УДК 681.513.8

МГУА И ВЕРОЯТНОСТНЫЕ МЕТОДЫ ПРИ ПОСТРОЕНИИ КЛАССИФИКАТОРОВ ДЛЯ МЕДИЦИНСКОЙ ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКИ

Н.В. Кондрашова

Международный научно-учебный центр информационных технологий и систем НАНУ и МОНМолодежи и спорта Украины,

NKondrashova@ukr.net

Інтеґрація методів дозволяє отримати хороші результати в медичній діагностиці. Зокрема, при створенні класифікаторів для диференціальної діагностики захворювань, пов'язаних з незсілістю крові, важко розпізнати діагнози за клінічними ознаками. У цій статті подано результати роботи класифікаторів, побудованих на основі МГУА та імовірнісних методів. Ідеї Байєса, Бордá, Кондорсé поряд з ідеєю самоорганізації моделей були використані при побудові класифікаторів за допомогою імовірнісних методів і МГУА з метою їх застосування для формування правила прийняття рішень. Запропоновано блок-схема алгоритму диференціальної діагностики на основі вищезазначених класифікаторів. українською мовою,

Ключові слова: Легкі випадки патології гемостазу, імовірнісний підхід, оцінки Бордá, метод парних порівнянь Кондорсé, формула Байєса, класифікація, диференційна діагностика, алгоритм МГУА, дерево рішень

Abstract. The integration of methods allows obtaining good results in medical diagnosis. Particularly, the creation of classifiers for differential diagnosis of diseases associated with blood incoagulability is difficult to detect diagnosis by clinical symptoms. This article presents the results of the classifiers that are based on GMDH and probabilistic methods. Ideas of Bayes, Bordá, Condorsé along with the idea of self-organizing models were used to create classifiers with help of probabilistic methods and GMDH for application to form a decision rule. A flowchart of the differential diagnosis based on the above classifications is proposed. in English.

Keywords: Mild cases hemostasis, pathology, probabilistic approach, Borda count, Condorcet's paired comparison method, Bayes' formula, classification, differential diagnostics, GMDH algorithm, decision trees.

Аннотация. Объединение методов позволяет получить хорошие результаты в медицинской диагностике. В частности, при создании классификаторов для дифференциальной диагностики заболеваний, связанных с несвертываемостью крови, трудно распознать диагнозы по клиническим признакам. В этой статье даны результаты работы классификаторов, построенных на основе МГУА и вероятностных методов. Идеи Байеса, Бордá, Кондорсé наряду с идеей самоорганизации моделей были использованы при построении классификаторов с помощью вероятностных методов и МГУА с целью их применения для формирования правила принятия решений. Предложена блок-схема алгоритма дифференциальной диагностики на основе вышеупомянутых классификаторов.

Ключевые слова: Легкие случаи патологии гемостаза, вероятностный подход, оценки Бордá, метод парных сравнений Кондорсé, формула Байеса, классификация, дифференциальная диагностика, алгоритм МГУА, дерево решений.

Вступление

Проблема принятия решений возникает в медицинских задачах, особенно в связи с постановкой диагноза, в частности, когда количество наблюдаемых признаков достаточно велико. Поставить диагноз по клиническим признакам

является трудно разрешимой задачей еще и потому, что не всегда их наличие связано с болезнью [1]. Особенностью и одновременно трудностью диагностики по значениям выделенных врачами-специалистами симптомов является трудность различения диагнозов из-за частичного совпадения клинических признаков. Т.е. один и тот же признак, а также одинаковое их сочетание, может наблюдаться при различных заболеваниях. Такая ситуация имеет место при заболеваниях крови, обусловленных ее плохой свертываемостью. Эти болезни известны под общим названием “легкие случаи патологии гемостаза” или иначе – “легкие формы коагулопатии и тромбоцитопатии (КиТ)” [2].

Как правило, у исследователей имеется некоторая статистика наблюдений за больными. Вероятностный подход используется потому, что исходные данные представляют собой сведения об априорных частотах проявления различных симптомов у больных с патологией гемостаза. Эти данные были сгруппированы в соответствии с принадлежностью к группам с учетом пола, возраста и диагнозов, поставленных на основании лабораторных исследований и результатов дорогостоящих тестов. Лишь для одной возрастной группы пациенток имелись данные о конкретных значениях клинических признаков, сопутствующих диагнозам.

В [3] на основании таблиц априорных частот решение о предрасположенности больных к тому или иному заболеванию крови принималось на основании критерия, использующего преобразованные значения этих частот. В [4] классификаторы по методу группового учета аргументов (МГУА) строились по многомерным выборкам данных в пространстве значений признаков. Результаты классификации на основе вероятностных методов можно сравнивать с результатами классификации по МГУА только на той группе пациенток, для которой были известны данные в полном объеме.

Исходя из характера данных, в [5] было рассмотрено два основных подхода получения решения: вероятностный и индуктивный. При этом диагноз для каждой пациентки определялся:

1) полученным ранее алгоритмом преобразования значений относительных частот (вероятностей) признаков в соответствии с наличием тех или иных признаков у конкретной больной;
2) с помощью построенных моделей связи признаков с диагнозами (признаки конкретных пациенток задаются своими значениями).

В первом случае модель будем называть алгоритмической, а классификаторы – вероятностными, во втором – модель функциональная, а классификаторы носят название в соответствии с методом, с помощью которого они построены. В данной работе в качестве функциональных использовались МГУА-классификаторы.

При создании вероятностных классификаторов были применены известные методы: оценок Борда (МОБ) [6], парных сравнений (МПС) Кондорсе [7] и метод, основанный на идее Байеса (МБ) [8]. По результатам классификации этими классификаторами, а также МГУА-классификаторами построена преференциальная система принятия решений (блок-схема будет представлена ниже).

Наиболее простой в употреблении – первый из упомянутых – метод на основе рейтинговых оценок Борда (Borda count method) был выбран как наиболее часто применяющийся метод именно для подготовки информации лицу, принимающему решение. Этот метод основан на прямом ранжировании. Далее для обоснования принятия решения используется статистический критерий проверки согласованности. Относительные частоты (далее просто частоты) легко интерпретируются, как ранги. Этот метод допускает наличие пересекающихся классов в пространстве признаков, что учитывается вычислением *показателя взаимосвязанности рангов.* По сравнению с методом Байеса в методе Борда используются «загрубленные» оценки, т.к. интервальные значения частот преобразуются в целочисленные значения рангов. В ряде случаев это позволяет для пациентов, получив ответ: «не знаю» избежать неправильной постановки диагноза. На основе ранговых коэффициентов можно ответить на вопрос: следует ли принять гипотезу H_0 об отсутствии различий и «согласованы» ли эксперты в своем решении относительно диагноза? При этом не всегда удается ответить на вопрос, в отношении какого именно диагноза имеется согласие. Это будет продемонстрировано на дальнейших примерах.

На самом деле в МОБ нулевой гипотезе H_0 соответствует гипотеза об отсутствии отличий от равномерного распределения при голосовании за тот или иной диагноз. Равномерный закон, соответствует случаю, когда эксперты не могут совместно отдать предпочтение какому либо диагнозу (классу), т.е. невозможно различить классы по результату голосования. При проверке данной гипотезы используется критерий хи-квадрат Пирсона, поскольку исходные данные для него могут быть получены в любой шкале.

Недостаток вышеуказанного метода анализа ранжировок иногда устраняет метод парных сравнений. В этом методе вместо гипотезы равномерного распределения рассматривается гипотеза однородности, т.е. вместо совпадения всех распределений с одним фиксированным (равномерным) проверяется лишь совпадение распределений мнений экспертов между собой, что естественно трактовать как согласованность их мнений в отношении какого либо диагноза. Таким образом, удается избавиться от неестественного предположения равномерности. На основе этого метода относительно просто получить подсказку «советчика» об изменении диагноза больного при изменении его состоянии, например, при добавлении нового симптома.

1. Формулировка задачи

Пусть в пространстве клинических признаков x_i, $i=1,...,m$ (m=18) заданы (k=4) классы (диагнозы): D_1– болезнь Виллебранда (БВ), D_2 – коагулопатия (КП), D_3 – дезагрегационная тромбоцитопатия (ДТ), D_4 – комбинированная патология системы гемостаза (КПСГ). Каждый из четырех диагнозов был установлен пациентам в клинической лаборатории при использовании дорогостоящих реактивов. Клинические признаки x_i принимают, как правило, целочисленные значения «да» (+1) или «нет» (-1), но для некоторых больных вводится тре-

тье значение – «не было условий для проявления данного признака» (0). В скобках, например (+1) или (-1), даются значения признаков, обозначенные в таблице 1 соответственно «+» или «-».

Экспертами данной предметной области выделены следующие девятнадцать геморрагических признаков: 1 – ювенальное маточное кровотечение (ЮМК); 2 – дисфункциональное маточное кровотечение (ДМК); 3 – носовое кровотечение (НК); 4 – кровоточивость десен (КД); 5 – кровотечение после экстракции зубов (КПЭЗ); 6 – интра и послеоперационное кровотечение (ПОК); 7 – послетравматическая гематома (ПТГ); 8 – кровотечение из поверхностных ран (КПР); 9 – продолжительное не заживление ран (ПНЗ); 10 – послетравматический гемартроз (ПГ); 11 – послеинъекционная гематома (ПИГ); 12 – кровотечение из-за травмы уздечка языка; 13 – желудочно-кишечное кровотечение (ЖКК); 14 – паховая гематома; 15 – кровотечение при прорезывании зубов; 16 – кефалогематома при рождении (КР); 17 – почечное кровотечение (ПК); 18 – послеродовое кровотечение (ПРК); 19 – геморрагический инсульт. Насколько полон набор из девятнадцати признаков для однозначного определения указанных диагнозов в данной работе не обсуждается.

Формулировка задачи заключается в следующем.

Имеется выборка наблюдений. Фрагмент выборки представлен в таблице 1. Наблюдаемые признаки x_i, $i=\overline{1,m}$ заболеваний (диагнозов) $D_j \in D$, $j=\overline{1,k}$ принимают целочисленные значения из конечного множества трех значений

$$x_i \in \{-1,0,1\}.$$

В соответствии с данными всей выборки получена таблица 2. Она содержит числа $p_{i,j}$, $j=\overline{1,k}$, $i=\overline{1,m}$, которые означают вероятности (относительные частоты) наблюдаемости признака x_i при наличии диагноза D_j (вычисляется как частное двух чисел $n_{i,j}/n_i$, где $n_{i,j}$ - число пациентов, у которых наблюдался признак x_i и был диагностирован D_j; n_i - объем общей выборки пациентов задачи дифференциальной диагностики (D={БВ, КП, ДТ, КПСГ}), имеющих признак x_i).

Пусть w: $X \to D$ означает функцию, которая по наблюдению $X_s \subseteq X$, $X_s = (x_1,...,x_s)$ принимает значение $w(X_s) \in \Re$. Функция w называется решающей функцией, т.к. по ее максимальному значению делается предварительное заключение о наличии того или иного диагноза:

$$D^* = \arg\max_{i=\overline{1,k}} w(X_s(D_i)), \quad D^* \in D. \qquad (1)$$

Задача состоит в том, чтобы при заданных множествах X, D и функции $p_{XD}: X \times D \to \Re$ найти решающее правило f: $X \to D$, которое максимизирует точность классификации диагнозов на всей выборке, а также на независимой (экзаменационной) выборке.

Решающее правило будем строить в виде дерева решающих правил [9]. В качестве «листьев» этого дерева будут вероятностные классификаторы и классификаторы, полученные по методу группового учета аргументов.

Если ввести пороговое значение приемлемой вероятности, то тогда появится вариант: отказ от классификации – “не классифицировано” (значение вероятности ниже порогового). Точность классификации будем вычислять, как отношение числа n_r правильно диагностированных пациентов к их общему числу n. В данной работе n – объем контрольной выборки.

Пусть f^* функция решающих правил (решающая функция), которая максимизирует точность классификации

$$f^* = \max_{f \in F,\, p \in [0,1]} n_r\big(f(p)\big)/n \qquad (2)$$

где F - множество функций, реализующих решающие правила рассматриваемого множества классификаторов (вероятностных и МГУА). Решением задачи является построение минимального дерева решающих правил, удовлетворяющего (2). Количество признаков m зависит от исследуемой группы пациентов и изменяется от 5 до 13. В женской группе в возрасте от 19 до 49 лет не все из перечисленных признаков имеют место, а лишь те, которые представлены в таблицах 1 и 2 (m =13).

Таблица 1.

Наличие геморрагических проявлений у женщин в возрасте от 19 до 49 лет с диагнозом (D)

№ пац	В-т лет	D	ЮМК 1	ДМК 2	НК 3	КД 4	КПЭЗ 5	ПОК 6	ПТГ 7	КПР 8	ПНЗ 9	ПГ 10	ПИГ 11	ЖКК 13	ПК 16	ПРК 18
70	38	ДТ	+1	+1	+1	+1	+1	+1	-1	-1	-1	-1	-1	-1	-1	0
79	43	ДТ	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1	-1	+1	-1	+1
73	21	КП	+1	+1	+1	+1	+1	0	+1	-1	-1	-1	-1	-1	-1	0
74	49	БВ	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1	-1	-1	-1	0
80	19	БВ	+1	+1	+1	-1	-1	-1	+1	+1	+1	-1	-1	-1	-1	-1
63	20	КП	+1	+1	+1	+1	-1	0	+1	+1	+1	-1	-1	-1	-1	-1
42	49	ДТ	-1	-1	+1	+1	+1	0	+1	-1	-1	-1	-1	-1	-1	-1
113	31	БВ	+1	+1	+1	+1	+1	0	+1	-1	-1	-1	-1	-1	-1	-1
68	49	КП	+1	+1	+1	-1	+1	-1	+1	+1	+1	-1	-1	-1	-1	0
57	21	ДТ	+1	+1	+1	+1	+1	0	+1	-1	-1	-1	-1	-1	-1	0
83	28	ДТ	+1	+1	+1	+1	+1	0	+1	+1	+1	-1	-1	-1	-1	+1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Обозначение: (В-т) – возраст

Таблица 2.

Относительная частота возникновения геморрагических проявлений у женщин в возрасте от 19 до 49 лет с диагнозами легкой формы КиТ

№ признаков	1	2	3	4	5	6	7	8	9	10	11	13	16	18
БВ	1	1	1	0,5	0,82	0,8	0,5	0,63	0,5	0,04	0,04	0	0	0,7
ДТ	0,91	0,91	0,84	0,56	0,67	0,68	0,59	0,56	0,37	0	0	0,03	0	0,5
КП	0,88	0,88	0,71	0,65	0,86	0,75	0,77	0,53	0,29	0,06	0,06	0	0	0,3
КПСГ	0,75	0,75	0,5	0,38	0,8	0,83	0,63	0,75	0,5	0,13	0,13	0	0,1	0,7

Примечание. Автор исходных данных для рассматриваемой задачи– д.м.н. Томилин В.В., ГУ "Институт гематологии и трансфузиологии" АМН Украины.

Используя данные таблицы 2, необходимо построить алгоритм предварительной постановки диагноза, и оценить точность работы системы с помощью данных таблицы 1. Классификаторы, на основе которых будет работать система принятия решений, должны будут для каждого пациента определять один правильный из четырех возможных диагнозов.

Анализ данных таблицы 1 выявил наличие пересекающихся наборов признаков с одинаковыми значениями, характеризующих различные диагнозы, а также «двойников» (№73, №57), т.е. людей с совпадающими значениями одинаковых наборов признаков и одинаковым возрастом, различающихся только названиями диагнозов. Понятно, что «двойников» никакой классификатор различить не может, но один из этих двух диагнозов классификатор вполне может идентифицировать, и, естественно, – тот, априорная вероятность у которого будет выше (в данном случае это диагноз пациента под номером 73). Множества больных с совпадающими наборами признаков должны быть обособлены и дополнительно исследованы (в данной выборке имеется пара таких больных).

2. Классификация по методу на основе оценок Борда

Отличительной особенностью классификатора является то, что *признаки* используются в качестве *экспертов*, которые будут «голосовать» за тот или иной диагноз в соответствии с *мерой*, которую они получили в результате обследования пациентов в специализированной клинике. Такой *мерой* является представленная в таблице 2 *относительная частота p* присутствия определенного признака при установлении каждого из диагнозов. Одной этой меры недостаточно для однозначного определения диагноза в силу пересечения множеств признаков, поэтому используется коллективное согласованное подтверждение.

Критика оснований применения статистических методов в экспертных системах, в частности, догмы согласованности в данном случае не может быть актуальной в части, касающейся реально имеющегося разделения экспертов на группы. Поскольку геморрагические признаки психологически не воздействуют друг на друга и потому не склонны к конформизму.

Относительная значимость тех или иных диагнозов при вероятностном подходе устанавливается с помощью коэффициента конкордации Кендалла-Смита, статистического критерия Пирсона χ^2 и результата парного сравнения. Единственность диагноза обеспечивается максимумом веса предпочтений (1), где вес w_ℓ вычисляется на основе вычисления рангов по формулам, приведенным при описании этих методов в [3]. Обозначим функцию, в соответствии с которой принимается решение о диагнозе каждым из методов, через f_ℓ. Нижний индекс ℓ порядковый номер метода, для МГУА ℓ=1, МОБ ℓ=2, МПС ℓ=3, МБ ℓ=4. Если решающая функция f_ℓ, полученная каким-либо методом ℓ,

имеет одинаковое значение сразу для нескольких диагнозов, то эту ситуацию назовем «конфликтом диагнозов».

3. Классификация по методу парных сравнений Кондорсе

Среди больных немало таких, для которых критерий χ^2=0, т.е. с помощью МОБ диагноз не выявляется, например, больные 74,75,73,42 в таблице 1.
Методом парных сравнений определим, какой именно диагноз имеет предпочтение. Особенностью применения МПС является переход от вероятностей к целочисленной шкале: 1, 2, 3, 4, и вычисление оценки предпочтений любого из диагнозов каждым экспертом в соответствии со значением частот. Подробно метод описан в [3]. Относительная значимость тех или иных диагнозов определяется весами $w_{j,\ell}$, единственность диагноза определяется по формуле (1).

Если наибольшие веса совпадают $w_{j,\ell} = w_{i,\ell} = ... = w_{v,\ell}, \quad v \neq i \neq j;$ $i, j \in \overline{1,4}$, то единственность диагноза обеспечивается максимумом показателя, который вычисляется, как произведение априорной вероятности и весового коэффициента диагноза:

$$D_\ell^+ = \arg\max_{j=\overline{1,4}}(p_j w_{j,\ell}),\ D_\ell^+ \in \{\text{БВ, ДТ, КП, КПСГ}\}, \qquad (3)$$

где априорная вероятность p_j есть доля имеющих диагноз D_j по отношению к общему количеству больных в данной возрастной группе женщин. Модификации методов МОБ и МПС, учитывающие априорную вероятность диагноза p_j, обозначаются как МОБ$^+$ и МПС$^+$. В случае МОБ-классификатора решающая функция $f_2 = w_2$, МОБ$^+$ $f_2^+ = w_2 p$. Для МПС $f_3 = w_3$, МПС$^+$ $f_3^+ = w_3 p$. Для устранения неопределенности, когда классификаторы МОБ и МПС и их модификации могут не «не дать» однозначного ответа: имеет место отказ от распознавания или оба метода приводят к конфликту диагнозов, либо результат их решения не согласован ($D_\ell^* = \arg\max_{i \in D} f_{\ell,i}$, $D_k^* = \arg\max_{i \in D} f_{k,i}$, $D_\ell^* \neq D_k^*$, k, ℓ – индексы различных методов), рассмотрим классификатор на основе Байесовской формулы.

4. Байесовская классификация

Для построения решающего правила воспользуемся формулой Байеса-Лапласа [10]:

$$p(D_j \mid X_s) = \frac{p(X_s, D_j)}{p(X_s)} = \frac{p(D_j)p(X_s \mid D_j)}{\sum_{j=1}^k p(D_j)p(X_s \mid D_j)} = \frac{p(D_j)p\left((x_1 \cup x_2 ... \cup x_s)\middle| D_j\right)}{\sum_{j=1}^k p(D_j)p(X_s \mid D_j)}, \qquad (4)$$

где $x_i \in X_s$ набор геморрагических признаков конкретного пациента, для которого оценивается вероятность диагноза $s \leq m$. Объем данных недостаточен, поэтому учесть взаимозависимость признаков не представляется возможным (нет возможности подсчитать $p(x_1,...,x_v \middle| x_q,...,x_w)$). Апостериорная вероятность

диагноза рассчитывается по формуле наивного Байесовского классификатора при условии независимости признаков $x_1,...,x_v,x_q,...,x_w$ и того, что известны условные вероятности $p(x_i \mid D_j)$ как:

$$p(D_j \mid X_s) = \frac{p(D_j) \prod\limits_{\forall x_i \in X_s} p(x_i \mid D_j)}{\sum_{j=1}^{k} p(D_j) \prod\limits_{\forall x_i \in X_s} p(x_i \mid D_j)}. \quad (5)$$

Поскольку знаменатель одинаков для всех диагнозов, то для выявления предпочтительности диагноза при их сравнении есть смысл оценивать только числитель формулы (4). Наиболее предпочтительный диагноз определяется в предположении, что известны условные вероятности как

$$D_4^* = \arg\max_{j=\overline{1,4}} p(D_j) p(D_j \mid X_s) = \arg\max_{j=\overline{1,4}} p(D_j) \prod_{i=\overline{1,s}} p_{i,j}.$$

Важным является трактовка чисел $p_{i,j}$ в таблице 2. В зависимости от смысла, вкладываемого в значение $p_{i,j}$, содержащихся в таблице 2, формула (4) при сравнении диагнозов может вычисляться, как с учетом $p(D_j)$, так и без него. Если $p_{i,j}$ - условные вероятности, то $p(D_j)$ следует учитывать, и в формуле (5) вместо $p(x_i \mid D_j)$ подставляются вероятности $p_{i,j}$ j-го диагноза и i-го признака. Если $p_{i,j}$ - совместная вероятность, то в (5) $p_{i,j} = p(x_i, D_j)$ и $p(D_j)$ учитывать не нужно. Тогда диагноз определяется как:

$$D_4^- = \arg\max_{j=\overline{1,4}} p(D_j, X_s) = \arg\max_{j=\overline{1,4}} \prod_{i=\overline{1,s}} p_{i,j}.$$

Модификация классификатора, основанного на формуле Байеса, не учитывающая априорную вероятность диагноза p_j обозначается как МБ$^-$. Для классификатора МБ $f_4 = p(D_j) \prod\limits_{i=\overline{1,s}} p(x_{i,j})$, для классификатора МБ$^-$ $f_4^- = \prod\limits_{i=\overline{1,s}} p(x_{i,j})$.

5. Результаты классификации вероятностными методами и МГУА

Проанализируем работу вероятностных классификаторов и МГУА-классификаторов. Результаты диагностирования группы, объединяющей наугад выбранных пациентов и пациентов с нераспознанными диагнозами, представлены в таблице 3.

По результатам проверки работы классификаторов на контрольной выборке видно, что при классификации по методу Борда в 80% имеется отказ от классификации и только модифицированный метод Борда в этом случае с вероятностью максимального диагноза позволяет независимо от набора признаков всегда давать ответ ДТ, который только в 50% случаев оказывается верным.

Особый интерес представляют результаты работы МОБ, МПС и МБ классификаторов на тех данных, которые относятся к множеству нераспознанных МГУА-классификаторами (это данные множества Ω пациенток под номерами 70, 74, 73(57), 79 в таблицах 1 и 3).

Таблица 3.

Результаты работы различных классификаторов

	№ п/п	1	2	3	4	5	6	7	8	9	10
	№ пациентки	70	79	73(57)	74	63	83	68	113	80	42
Вероятностные классификаторы	МОБ	-	+	??	??	??	??	??	??	??	??
	$МОБ^{+}$	+	+	+(-)	-	-	+	-	-	-	+
	МПС	-	-	-	+	-	-	-	?	+	-
	$МПС^{+}$	+	-	?	+	-	-	-	?	+	-
	$МБ^{-}$	-	-	-(+)	+	-	-	-	-	+	-
	МБ	-	-	+(-)	+	-	-	-	-	+	+
МГУА- классификаторы		-	-	-	-	+	?	+	+	+	+

Обозначения: «?» – конфликт диагнозов, «-» – неправильное решение, «+» – правильное решение, «??» – хи-квадрат равен нулю (согласованный отказ от классификации), «+ (-)» – правильный диагноз для пациента №73 и неправильный диагноз для пациента № 57, а также наоборот «-(+)».

Оказалось, что для пациенток №74 $\in \Omega$ и №73(57) $\in \Omega$ метод, основанный на оценках Борда (МОБ) не может определить диагноз, т.к. критерий хи-квадрат равен нулю (обозначено как «??» в таблице 3). Однако метод парных сравнений для 74-й пациентки позволяет определить правильный диагноз, а $МОБ^{+}$ – правильный диагноз для одного из двойников №73 и еще для двоих (70 и 79) из числа нераспознанных по МГУА пациенток. Для пациенток 70, 73(57) и 79 метод парных сравнений неправильно определяет результат, но метод МПС+ для 74 и 70 «дает» правильный диагноз. При голосовании по большинству среди различных классификаторов для №74 пациентки имеется перспектива избегнуть лабораторной проверки диагноза, если еще один классификатор даст подтверждение правильного диагноза т. к. это решение поддержано также классификатором МБ (кроме того имеется большинство при голосовании среди двух классификаторов и их модификаций). Для пациентки 70 – аналогично желательно иметь результат классификации еще одним классификатором, т.к. по двум модификациям $МОБ^{+}$ и $МПС^{+}$ уже имеется правильный результат.

По данным пациентки №79 имеется одно правильное решение классификатора на основе оценок Борда (МОБ, МОБ+), а значит, эта ошибка при проверке согласованности голосования по всем 4-м классификаторам не определяется и является устранимой только лабораторным тестированием.

Таким образом, классификаторы вероятностного подхода (МОБ, МПС и МБ) позволяют обнаружить все четыре нераспознанные ошибки из класса нераспознанных МГУА-классификатором, причем две устранить – по правилу

большинства, одну как результат лабораторного тестирования, а одну с помощью либо дополнительного классификатора, либо лабораторного тестирования.

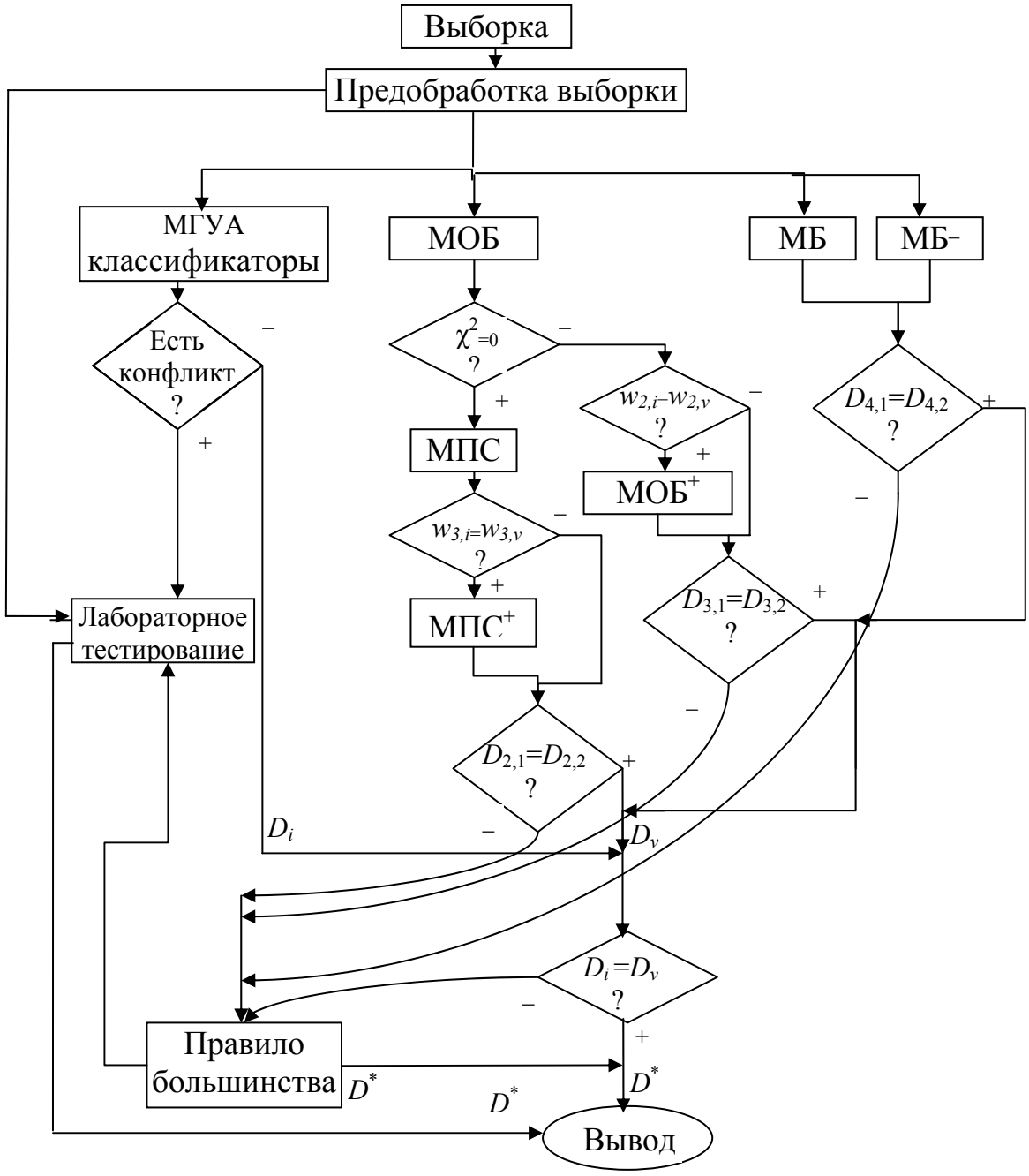

Рис. 1 Блок-схема мажоритарной системы дифференциальной диагностики

На рис. 1 представлена блок-схема мажоритарной системы дифференциальной диагностики. По результатам решений классификаторов МГУА и различных вероятностных методов стоится дерево принятия решений. Исходя из результатов проверки работы классификаторов на контрольной выборке в соответствии с такой схемой принятия решений (по большинству) только лабораторное тестирование пациентов 79, 63, 83, 68, 113 сможет устранить ошибки системы. Схема иллюстрирует предложение по реализации идеи голосования классификаторов (комитет решающих правил) и подлежит дальнейшей корректировке.

Контрольная выборка состояла частично из пациенток с нераспознанными МГУА-классификаторами диагнозами и пациенток выбранных случайным образом из общей выборки. В целом, на множестве из десяти пациенток контрольной выборки классификаторы вероятностного подхода «имели» 3, а МГУА-классификаторы – 4 нераспознаваемые ошибки. При этом не было ни одной пациентки, при диагностике которой все бы «согласованно ошибались». Все вероятностные классификаторы «ошибались» в случаях диагностики пациенток под номерами 63, 68 и 113, для которых МГУА-классификаторы «не ошибались».

Для всех трех вероятностных классификаторов учет априорной вероятности диагнозов $p(D_j)$ увеличивает процент правильной классификации.

Различные классификаторы при постановке диагноза десяти пациенткам контрольной выборки «неустранимо ошибались» в следующих случаях: классификатор по методу парных сравнений (вместе с МПС$^+$) и Байесовский классификатор (вместе с МБ$^-$) – 6 раз; классификатор, основанный на оценках Борда (вместе с МОБ$^+$) - 5 раз; МГУА-классификатор – 4 раза.

Метод, основанный на оценках Борда, является самым «осторожным» в оценках, «дает» больше, чем другие методы «отказов от распознавания» (в 80% случаев). МОБ и МОБ$^+$ «согласованно определили» один раз правильный диагноз.

Метод парных сравнений и его модификация в одном случае «не могли отдать» предпочтение двум альтернативным диагнозам – (?), и в двух случаях МПС и МПС$^+$ «согласованно определили» правильный диагноз.

Байесовский классификатор (МБ и МБ$^-$) в трех случаях «согласованно определил» правильный диагноз. Довольно низкие показатели байесовского классификатора объясняются тем, что вычисления производятся в предположении о взаимной независимости признаков, чего в реальности не бывает, а при построении МГУА-классификатора зависимость признаков не является столь критичной.

Выводы

1. Ни один классификатор без применения других не следует использовать для безошибочного диагностирования легких форм патологий гемостаза.
2. На контрольной выборке самый высокий процент правильной классификации диагнозов у МГУА-классификаторов (не менее 50%).
3. Все классификаторы на выборке 80-ти пациенток ни разу «не ошиблись согласованно».
4. С учетом несогласованности результатов МГУА-классификаторов с результатами мажоритарного голосования классификаторов вероятностного подхода можно полностью устранить неопределенность путем лабораторного тестирования для найденной группы пациентов и получить 100% точность.

Дальнейшее развитие системы дифференциальной диагностики на основании классификаторов диагнозов БВ, КП, ДТ и КПСГ может осуществляться по двум направлениям:

1.Включение в перечень используемых признаков дополнительных характеристик диагностируемых патологий и на основе нового состава признаков разработка более надежных классификаторов.

2. Совершенствование диагностической системы принятия решений.

Выражение признательности

Автор благодарен Томилину В. В. за предоставленные данные результатов лабораторного диагностического тестирования.

Литература

1. Баркаган З.С. Геморрагические заболевания и синдромы - переработанное и дополненное. – 2-е изд. – М: Медицина. – 1988. – 528 с.
2. Томілін В. В. Етіологія, прогнозування, профілактика та лікування геморагічних ускладнень при легких формах коагулопатій і тромбоцитопатій [Текст] : автореф. дис. ... доктора мед. наук : 14.01.31; АМН України, ДУ "Ін-т гематології та трансфузіології". – К., 2011. – 39 с.
3. Кондрашова Н.В., Томилин В.В. Решение задачи диагностики заболеваний легкой формой коагулопатии и тромбоцитопатии на основе методов экспертных оценок // Системные технологии. Межвузовский сборник научных работ. – Дн., – 2010.– Вып. 6. – С.104-114.
4. Павлов А.В. Павлов В.А. Томилин В.В. Синтез классификаторов дифференциальной диагностики заболеваний легких форм гемостазиопатий методом группового учета аргументов // Восточно-Европейський журнал передовых технологий. – Харьков, 2011. – № 2/2(50). – С.42-48.
5. Kondrashova N.V. About algorithm of decision-making in the medical differential diagnosis // Proc. of 5th International Workshop on Inductive Modeling IWIM 2012, Kyiv-Zhukyn, Ukraine. – Kyiv: IRTC ITS NASU, – 2012. – p. 7-14.
6. Jean-Charles de Borda [Електронний ресурс] — 2011. — Режим доступа: http://en.wikipedia.org/wiki/Jean-Charles_de_Borda.
7. Condorcet M. J. A. N. Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix. — 1785.
8. Bayes Th. An essay towards solving a Problem in the Doctrine of Chances. – Philosophical Transactions of the Royal Society of London 53. –1763.
9. Breiman L., Fiedman J.H., Olshen R.A. & Stone C.J. Classification and regression trees. Monterey. CA: Wadswort & Books/Cole Advanced Books & Software. 1984. ISBN 978-0-412-04841-8.
10. Гнеденко Б.В. Очерки истории вероятностей 69с. (Гнеденко Б. В. Курс теории вероятностей: Учебник – переработанное и дополненное – 6-е изд. – М: Наука. Гл. ред. физ-мат. лит. – 1988. – 448 с.)