TP+\Phi},\quad AC_e = \frac{TN}{TN+\Phi}.$$

Если оценка *PC* представляет интерес для врача, то оценка *AC* интересна прежде всего пациентам, многие из которых с предубеждением относятся к электронным советчикам и трудно переживают последствия постановки ложных диагнозов. При создании *советчика врача* имеет смысл использовать *PC* в качестве критерия, который необходимо максимизировать для того, чтобы получить наиболее точный классификатор диагноза. Увеличением *AC* минимизируется моральный и материальный ущерб *здоровому* человеку от неверной постановки диагноза.

Аналогично формуле (5) можно ввести следующие целесообразные формы критериев:

$$S = \frac{1}{\beta\frac{1}{Se} + (1-\beta)\frac{1}{Sp}},\quad \beta \in [0,1],$$

$$\text{а при } \beta = 0{,}5 \quad S_1 = \frac{2SeSp}{Se+Sp},$$

$$T = \frac{1}{\gamma\frac{1}{PPV} + (1-\gamma)\frac{1}{NPV}},\quad \gamma \in [0,1],$$

$$\text{а при } \gamma = 0{,}5 \quad T_1 = \frac{2PPV\cdot NPV}{PPV+NPV}.$$

Особенности построения МГУА-классификаторов

После того как построены функции принадлежности $y_\ell^*(\mathbf{v}, \widehat{\Theta}^*)$, $\ell = \overline{1,4}$, необходимо провести границы, разделяющие классы – отделяющие больных каждого из диагнозов от всех остальных. Для построения решающих правил:

$$\psi_\ell(\Theta, \mathbf{v}_j) = y_\ell(\Theta, \mathbf{v}_j) - A_\ell = 0,\ j = \overline{1, n_W}$$

необходимо найти пороговое значение A_ℓ. Оно определяется с помощью максимизации критерия (5) при минимизации равнозначных ошибок первого и второго рода.

Задается номер класса (диагноза) ℓ=1.

Процедура определения порога ℓ-го классификатора.

Мощности множеств: ошибок первого рода $|FP| = 0$, ошибок второго рода $|FN| = 0$ и $|\Sigma| = 0$; $q = 0$ полагаются равными нулю.

Ш а г 1. По данным тестовой выборки находится $y_{\min_\ell} = \min_i y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*)$, $\forall i \in \Omega_\ell$, а также $y_{\max_\ell} = \max_i y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*)$, $\forall i \notin \Omega_\ell\ i \in \Omega_\ell$, где Ω_ℓ – множество пациентов, имеющих диагноз ℓ.

Ш а г 2. Если $y_{\min_\ell} - y_{\max_\ell} \geq 0$, то вывод значений $\Phi = q$; $A_\ell = 0{,}5(y_{\min_\ell} - y_{\max_\ell})$, повторить *процедуру определения порога ℓ-го классификатора* для $\ell = 2; 3; 4$.

Иначе, если $|\Sigma| = 2q$, то $|FN| := |FN| + 1$; $q := q + 1$; $|\Sigma| := |\Sigma| + 1$ точка i, соответствующая $y_{\max_\ell}$, исключается из множества Ω_ℓ, иначе $|FP| := |FP| + 1$; $|\Sigma| := |\Sigma| + 1$, точка (пациент) i, соответствующая значению $y_{\min_\ell}$, исключается из множества Ω_ℓ, где $|\bullet|$ означает мощность множества, а := знак присвоения. Перейти к выполнению шага 1.

Решающее правило имеет вид:

Если $(y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*) - A_\ell) \geq 0$, то i-й пациент имеет диагноз ℓ.

Если $(y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*) - A_\ell) < 0$, то $i \notin \Omega_\ell$.

В ходе решения этой задачи используется понятие близости пациента к классу (диагнозу). Классификация реализуется через процедуру отнесения пациента к одному из классов диагнозов в зависимости от значений функций принадлежности и порогов.

В системе бинарных классификаторов возможны ситуации «конфликта» и отказа от распознавания. Для их устранения решающее правило для i-го пациента имеет вид: $i \in \Omega_{\ell^*}$, где Ω_{ℓ^*} – множество пациентов, имеющих диагноз ℓ^*, где

$$\ell^* = \arg\max_{\ell=\overline{1,4}} (y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*) - y_-).$$

Это означает, что пациент i принадлежит классу ℓ^*, если при подстановке вектора его признаков $\mathbf{v}_i$ в соответствующую модель (функцию принадлежности) $y_\ell^*(\mathbf{v}_i, \widehat{\Theta}^*)$, ее значение максимально отстоит от значения y_-, означающее, что пациент не принадлежит к классу (y_- одинаковы у всех классификаторов), а значения признаков принадлежности к классу в различных классификаторах равны y^+.

Канонический дискриминантный анализ

КДА иначе называют множественным или линейным дискриминантным анализом многомерных данных K классов (*Linear Discriminant Analysis*), частным случаем которого для двух классов есть линейный дискриминант Фишера (*Fisher's Linear Discriminant*). К преимуществам применения КДА следует отнести возможность анализа при:

- большой размерности пространства исходных признаков;
- не выполнении условия равенства ковариационных матриц различных классов;
- отсутствии требования нормального многомерного распределения;
- плохой обусловленности матрицы ковариаций исходного пространства признаков.

Вначале набор $\mathbf{x}_1, ..., \mathbf{x}_m$ (исходных признаков) оптимизируется с помощью шаговой процедуры, $\dim \mathbf{x}_j = n_W \times 1$, $j = \overline{1, m}$. Далее получают уменьшенной размерности $q \leq m - 1$ новое пространство ортогонализованных признаков – канонических дискриминантных функций (КДФ) – $\mathbf{z}_i(x)$, $i = \overline{1, q}$, $\dim \mathbf{z}_i = n_W \times 1$, используя максимизацию критерия:

$$\lambda = |\mathbf{V}^T S_B \mathbf{V}| / |\mathbf{V}^T S_I \mathbf{V}| = |\tilde{S}_B| / |\tilde{S}_I|,$$

где $\mathbf{V}$ – матрица проектирования векторов из пространства признаков в пространство КДФ. $\dim V = m \times m$. S_B, $\widetilde{S}_B$ – матрицы межгруппового разброса в пространстве признаков и в пространстве КДФ. $\dim S_B = m \times m$, $\dim \widetilde{S}_B = m \times m$, причем ранг матрицы $\widetilde{S}_B$ равен q. Здесь $|\bullet|$ обозначает детерминант матрицы. Аналогично

S_I, $\tilde{S}_I$ – матрицы внутригруппового разброса в исходном и преобразованном пространствах. КДА вычисляет проекцию в подпространство признаков таким образом, чтобы минимизировать внутриклассовый и максимизировать межклассовый разброс проекций векторов.

λ – аналог отношения межгрупповой дисперсия $\mathbf{V}^T S_B \mathbf{V}$ к внутригрупповой – $\mathbf{V}^T S_I \mathbf{V}$. Для двумерного случая, когда $\mathbf{v}$ – двумерный собственный вектор, λ совпадает с единственным собственным значением, будучи отношением межгрупповой дисперсии $\mathbf{v}^T S_B \mathbf{v}$ к внутригрупповой – $\mathbf{v}^T S_I \mathbf{v}$.

В уменьшенной размерности $q \le m-1$ пространства векторов КДФ $\mathbf{Z} = \mathbf{V}^T \mathbf{X}$ с компонентами $\mathbf{z}_i$, $i = \overline{1,q}$ можно более точно оценить отдельные ковариационные матрицы для каждого класса, использовать допущение об общем нормальном многомерном распределении, что невозможно было в исходном пространстве признаков (в силу его большой размерности). Становятся более эффективными процедуры расчета наибольшей вероятности класса для наблюдаемого объекта или расстояния Махаланобиса M:

$$\min M^2(\mathbf{z}-\overline{\mathbf{z}}_\ell) = (\mathbf{z}-\overline{\mathbf{z}}_\ell)^T \Sigma_\ell^{-1} (\mathbf{z}-\overline{\mathbf{z}}_\ell),\ \ell = \overline{1,L} \quad (8)$$

для определения принадлежности к классу некоторого наблюдения. Здесь Σ_ℓ^{-1} – матрица, обратная ковариационной матрице КДФ, $\dim \Sigma_\ell^{-1} = q \times q$.

Центр ℓ-го класса определяется как вектор средних значений: $\overline{\mathbf{z}}_\ell = \left(\overline{z}_{\ell,1}, \overline{z}_{\ell,2}, ..., \overline{z}_{\ell,q}\right)$, $\ell = \overline{1,L}$, где средние значения: $\overline{z}_{\ell,i} = \dfrac{1}{n_\ell} \sum\limits_{z_{\bullet,i} \in \mathbf{z}_\ell} z_{\bullet,i}$ $i = \overline{1,q}$.

Кроме минимизации расстояния Махаланобиса (8) возможны еще несколько способов разделения на классы, например, с помощью нормальных линейных функций Фишера, так как с учетом центральной предельной теоремы в новом пространстве признаков их многомерное распределение $P(\mathbf{z}_i(x))$, $i = \overline{1,q}$ можно считать нормальным. С их помощью определяется наиболее вероятный класс, к которому может быть отнесен каждый объект. Имеется столько же функций классификации, сколько классов. Каждая функция классификации позволяет для каждого наблюдения (пациента) и для каждого класса ℓ вычислить результат показателя классификации Q_ℓ по формуле:

$$Q_\ell = w_{\ell,0} + w_{\ell,1} z_1 + w_{\ell,2} z_2 + ... + w_{\ell,q} z_q, \quad (9)$$

где $\ell = \overline{1,L}$ обозначает номер класса; 1, 2, ..., q – индексы переменных; $w_{\ell,0}$ константа ℓ-го класса; $w_{\ell i}$ – параметр при значении i-й переменной в функции Фишера, используемой для вычисления показателя классификации; z_i – значение i-й КДФ, соответствующей преобразованию исходных переменных, каждого наблюдения (пациента). Принадлежность к определенному классу определяется по максимуму значения (9) как: $\ell^* = \arg \max\limits_{\ell=\overline{1,L}} Q_\ell$, которое соответствует максимуму вероятности принадлежности к классу.

Результаты кластеризации МГУА и КДА

Для проведения корректного сравнения рассмотрим результаты классификации с использованием линейных по входным переменным моделей. Результаты получены в одинаковых условиях использования исходной выборки W. Из 80 точек исходной выборки в экзамен C выделено 10 точек ($W = A \cup B \cup C$, $A \cap B = \varnothing$, $(A \cup B) \cap C = \varnothing$). В сдвоенной табл. 3 представлены результаты линейных классификаторов МГУА и КДА. На диагоналях указаны цифры (соответственно, результат МГУА полужирным шрифтом, а в скобках – результат КДА) в абсолютных величинах и в процентах, отражающие правильную классификацию пациентов с различными диагнозами. Наилучшим образом классифицируются: алгоритмом МГУА – пациенты с диагнозом ДТ (74,2%), КДА-классификатором – с диагнозом БВ (70,8%); худший результат МГУА – пациенты с диагнозом КПСГ (12,5%), а КДА – пациенты с диагнозом КП (23,5%). Из табл. 3 видно, что у обоих методов результаты классификации всей выборки пациентов на четыре класса в среднем одинаковы и неудовлетворительны, так как только

в 51,3 процентах случаев (среднее диагональных значений) исходные наблюдения классифицированы правильно (41 из 80). Поэтому для обоих подходов МГУА и КДА были построены четыре бинарных классификатора (по числу диагнозов), отделяющих пациентов каждого из диагнозов от всех остальных. В пакете, реализующем КДА, использован метод шаговой регрессии включения–исключения для оптимизации набора признаков, а значения порогов (значений критерия Фишера) на включение–исключение оптимизированы с учетом минимизации ошибки классификации на экзаменационной выборке *C*, т.е. с применением принципа МГУА.

Таблица 3. Результаты классификации МГУА (КДА) всей выборки наблюдений на четыре класса

		Количество пациентов, классифицированных по принадлежности к диагнозу (абсолютное количество)				Всего
	Диагнозы	БВ	ДТ	КП	КПСГ	
Наблюдения	БВ	**12** (17)	**11** (2)	**1** (5)	**0** (0)	24
	ДТ	**5** (13)	**23** (11)	**2** (6)	**1** (1)	31
	КП	**2** (3)	**9** (2)	**5** (8)	**1** (4)	17
	КПСГ	**2** (2)	**3** (1)	**2** (0)	**1** (5)	8
		Доля пациентов, классифицированных по принадлежности к диагнозу (%)				
Наблюдения	БВ	**50,0** (70,8)	**45,8** (8,3)	**4,2** (20,8)	**0,0** (0,0)	100,0
	ДТ	**16,1** (41,9)	**74,2** (35,5)	**6,3** (19,4)	**3,2** (3,2)	100,0
	КП	**11,8** (17,6)	**52,9** (11,8)	**29,4** (47,1)	**5,9** (23,5)	100,0
	КПСГ	**25,0** (25,0)	**3,75** (12,5)	**25,0** (0,0)	**12,5** (62,5)	100,0

Априорные вероятности классов в преобразованном пространстве считались равными. Классификация осуществлялась по расстоянию Махаланобиса для двух вариантов ковариационных матриц:

- с равенством внутри классов (если все точки имеют одну общую ковариационную матрицу);
- каждый класс имеет собственную ковариационную матрицу.

Лучший результат КДА с разными вариантами ковариационных матриц также был выбран с учетом минимизации ошибок классификации на экзаменационной выборке. Результаты бинарных классификаторов МГУА и КДА для контрольной выборки представлены в табл. 4. Контрольная выборка составлена из случайно выбранных пациентов и группы, нераспознанных по МГУА. При этом были посчитаны ошибки первого и второго рода для различных бинарных классификаторов (не показаны). Среди них найдены нераспознаваемые и распознаваемые ошибки, последние есть причиной «конфликта» диагнозов и могут быть устранены путем лабораторного тестирования.

Таблица 4. Результаты работы различных бинарных классификаторов на контрольной выборке

№ пациенток	70	79	75	74	22	42	30	10	1	52	35
КДА-классификаторы	?	?	+	+	+	–	+	–	–	+	–
МГУА-классификаторы	–	–	–	–	+	?	+	?	+	+	+

Обозначения: «?» – конфликт диагнозов (распознаваемые ошибки), «–» – неправильное решение (нераспознаваемые ошибки), «+» – правильное решение.

Поскольку точность линейных классификаторов как МГУА, так и КДА достаточно низкая, то необходимо применять нелинейные модели, реализованные в алгоритмах МГУА, имеющие более высокие показатели точности [5]. Несмотря на это идеи, заложенные в КДА, имеют хорошие перспективы использования. Для устранения нераспознаваемых ошибок необходимо применять ансамблевые классификаторы и коллектив решающих правил.

Заключение. По полученным результатам можно сделать следующий вывод: поскольку точность распознавания по рассмотренным методам анализа данных совпадает в среднем, но не совпадает в частности – по отдельным классификаторам и персоналиям, то эти методы можно использовать с другими, например, вероятностного подхода для повышения точности системы классификаторов в решающих правилах.

1. *Ивахненко А.Г.* Метод группового учета аргументов – конкурент метода стохастической аппроксимации // Автоматика. – 1968. – № 3. – С. 58–72.
2. *Ивахненко А.Г.*, *Степашко В.С.* Помехоустойчивость моделирования. – Киев: Наук. думка, 1985. – 214 с.
3. *Павлов А.В.*, *Павлов В.А.*, *Томилин В.В.* Синтез классификаторов дифференциальной диагностики заболеваний легких форм гемостазиопатий методом группового учета аргументов // Восточно-Европейский журнал передовых технологий. – 2011. – № 2/2(50). – С. 42–48.
4. *Кондрашова Н.В.*, *Томилин В.В.* Решение задачи диагностики заболеваний легкой формой коагулопатии и тромбоцитопатии на основе методов экспертных

оценок // Системные технологии: Межвуз. сб. научн. работ. – 2010. – **6**. – С. 104–114.
5. *Кондрашова Н.В.* МГУА и вероятностные методы при построении классификаторов медицинской дифференциальной диагностики // Індуктивне моделювання складних систем: Зб. наук. праць. – К.: МННЦІТ та С НАНУ, 2012. – **4**. – С. 78–89.
6. *Дуда Р.О.*, *Харт П.* Распознавание образов и анализ сцен. – М.: Мир, 1976. – 511 с.
7. *Павлов А.В.* Технологія побудови регресійних моделей на основі ітераційного алгоритму з рекурентними обчисленнями: Автореф. дис. … канд. тех. наук. – К., 2012. – 20 с.
8. *Сарычев А.П.* Идентификация состояний структурно-неопределенных систем. – Днепропетровск: Институт технической механики НАН Украины и НКА Украины, 2008. – 268 с.
9. *Павлов А.В.* Обобщенный релаксационный итерационный алгоритм МГУА // Індуктивне моделювання складних систем: Зб. наук. праць, – К.: МННЦІТ та С НАНУ, 2011. – **2**. – С. 95–108.
10. *Томілін В.В.* Етіологія, прогнозування, профілактика та лікування геморагічних ускладнень при легких формах коагулопатій і тромбоцитопатій: Автореф. дис. … д-ра. мед. наук. – К.: Ін-т гематології та трансфузіології АМН України, 2011. – 39 с.
11. *Баркаган З.С.* Геморрагические заболевания и синдромы. – М.: Медицина, 1988. – 528 с.
12. *Томілін В.В.* Активне виявлення легких форм коагулопатій та тромбоцитопатій // Гематологія і переливання крові: Міжвід. зб. – К.: Атіка-Н, 2010. – **35**. – С. 113–117.
13. *Степашко В.С.* Структурная идентификация прогнозирующих моделей в условиях планируемого эксперимента // Автоматика. – 1992. – № 1. – С. 26–35.

Тел. для справок: +38 044 526-3028, 412-0597, 236-5639, +38 095 169-2910, +38 063 686-3474, +38 050 559-7954 (Киев)
E-mail: *helab@i.com.ua*