УДК 006.78

СИНТЕЗ КЛАССИФИКАТОРОВ ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ ЛЕГКИХ ФОРМ ГЕМОСТАЗИОПАТИЙ ПО МГУА

А.В. Павлов
Аспирант
Отдел информационных технологий и индуктивного моделирования
Международный научно-исследовательский учебный центр информационных технологий и систем НАН и МОН Украины
пр. Глушкова, 40, г. Киев, Украина, 03680
Контактный тел.: (044) 526-15-70; (044) 412-05-97

В.А. Павлов
Кандидат технических наук, доцент
Кафедра компьютерного эколого-экономического мониторинга
Открытый международный университет развития человека «Украина»
ул. Хорева, 1-Г, г. Киев, Украина, 04071
Контактный тел.: (044) 424-62-74, 050-559-79-54
E-mail: vpavlo@bk.ru

В.В. Томилин
Кандидат медицинских наук, старший научный сотрудник
Отделение хирургической гематологии и гемостазиологии
ДУ Институт гематологии и трансфузиологии АМН Украины
ул. Максима Берлинского, 12, г. Киев, Украина, 04060
Контактный тел.: (044) 440-27-44; 050-330-96-17

У роботі отримано класифікатори диференціальної діагностики захворювань легкою формою коагулопатій і тромбоцитопатій методом групового урахування аргументів. Точність класифікаторів отримана не менше 90%. Запропоновано шляхи розвитку діагностичної системи

Ключові слова: класифікація, діагноз, МГУА

В работе получены классификаторы дифференциальной диагностики заболеваний легкой формой коагулопатий и тромбоцитопатий методом группового учета аргументов. Точность распознавания получена не менее 90%. Предложены пути развития диагностической системы

Ключевые слова: классификация, диагноз, МГУА

In the paper classifiers for light form of coagulopathy and thrombocytopathy differential diagnostics has been obtained using GMDH. Recognition accuracies were not less then 90% at whole data set. New ways of diagnostic system development were suggested

Keywords: classification, diagnosis, GMDH

1. Введение

Интенсивное развитие методов индуктивного моделирования позволяет получать хорошие результаты в различных прикладных областях. Связано это с тем, что методы данного направления, решая обе задачи моделирования (структурная и параметрическая идентификация) позволяют получать модели опти-

мальной структуры в смысле минимума внешнего критерия. Ниже рассмотрена задача построения классификаторов дифференциальной диагностики для четырех заболеваний, достаточно трудно различимых по клиническим признакам в врачебной практике.

2. Постановка задачи

Пусть в пространстве клинических признаков x_i, i=1,...m (рис. 1) заданы 4 класса (диагнозы): $Д_1$ – болезнь Виллебранда (БВ) , $Д_2$ – коагулопатия (КП), $Д_3$ – дезагрегационная тромбоцитопатия (ДТ), $Д_4$ – комбинированная патология системы гемостаза (КПСГ).

Клинические признаки (m=12) принимают, как правило, бинарные значения «да» или «нет», однако для некоторых больных необходимо вводить третье значение – «не было условий для проявления данного признака».

Например, признак – «кровотечение после операций» не наблюдался, однако и самих операций у некоторых пациентов не было.

Для построения классификаторов выбрана группа пациентов – женщины, возраст – в пределах 19-49 лет. Для наглядности, на рис. 1 приведена постановка задачи распознавания диагнозов в плоскости двух признаков x_1 и x_2. При этом признаки изображены для общего случая задачи распознавания, как имеющие непрерывную область определения, а не как бинарные, в нашем случае. Классы в плоскости признаков отображены на рисунке различными значками.

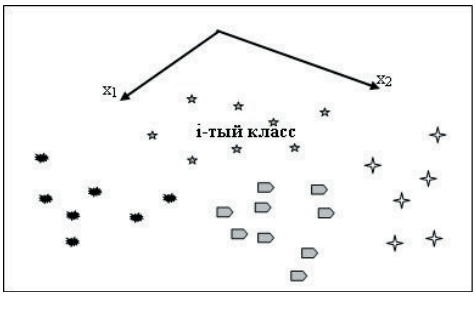

Рис. 1. Проекции точек различных классов (диагнозов) в плоскость двух произвольно выбранных признаков

Будем искать уравнения границ заданных классов $y_i(x)=c_i$, i=1,...,4, где x - вектор признаков, как задачу «один против всех», отделяя каждый i-тый класс от прочих классов своим i-тым классификатором $y_i(x)-c_i=0$ (рис. 2).

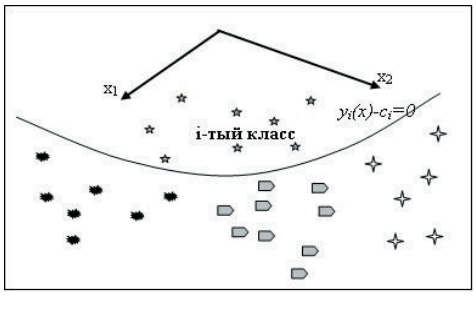

Рис. 2. Проекции точек и разделяющей поверхности

Если удастся построить некоторые функции $y_i(x)-c_i$ такие, что они будут разделять классы так, как на рис. 2, то тогда очевидно следующее: на точках (значениях клинических признаков пациента) i-того диагноза функция $y_i(x)-c_i$ будет принимать значения одного определенного знака (допустим, в верхней части рисунка – положительные), а на точках других диагнозов – значения противоположного знака (отрицательные). На точках, расположенных на самой линии раздела $y_i(x)-c_i=0$. Таким образом функция $y_i(x)-c_i$, принимая на точках i-того диагноза только положительные значение, а на прочих – неположительные, будет являться индикатором (классификатором) i-того диагноза.

Сформулируем наши требования к классификаторам $y_i(x)-c_i$, i=1,...,4. Требуется получить такие классификаторы $y_i(x)-c_i$, которые, будучи синтезированы на имеющейся у нас выборке пациентов были бы максимально чувствительны и специфичны не только к данной выборке но и на других выборках пациентов данной группы.

3. Обоснование метода решения задачи

Один из распространенных способов решения задач такого типа состоит в том, чтобы задать для каждой функции $y_i(x)$ вид идеального классификатора $P_i(x)$ и затем применить подходящие методы моделирования для того, чтобы построить реальный классификатор $y_i(x)$, максимально похожий (насколько это позволит аппарат моделирования) на идеальный.

В качестве идеального классификатора обычно задают пороговую функцию $P_i(x)=\begin{cases}A, \forall E \in Д_i \\ B, \forall E \notin Д_i\end{cases}$, которая разрезает (т+1)-мерное пространство переменных (на рис. 3 т=2, плоскость двух признаков x_1, x_2) по плоскости $P_i(x)= c_i$.

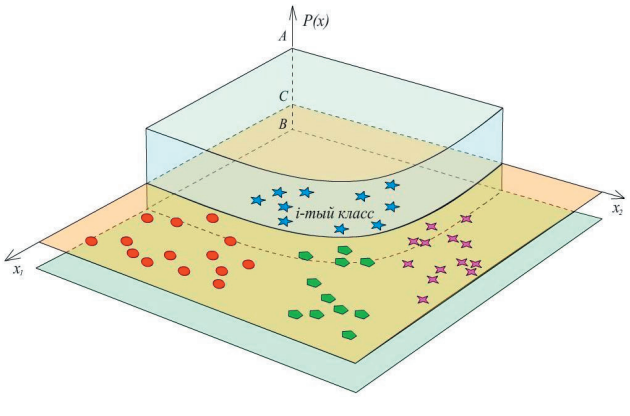

Рис. 3. Трехмерное изображение идеального классификатора

Сформулированное выше требование к классификаторам делает необходимым учитывать следующее.

Если просто «заставить» некоторую функцию $y_i(x,a)$ с помощью выбора подходящего значения вектора параметров a, принимать положительные значения на точках i-того диагноза на всей имеющейся у нас выборке пациентов (обозначим ее X^W), то нет

никаких гарантий, что на «свежих», новых точках (пациентах), с другим набором значений клинических признаков, классификатор даст правильный результат.

Для построения структур классификаторов, наилучшим образом «угадывающих» диагноз и для новых пациентов разработан специальный метод – метод группового учета аргументов [1,2].

Суть метода в том, что вся выборка пациентов X^W делится на две непересекающиеся части: $X^A \subset X^W$, – обучающая выборка и $X^B \subset X^W$, –проверочная выборка, $X^A \cap X^B = \varnothing$. Затем применяется следующий алгоритм синтеза функции $y_i(x,a)$.

Генерируют различные подходящие структуры-претенденты для $y_i(x,a)$, и для каждой из них рассчитывают наилучший вектор параметров a на множестве пациентов X^A с точки зрения сходства с идеальным классификатором $P_i(x)$. Затем из полученного множества претендентов выбирается тот, который лучше всего разделил диагнозы на выборке X^B, не принимавшей участие в настройке параметров классификатора.

Таким образом, мы определяем структуру классификатора, способную наилучшим образом работать на новых пациентах, не принимавших участие в настройке классификатора.

Как правило, для того чтобы иметь объективную оценку классификатора выделяют не две, а три выборки. Дополнительная выборка X^C – экзаменационная выборка, при синтезе классификатора не участвует, и потому позволяет получить полностью объективную оценку классификатора. Получив такую оценку (чувствительность и специфичность на экзамене) обычно выборку X^C включают в состав рабочей выборки и с учетом точек X^C пересчитывают классификатор, который затем и используют как «советчик» в принятии решения о диагнозе пациента, принимая во внимание, что реальные чувствительность и специфичность могут быть лучше, чем на экзамене

4. Решение задачи

Исключив из общего набора клинических признаков малоинформативные и дублирующие, введем для остальных обозначения:

x_1 – носовое кровотечение; x_2 – кровоточивость десен; x_3 – кровотечение после экстракции зубов; x_4 – интра и послеоперационное кровотечение; x_5 – послетравматическая гематома; x_6 – кровотечение из поверхностных ран; x_7 – продолжительное не заживление ран; x_8 – послетравматический гемартроз; x_9 – послеинъекционная гематома; x_{10} – послеродовое кровотечение; x_{11} – ювенильное маточное кровотечение; x_{12} – возраст.

С различными реализациями метода МГУА, которые были использованы при нахождении наилучших классификаторов $y_i(x)$-c_i i=1,..,4 можно познакомится в [3,4].

Уровень идеального классификатора $P_i(x)$, соответствующий принятию «своего» диагноза принят (см. рис. 3) A=220, уровень идеального классификатора $P_i(x)$ соответствующий принятию «прочих» диагнозов принят B=100. Уровни значения клинических признаков, соответствующих их наличию в анамнезе приняты равными числу 25, отсутствию признака в анамнезе – числу (-5), отсутствию условий для проявления признака – числу 1. Значения уровней приняты достаточно произвольно, но во взаимосвязи друг с другом.

Общее количество пациентов в группе – 80, распределение пациентов каждого диагноза в обучающей, проверочной и экзаменационной выборке следующее:

БВ - всего точек - 24 , в обучении - 17 , в проверке - 3 , экзамен – 4,

КП - всего точек - 17 , в обучении - 13 , в проверке - 2 , экзамен – 2,

ДТ - всего точек - 31 , в обучении - 24 , в проверке - 3 , экзамен – 4,

КПСГ - всего точек - 8, в обучении – 8 или 6, в проверке 0 или 2, экзамен –0.

В связи с малым объемом выборки КПСГ при расчете классификаторов БВ, КП в обучении оставлены все 8 точек КПСГ, в проверке и экзамене 0 точек КПСГ. При расчете классификатора ДТ и КПСГ в обучении оставлено 6 точек, в проверку 2 точки, в экзамен – 0.

При расчете классификатора БВ обучающая выборка по всем диагнозам – (выборка А) – 62 точек, проверочная (выборка В) – 8 точек, экзамен (выборка В) –10 точек.

При расчете классификатора КП обучающая выборка по всем диагнозам – (выборка А) –62 точек, проверочная (выборка В) – 8 точек, экзамен (выборка В) – 10 точек.

При расчете классификатора ДТ обучающая выборка по всем диагнозам – (выборка А) – 60 точек, проверочная (выборка В) – 10 точек, экзамен (выборка В) – 10 точек.

При расчете классификатора КПСГ обучающая выборка по всем диагнозам (выборка А) – 60 точек, проверочная (выборка В) – 10 точек, экзамен – 10 точек.

5. Результаты синтеза

1. Формула классификатора, дифференцирующего диагноз БВ:

У1(x)= 108.4733557 - 440.7850808/(x4*x5*x10) +
+ 1.6562130*x5*x11/(x9 *x6) +
+ 112.0633706*x7/x3*x9*x8) +
+ 0.0001365*x10*x1*x4*x10 +
+ 0.0006620*x12*11*x12 /x4 -
- 0.4547683*x6*x6/(x1*x4) +
+ 0.0001007*x12*x2*x10*x7 -
- 0.0456770*x12* *x12/(x6*x4) +
+ 5407.5841870/(x12*x5*x10*x3) -
- 0.0659565 *x4*x6*x10/x12

График классификатора БВ приведен ниже на рис. 4.

На графике:

- черная (пороговая) функция – идеальный классификатор БВ,

- серая кривая – классификатор БВ, формула для которого приведена выше,

- белая линия - порог распознавания A=150: все точки серой кривой, находящиеся выше данного порога будем относим к диагнозу БВ, прочие – не являются пациентами с данным диагнозом.

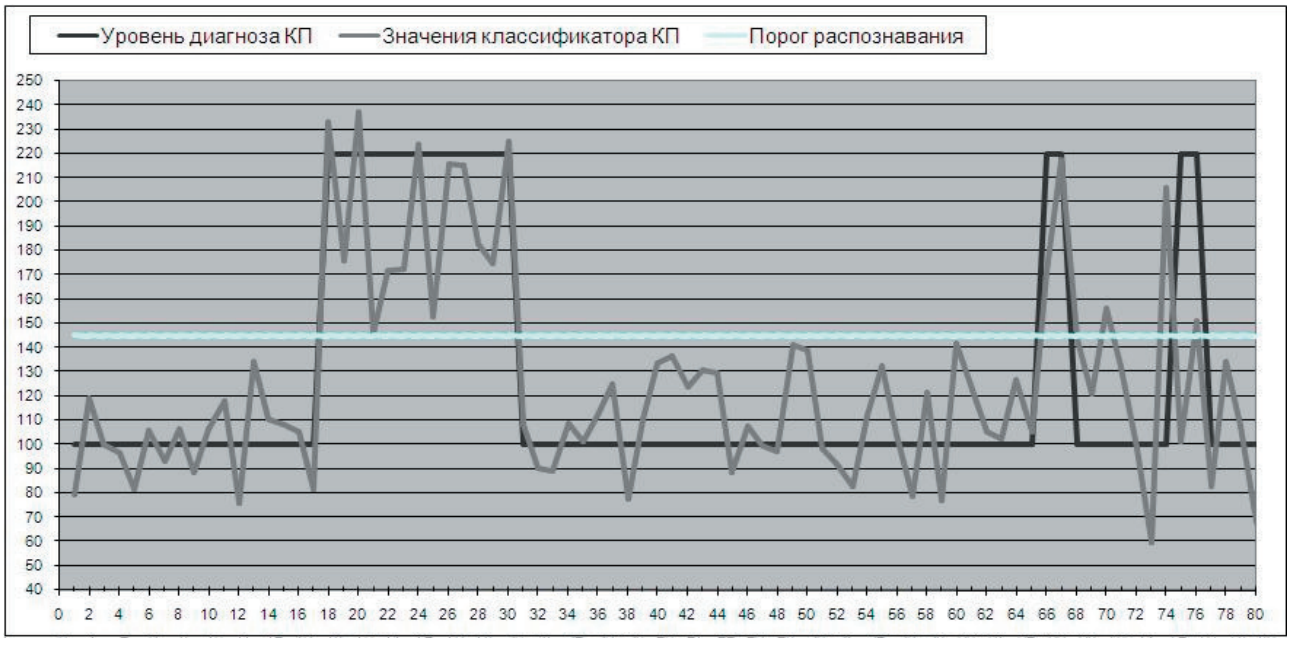

Рис. 4. График классификатора БВ

Таблица 1

Характеристики полученного классификатора БВ

		%распознавания	чувствительность	специфичность
Вся выборка	A+B+C	97,50%	0,953	0,953
Рабочая выборка	W=A+B	100,00%	1	1
Обучение	A	100,00%	1	1
Проверка	B	100,00%	1	1
Экзамен	C	80%	0,75	0,833

Для всей выборки (A+B+C) качественные характеристики классификатора устойчивы и не меняются в полосе порога от 148,88 до 164,03.

2. Формула классификатора, дифференцирующего диагноз КП:

У2(x)= 98.7962916 -
- 41.4715530*x11/(*x3*x12) +
+ 1037.1013717/x12 - 0.5270360 *x9*x5/(x4*x7) +
+ 109.3761181*x3/(x1*x6)-0.8897469*x10 +
+ 0.00123 79* x12* x3* x12 /x10 -
- 0.0527665 *x3*x12/(x10*x4) +
+ 111.8641766/(x2*x4) + 0.8899826 *x10/x3 +
+ 0.4468907 *x4*x3/(x2*x10).

График классификатора КП приведён на рис. 5.

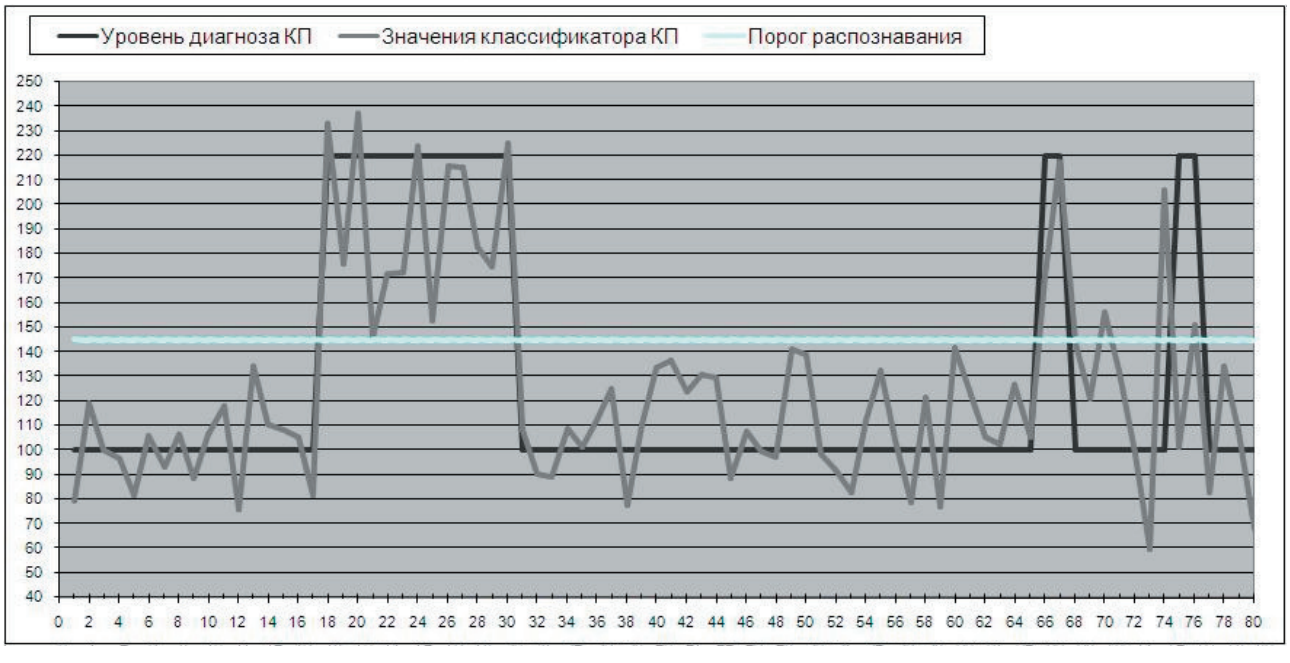

Рис. 5. График классификатора КП

На графике:

- черная (пороговая) функция – идеальный классификатор КП,

- серая кривая – классификатор КП, формула для которого приведена выше,

- белая линия - порог распознавания А=145: все точки серой кривой, находящиеся выше данного порога будем относим к диагнозу КП, прочие – не являются пациентами с данным диагнозом.

Таблица 2

Характеристики полученного классификатора КП

		%распознавания	чувствительность	специфичность
Вся выборка	A+B+C	96,25%	0,941	0,962
Рабочая выборка	W=A+B	98,57%	1	0,982
Обучение	A	100,00%	1	1
Проверка	B	87,50%	1	0,833
Экзамен	C	80%	0,5	0,875

3. Формула классификатора, дифференцирующего диагноз ДТ:

У3(x)= 146.2008995 +
+ 428.5747691/(x10*x5) -
- 10.3802641*x1*x1/(x5*x12) +
+ 1274.9558858/(x9*x7*x3*x3) +
+ 0.0435972*x1*x3 - 0.0001708* x5*x4*x2*x3 +
+ 37.5828185*x6/(*x7*x9) -
- 0.0759872*x9*x5/x4 +
+ 848.9092002*x2/(x1*x12*x9) +
+ 7326.7438409/(x1*x4*x2*x2) -
- 2.4606239*x10*x11/(x3*x12) -
- 4.2983931*x4 /(x10*x6).

График классификатора ДТ приведен на рис. 6.

На графике:

- черная (пороговая) функция – идеальный классификатор ДТ,

- серая кривая – классификатор ДТ, формула для которого приведена выше,

- белая линия – порог распознавания А=159,5: все точки серой кривой, находящиеся выше данного порога будем относим к диагнозу ДТ , прочие – не являются пациентами с данным диагнозом.

Таблица 3

Характеристики полученного классификатора ДТ

		%распознавания	чувствительность	специфичность
Вся выборка	A+B+C	90,00%	0,871	0,918
Рабочая выборка	W=A+B	95,71%	0,926	0,977
Обучение	A	96,67%	0,958	0,972
Проверка	B	90,00%	0,67	1
Экзамен	C	50,00%	0,5	0,5

4. Формула классификатора, дифференцирующего диагноз КФПГ:

У4(x) = 94.4941392 +
+ 1456.3783323/(x1*x2) -
- 32.7076625*x6/(x10*x2*x3) +
+ 0.0016226*x4*x6*x4 -
- 3418.9997737/(*x12*x3*x9) -
- 0.0185404*x12*x7*x4/x1 +
+ 0.0615398*x4*x10*x2/x12 +
+ 652.0079631 *x1/(x5*x12*x7) -
- 0.1108216 *x6*x3/x7 -
- 0.0112632 *x12*x12*/x3 -
- 48.1555120 *x3/(x1*x2*x4)

График классификатора КФПГ приведен ниже на рис. 7.

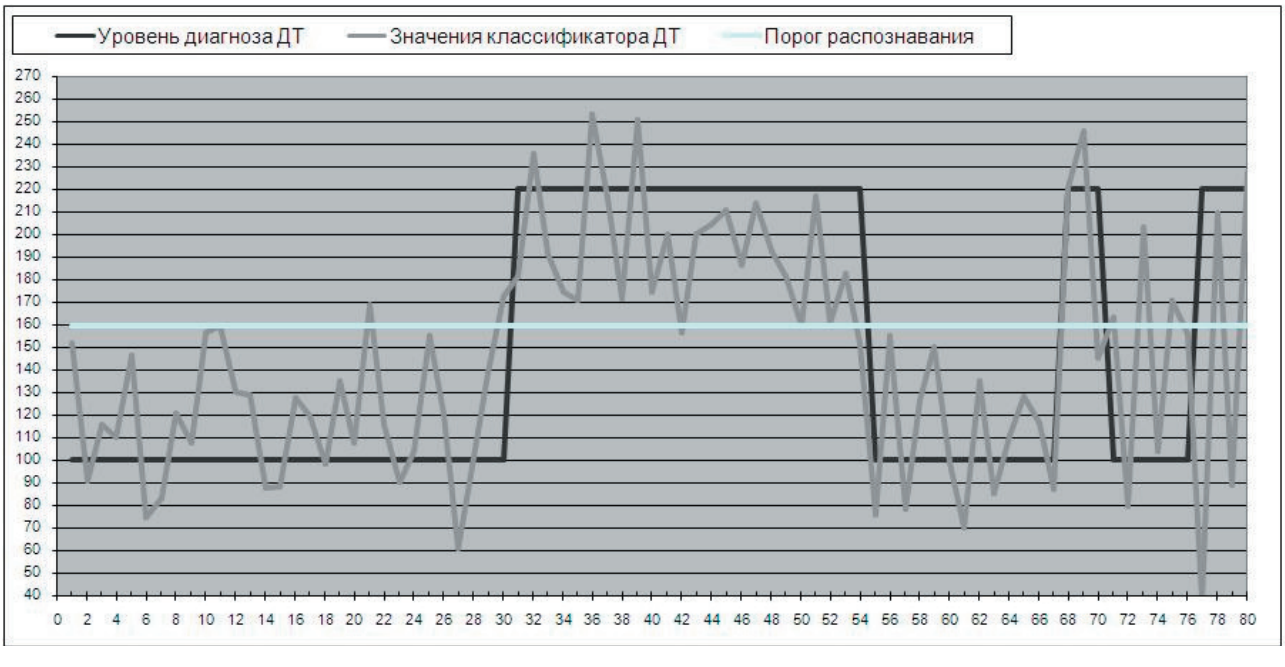

Рис. 6. График классификатора ДТ

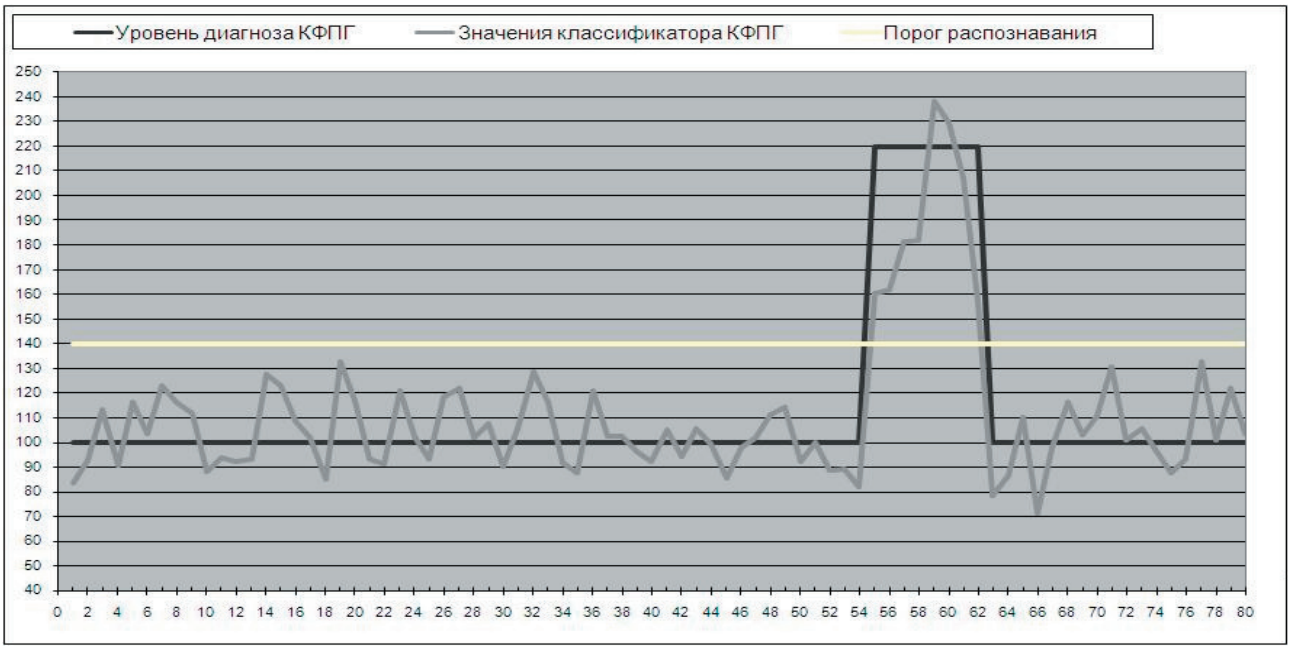

Рис. 7. График классификатора КФПГ

На графике:

- черная (пороговая) функция – идеальный классификатор КФПГ,
- серая кривая – классификатор КФПГ, формула для которого приведена выше;
- белая линия – порог распознавания A=140: все точки серой кривой, находящиеся выше данного порога будем относим к диагнозу КФПГ, прочие – не являются пациентами с данным диагнозом.

Таблица 4

Характеристики полученного классификатора КФПГ

		%распознавания	чувствительность	специфичность
Вся выборка	A+B+C	100,00%	1	1
Рабочая выборка	W=A+B	100,00%	1	1
Обучение	A	100,00%	1	1
Проверка	B	100,00%	1	1
Экзамен	C	100,00%		1

6. Выводы

1. Высокие показатели классификатора КФПГ позволяют рекомендовать его для дифференциальной диагностики. При этом, по-видимому, потребуется определенная его доработка с учетом новых данных ввиду крайне малой выборки (8 пациентов с диагнозом КФПГ) в данной группе.

2. Классификатор БВ допустил всего 2 ошибки. Однако оба неверно диагностированных пациента – номера: 74 – ложный отказ от диагноза и 79 – ложно установленный диагноз БВ, были согласованно ложно диагностированы классификаторами КП – ложный диагноз для 74 и ДТ – ложный отказ от диагноза для 79. Таким образом, двойная ошибка классификаторов делает ошибку классификатора БВ неустранимой с помощью других используемых здесь классификаторов.

4. Схожая ситуация получена для классификатора КП – получены неустранимые ошибки диагнозов для пациентов 70, 74 и 75 и классификатора ДТ – неустранимые ошибки диагнозов для пациентов 70 и 79.

5. Всего классификаторы на выборке 80-ти пациентов ошиблись на 9 пациентах – 21,42,70,71,73,74,75,77,79 при этом получено 4 неустранимых ошибки диагноза –номера 70,74,75,79. Для остальных пяти пациентов ввиду конфликта классификаторов или согласованного их отказа от диагноза, следует провести уточнение диагноза другими методами.

7. Заключение

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

1. Включение в перечень используемых признаков дополнительные характеристики диагностируемых патологий и на основе нового состава признаков разработать более надежные классификаторы.

2. Выделение отдельного класса пациентов с неустранимыми ошибками диагнозов уже разработанных классификаторов, накопление статистики таких пациентов и построение дополнительного классификатора для выделения такой специальной области значений признаков. Тогда бесконфликтная классификация диагноза будет свидетельствовать о решении задачи дифференциальной диагностики, Конфликт же диагнозов либо отнесение пациента к дополнительному классу ложных диагнозов будет однозначно означать необходимость проведения процедуры уточнения диагноза.

8. Литература

1. Ивахненко А.Г. Мюллер Й.А. Самоорганизация прогнозирующих моделей. Киев: Техника, 1985. 219 с.
2. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. — Киев: «Наук.думка», 1985, - 216 с.
3. Многорядный алгоритм веерных решений. Кондрашова Н.В., Павлов В.А., Павлов А.В. -Вісник національного технічного університету України «КПІ». Інформатика, управління та обчислювальна техніка. №45, 2006, с. 218-228
4. Павлов А.В. "Модифицированный алгоритм с комбинаторной селекцией переменных и его анализ", стр. 130-139. Збірник наукових праць "Індуктивне моделювання складних систем", Випуск 2 , Київ 2010.

У статті розглядаються переваги застосування онтологічного підходу під час проектування інформаційних порталів. Виконано аналіз останніх досліджень та публікацій за проблематикою, наведена модель пошуку інформації, що враховує сферу інтересів користувача порталу

Ключові слова: інформаційний портал, онтологічній підхід, переваги застосування

В статье рассматриваются преимущества применения онтологічного похода при проектировании информационных порталов. Выполнен анализ последних исследований и публикаций по проблематике, приведена модель поиска информации, учитывающая сферу интересов пользователя портала

Ключевые слова: информационный портал, онтологический подход, преимущества использования

In the article the advantages of using the ontological approach to design informational portals were considered. An analysis of recent studies and publications on problematic was done, the model of search of information is presented which takes into account the interests of the user portal

Key words: informational portal, the ontological approach, advantages of using

УДК 025.4.03

ПРОЕКТУВАННЯ ІНФОРМАЦІЙНИХ ПОРТАЛІВ — ПЕРЕВАГИ ЗАСТОСУВАННЯ ОНТОЛОГІЧНОГО ПІДХОДУ

Н.А. Хміль
Кандидат педагогічних наук, доцент*
Контактний тел.: (057) 702-15-91
E-mail: abc250@yandex.ru

А.В. Прилепо
Асистент*
*Кафедра соціальної інформатики
Харківський національний університет радіоелектроніки
пр. Леніна, 14, м. Харків, Україна, 61108
Контактний тел.: (057) 702-15-91
E-mail: si@kture.kharkov.ua

1. Вступ

Однією з актуальних задач сучасного розвитку інформаційного суспільства є проектування інформаційних порталів, які на сьогодні можна вважати одним з домінуючих Інтернет-рішень для систематизації інформації, доступу до неї та ефективного її використання.

Останнім часом усе частіше для проектування таких порталів стали використовуватися онтології, які здатні точно і ефективно описувати семантику даних для деякої предметної галузі і вирішувати проблему несумісності і суперечності понять. Так, наприклад, онтології в мережі Інтернет варіюються від великих таксономій, які категорізують веб-сайти (як на сайті Yahoo!), до категоризації товарів, які продаються та їх характеристик (як на сайті Amazon.com) [1].

2. Аналіз останніх досліджень і публікацій

Серед науковців, які досліджували різні аспекти проектування порталів можна виділити В. Г. Грищен-