ПРИКЛАДНАЯ МАТЕМАТИКА

УДК 519.3+519.7

В. В. Демьянова

ПРОГНОЗИРОВАНИЕ ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ СПОСОБОВ ЛЕЧЕНИЯ

1. Введение. В статьях [1–3] была описана методика прогнозирования эффективности применения различных способов обучения или лечения. В настоящей работе эта методика применяется для прогнозирования эффективности применения химио- и/или гормональной терапии при лечении онкологических заболеваний. В качестве примера взята база данных CHEMO-253 (см. [4]), хранящаяся в репозитории Висконсинского университета и широко используемая исследователями (см. [5]).

Эта база содержит сведения о 253 пациентах, больных раком молочной железы, которым была сделана хирургическая операция по удалению опухоли. Каждый из пациентов описан точкой в 39-мерном пространстве, представляющей информацию о результатах анализов (30 параметров), о том, какой курс лечения был проведен после операции (был ли пациент подвергнут химио- или гормональной терапии или нет), а также сведения о продолжительности его жизни после операции в месяцах (наблюдения велись в течение 13 лет), размере опухоли, наличии и количестве метастаз.

Для каждого пациента известны результаты применения (или неприменения) химио- и гормотерапии. В результате были получены четыре базы данных о пациентах, подвергшихся хирургической операции: о тех, кому не была сделана ни химио-, ни гормональная терапия; о прошедших курс только химиотерапии; о прошедших курс только гормональной терапии; о прошедших курс и химио-, и гормональной терапии. Предполагается, что эти базы представляют собой репрезентативные выборки из множества пациентов, подвергшихся хирургической операции.

Для каждой из баз (являющейся *обучающей выборкой*) строится критерий (называемый *идентификатором*, или решающим правилом – РП), позволяющий предсказать результат лечения. При этом получаются и вероятности этих предсказаний. Данный критерий затем применяется к другим базам (служащим *контрольными выборками*). В результате все пациенты делятся на 16 прогностических групп. Для каждой группы даются вероятности благоприятного исхода в случае и применения химио- и/или гормотерапии, и неприменения.

Для построения РП выбирается несколько наиболее информативных признаков, и в пространстве этих признаков проводится разделение множеств пациентов с благоприятным и неблагоприятным исходами.

В работе [3] обсуждался случай, когда химио- и гормональная терапии рассматривались как один тип терапии (именно так эта задача обычно и изучается – см. [4, 5]). Ниже показывается, что раздельное исследование каждой из возможных терапий

позволяет получить более точный, чем в [3], прогноз и увеличить количество пациентов с благоприятным исходом.

2. Постановка задачи. Приведем результаты анализа базы данных CHEMO-253 (см. [4]).

К пациенту, которому сделана хирургическая операция по удалению опухоли, может быть либо применена, либо нет химио- и/или гормональная терапия; или не применена ни химио-, ни гормональная терапия. Как уже отмечалось, имеются четыре базы данных о пациентах, подвергшихся хирургической операции. Обозначим базу данных о пациентах, которым не была сделана ни химио-, ни гормональная терапия, WCT-113 (или база I), базу данных о пациентах, прошедших курс только химиотерапии, – CT-33 (или база II), базу данных о пациентах, которым назначен курс только гормональной терапии, – HT-49 (или база III), базу данных о пациентах, получивших курс и химиотерапии, и гормональной терапии, – CHT-58 (или база IV).

Будем считать, что операция (с гормо- и/или химиотерапией или без нее) прошла успешно, если срок жизни пациента после операции не менее 5 лет (60 месяцев), и неудачно – если он был меньше.

База I (WCT-113) содержит сведения о 113 пациентах, из них 54 жили не менее 5 лет (множество этих пациентов обозначим A_1), а 59 – менее 5 лет (их множество – B_1).

База II (CT-33) включает сведения о 33 пациентах, из них 6 жили не менее 5 лет (множество этих пациентов обозначим A_2), а 27 – менее 5 лет (их множество – B_2).

База III (HT-49) содержит сведения о 49 пациентах, из них 23 жили не менее 5 лет (множество этих пациентов обозначим A_3), а 26 – менее 5 лет (их множество – B_3).

База IV (CHT-58) включает сведения о 58 пациентах, из них 32 жили не менее 5 лет (множество этих пациентов обозначим A_4), а 26 – менее 5 лет (их множество – B_4).

Информация о базах I–IV сведена в табл. 1. Применение или неприменение химио- и гормональной терапии указано соответственно числами 1 или 0 в соответствующей графе. Предполагается, что эти базы представляют собой репрезентативные выборки из множества пациентов, подвергшихся хирургической операции. К сожалению, находящаяся в нашем распоряжении база CHEMO-253 не может считаться достаточно репрезентативной (особенно база CT-33), поэтому выводы, которые будут сделаны ниже, имеют только иллюстративный характер. Для практического использования рекомендаций необходимо взять более обширную базу данных. Однако мы выбрали именно базу CHEMO-253, поскольку она общедоступна и используется многими исследователями для сравнения эффективности методов идентификации.

В работе рассматривается следующая задача: найти критерий, с помощью которого для каждого конкретного пациента можно определить, следует ли ему рекомендовать химио- и/или гормотерапию, либо эти процедуры ему противопоказаны (т. е. дать прогноз о продолжительности жизни в случаях, если химио(гормо)терапия будет назначена и когда не назначена).

3. Разделение баз I–IV с помощью параметров 30 и 35. Как и в [5], в проводимом нами исследовании было выделено несколько наиболее информативных параметров. В настоящем разделе приводятся данные, полученные с помощью двух параметров: наибольшая гладкость (worst smoothness) и наибольшая фрактальность (worst fractal dimension). Эти параметры были найдены по методике, описанной в [6, 7], они отличались от тех, которые использовались в [5]. Выбранные параметры в базе [4] имеют соответственно номера 30 и 35. В силу сказанного выше, WCT-113 $= A_1 \cup B_1$, CT-33 $= A_2 \cup B_2$, HT-49 $= A_3 \cup B_3$, CHT-58 $= A_4 \cup B_4$.

Таблица 1. **Разбиение базы CHEMO-253 на подбазы I–IV**

База	chemo	hormo	$\geqslant 60$	< 60	Σ
WCT-113 (I)	0	0	54	59	113
CT-33 (II)	1	0	6	27	33
HT-49 (III)	0	1	23	26	49
CHT-58 (IV)	1	1	32	26	58

Для каждой из баз была найдена гиперплоскость (в данном случае прямая в двумерном пространстве), наилучшим способом (в смысле общего количества неверно идентифицированных точек – см. [8]) разделяющая множества A_i и B_i.

Для базы WCT-113 была построена прямая $L_1 = \{x \in R^2 \mid h_1(x) = 0\}$, где $h_1(x)$ – линейная функция:

$$h_1(x) = (x, y_1) + d_1,\ x \in R^2,\ y_1 \in R^2,\ d_1 \in R,$$

$$y_1 = (-0,744728,\ -0,667374),\ d_1 = 0,163306.$$

С помощью этой прямой проводилась идентификация точек множества WCT-113 следующим образом:

если $h_1(c) \leqslant 0$, то считаем, что $c \in A_1$,

если $h_1(c) > 0$, то считаем, что $c \in B_1$.

Аналогично для базы CT-33 была построена прямая $L_2 = \{x \in R^2 \mid h_2(x) = 0\}$, где

$$h_2(x) = (x, y_2) + d_2,\ x \in R^2,\ y_2 \in R^2,\ d_2 \in R,$$

$$y_2 = (-0,220719,\ -0,97534),\ d_2 = 0,1111885.$$

С ее помощью проводилась идентификация точек множества CT-33:

если $h_2(c) \leqslant 0$, то считаем, что $c \in A_2$,

если $h_2(c) > 0$, то считаем, что $c \in B_2$.

Для базы HT-49 была построена прямая $L_3 = \{x \in R^2 \mid h_3(x) = 0\}$, где

$$h_3(x) = (x, y_3) + d_3,\ x \in R^2,\ y_3 \in R^2,\ d_3 \in R,$$

$$y_3 = (-0,730601,\ -0,68280),\ d_3 = 0,151446,$$

с помощью которой проводилась идентификация точек множества HT-49:

если $h_3(c) \leqslant 0$, то считаем, что $c \in A_3$,

если $h_3(c) > 0$, то считаем, что $c \in B_3$.

Для базы CHT-58 была построена прямая $L_4 = \{x \in R^2 \mid h_4(x) = 0\}$, где

$$h_4(x) = (x, y_4) + d_4,\ x \in R^2,\ y_4 \in R^2,\ d_4 \in R,$$

$$y_4 = (-0,321903,\ -0,946772),\ d_4 = 0,131791.$$

С ее помощью проводилась идентификация точек множества CHT-58:

если $h_4(c) \leqslant 0$, то считаем, что $c \in A_4$,

если $h_4(c) > 0$, то считаем, что $c \in B_4$.

Прямые $L_i (i \in 1:4)$ разделяли соответствующие множества неточно (см. табл. 2–9 ниже). Заметим, что угол между прямыми L_1 и L_2 оказался равным $35,38°$, между

прямыми L_1 и L_3 – $1,19648°$, между прямыми L_1 и L_4 – $29,357°$, между прямыми L_2 и L_3 – $34,186°$, между прямыми L_2 и L_4 – $6,029°$, между прямыми L_3 и L_4 – $28,159°$.

3.1. Разделение базы WCT-113. В табл. 2 приведены результаты разделения базы WCT-113 с помощью прямой L_1 и указанного выше идентификационного правила. Установлено, что

из 54 точек множества A_1 значения функции h_1 оказались положительными для 20 и отрицательными – для 34 (т. е. правильно были идентифицированы 34 точки, или 63%, а неправильно – 20, или 37%);

из 59 точек множества B_1 значения функции h_1 оказались положительными для 37 и отрицательными – для 22 (т.е. правильно были идентифицированы 37 точек, или 62,7%, а неправильно – 22, или 37,3%).

Таблица 2. **Результаты разделения базы WCT-113 с помощью прямой L_1**

Множество	h_1+	h_1-	Σ	Количество, %
A_1	20	34	54	63
B_1	37	22	59	62,7
Σ	**57**	**56**	**113**	

Таблица 3. **Данные идентификации (в %) в группах h_1+ и h_1- базы WCT-113 с помощью прямой L_1 как точек множеств A_1 и B_1**

Множество	h_1+	h_1-
A_1	35,1	60,7
B_1	64,9	39,3
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_1 и B_1 в группах h_1+ и h_1- для базы WCT-113 с помощью прямой L_1 приведено в табл. 3.

3.2. Разделение базы CT-33. В табл. 4 указаны результаты разделения базы CT-33 с помощью прямой L_2 и указанного выше идентификационного правила:

из 6 точек множества A_2 значения функции h_2 оказались положительными для 1 и отрицательными – для 5 (т. е. правильно были идентифицированы 5 точек, или 83,3%, а неправильно – 1, или 16,7%);

из 27 точек множества B_2 значения функции h_2 оказались положительными для 22 и отрицательными – для 5 (т. е. правильно были идентифицированы 22 точки, или 81,5%, а неправильно – 5, или 18,5%).

Таблица 4. **Результаты разделения базы CT-33 с помощью прямой L_2**

Множество	h_2+	h_2-	Σ	Количество, %
A_2	1	5	6	83,3
B_2	22	5	27	81,5
Σ	**23**	**10**	**33**	

Таблица 5. **Данные идентификации (в %) в группах h_2+ и h_2- базы CT-33 с помощью прямой L_2 как точек множеств A_2 и B_2**

Множество	h_2+	h_2-
A_2	4,3	50
B_2	95,7	50
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_2 и B_2 в группах h_2+ и h_2- для базы CT-33 с помощью прямой L_2 приведено в табл. 5.

3.3. Разделение базы HT-49. В табл. 6 указаны результаты разделения базы HT-49 с помощью прямой L_3 и указанного выше идентификационного правила. Было установлено, что

Таблица 6. **Результаты разделения базы НТ-49 с помощью прямой L_3**

Множество	h_3+	h_3-	Σ	Количество, %
A_3	10	13	23	56,5
B_3	15	11	26	57,7
Σ	**25**	**24**	**49**	

Таблица 7. **Данные идентификации (в %) в группах h_3+ и h_3- базы НТ-49 с помощью прямой L_3 как точек множеств A_3 и B_3**

Множество	h_3+	h_3-
A_3	40,0	54,2
B_3	60,0	45,8
Σ	**100**	**100**

из 23 точек множества A_3 значения функции h_3 оказались положительными для 10 и отрицательными – для 13 (т. е. правильно были идентифицированы 13 точек, или 56,5%, а неправильно – 10, или 43,5%);

из 26 точек множества B_3 значения функции h_3 оказались положительными для 15 и отрицательными – для 11 (т. е. правильно были идентифицированы 15 точек, или 57,7%, а неправильно – 11, или 42,3%).

Количество правильно идентифицированных точек как точек множеств A_3 и B_3 в группах h_3+ и h_3- для базы НТ-49 с помощью прямой L_3 приведено в табл. 7.

3.4. Разделение базы СНТ-58. В табл. 8 указаны результаты разделения базы СНТ-58 с помощью прямой L_4 и указанного выше идентификационного правила. Установлено, что

из 32 точек множества A_4 значения функции h_4 оказались положительными для 12 и отрицательными – для 20 (т. е. правильно были идентифицированы 20 точек, или 62,5%, а неправильно – 12, или 37,5%);

из 26 точек множества B_4 значения функции h_4 оказались положительными для 17 и отрицательными – для 9 (т. е. правильно были идентифицированы 17 точек, или 65,4%, а неправильно – 9, или 34,6%).

Таблица 8. **Результаты разделения базы СНТ-58 с помощью прямой L_4**

Множество	h_4+	h_4-	Σ	Количество, %
A_4	12	20	32	62,5
B_4	17	9	26	65,4
Σ	**29**	**29**	**58**	

Таблица 9. **Данные идентификации (в %) в группах h_4+ и h_4- базы СНТ-58 с помощью прямой L_4 как точек множеств A_4 и B_4**

Множество	h_4+	h_4-
A_4	41,4	69,0
B_4	58,6	31,0
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_4 и B_4 в группах h_4+ и h_4- для базы СНТ-58 с помощью прямой L_4 приведено в табл. 9.

3.5. Перекрестное исследование баз I–IV с помощью прямых $L_1 - L_4$. Каждую из баз I–IV исследуем теперь с помощью других прямых, т. е. выясним, каков прогноз эффективности применения других способов лечения. В результате все 253 пациента будут разделены на 16 прогностических групп:

группа 0000 – пациенты, для которых все четыре прогноза неблагоприятны;

группа 0001 – пациенты, для которых прогноз благоприятен только в случае одновременного применения и гормо-, и химиотерапии;

Таблица 10. **Результаты разделения баз I–IV с помощью прямых $L_1 - L_4$**

Группа	A_1	B_1	A_2	B_2	A_3	B_3	A_4	B_4	Σ	p	T	E	**A**	Δ
0000	8	23	1	16	10	15	**6**	**7**	86	0,414	IV	35,6	25	10,6
0001	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
0010	6	11	0	4	**2**	**1**	2	4	30	0,542	III	16,25	10	6
0011	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
0100	1	0	**0**	**0**	0	0	1	0	2	0,500	II	1	2	-1
0101	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
0110	4	2	2	1	**2**	**3**	0	1	15	0,542	III	8,13	8	0
0111	1	1	0	0	0	0	**0**	**1**	3	0,690	IV	2,07	1	1
1000	**0**	**0**	0	0	0	0	0	0	0	0,607	I	0	0	
1001	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
1010	**1**	**1**	0	2	1	0	1	2	8	0,607	I	4,86	3	2
1011	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
1100	**0**	**0**	0	0	0	0	0	0	0	0,607	I	0	0	
1101	0	0	0	0	0	0	**0**	**0**	0	0,690	IV	0	0	
1110	**3**	**5**	0	1	1	2	2	3	17	0,607	I	10,32	6	4
1111	30	16	3	3	7	5	**20**	**8**	92	0,690	IV	63,45	60	3
Σ	**54**	**59**	**6**	**27**	**23**	**26**	**32**	**26**	**253**			141,6	115	26,6
p_i	**0,61**	0,35	**0,5**	0,04	**0,54**	0,4	**0,69**	0,41						

группа 0010 – пациенты, для которых прогноз благоприятен только при применении гормональной терапии;

группа 0011 – пациенты, для которых прогноз благоприятен в случаях применения гормональной терапии, одновременного применения гормо- и химиотерапии, и неблагоприятен в случае отсутствия терапии или применения химиотерапии;

группа 0100 – пациенты, для которых прогноз благоприятен только при применении химиотерапии;

группа 0101 – пациенты, для которых прогноз благоприятен в случаях применения химиотерапии, одновременного применения гормо- и химиотерапии, в случае же отсутствия терапии или применения гормональной терапии он неблагоприятный;

группа 0110 – пациенты, для которых прогноз благоприятен в случаях применения и химио-, и гормональной терапии, когда же отсутствует терапия или применяется одновременно и гормональная, и химиотерапия, прогноз неблагоприятный;

группа 0111 – пациенты, для которых прогноз благоприятен при применении и химио-, и гормональной терапии, и одновременного применения гормо- и химиотерапии, когда же терапия не назначена, прогноз неблагоприятный;

группа 1000 – пациенты, для которых прогноз благоприятен только в случае, когда не применяется ни гормональная, ни химиотерапия;

группа 1001 – пациенты, для которых прогноз благоприятен в случаях и отсутствия, и одновременного применения гормо- и химиотерапии;

группа 1010 – пациенты, для которых прогноз благоприятен в случае неприменения терапии и применения гормональной терапии;

группа 1011 – пациенты, для которых прогноз благоприятен в случаях отсутствия терапии, применения гормональной терапии, одновременного применения гормо- и химиотерапии, и неблагоприятен при применении химиотерапии;

группа 1100 – пациенты, для которых прогноз благоприятен при отсутствии терапии и в случае применения химиотерапии, когда же применяется гормональная терапия или одновременно гормо- и химиотерапия – прогноз неблагоприятный;

группа 1101 – пациенты, для которых прогноз благоприятен в случаях отсутствия терапии, применения химиотерапии, одновременного применения гормо- и химиотерапии, в случае же применения гормональной терапии прогноз неблагоприятный;

группа 1110 – пациенты, для которых прогноз благоприятен и при отсутствии терапии, и в случаях применения химио- или гормональной терапии, когда же применяется одновременно гормо- и химиотерапия, то прогноз неблагоприятный;

группа 1111 – пациенты, для которых прогноз благоприятен во всех случаях.

Результаты разделения баз I–IV плоскостями $L_1 - L_4$ сведены в табл. 10. В последней строке этой таблицы указаны вероятности благоприятного исхода в случае попадания в соответствующее подмножество (A_i или B_i) (см. табл. 3, 5, 7 и 9). Жирным шрифтом дается вероятность благоприятного исхода, если пациент оказался в соответствующем подмножестве с благоприятным исходом (т. е. в подмножестве A_i) при применении данной терапии (i-й). Так, если пациент имеет благоприятный прогноз при применении только гормональной терапии (т. е. в четырехзначном номере его группы на третьем месте стоит 1), то вероятность благоприятного исхода равна 0,542 (см. табл. 7); если же он неблагоприятный (т. е. в номере его группы на третьем месте стоит 0), то вероятность благоприятного исхода равна 0,4.

Для каждой группы (0000, 0001, 0010 и т. д.) в соответствующей ей строке указано количество пациентов из каждого подмножества (A_i и B_i) базы $i, i \in$ {I, II, III, IV}, идентифицированных как пациенты этой группы.

Так, в группе 0000 (т. е. в группе пациентов, для которых любой способ лечения имеет неблагоприятный прогноз) из подмножества A_1 базы I оказалось 8 пациентов, из подмножества B_1 базы I – 23; из подмножества A_2 базы II – 1 пациент, из подмножества B_2 базы II – 16; из подмножества A_3 базы III – 10 пациентов, из подмножества B_3 базы III – 15; из подмножества A_4 базы IV – 6 пациентов, из подмножества B_4 базы IV – 7. Общее количество пациентов, попавших в группу 0000, – 86. Поскольку любой способ лечения для них имеет неблагоприятный прогноз, следует выбрать способ лечения, имеющий наибольшую вероятность успеха. В данном случае вероятность благоприятного исхода первого способа лечения (никакой терапии) 0,351, второго способа лечения (только химиотерапия) – 0,043, третьего способа (только гормональная терапия) – 0,4, четвертого способа (и гормо-, и химиотерапия) – 0,414. Максимальную вероятность успеха имеет четвертый способ лечения (0,414). Она и приведена в столбце p. В столбце T указан рекомендуемый способ лечения (IV). В столбце E дается математическое ожидание количества пациентов с благоприятным исходом ($86 \times 0{,}414 = 35{,}6$); в столбце **А** – количество пациентов с благоприятным исходом при применении тех методов лечения, которым они были реально подвергнуты ($8 + 1 + 10 + 6 = 25$). Наконец, в столбце Δ приводится математическое ожидание прироста количества пациентов с благоприятным исходом в случае применения рекомендуемого (IV) способа лечения (в данном случае прирост равен 10,59).

Рассмотрим еще группу 0110 (т. е. пациентов, для которых благоприятный исход прогнозируется в случае применения только либо химиотерапии, либо гормональной терапии; в случае отсутствия терапии или одновременного применения и химио-, и гормональной терапии прогноз неблагоприятный). В этой группе из подмножества A_1 базы I оказалось 4 пациента, из подмножества B_1 базы I – 2 пациента; из подмножества A_2 базы II – 2 пациента, из подмножества B_2 базы II – 1; из подмножества A_3 базы III – 2 пациента, из подмножества B_3 базы III – 3; из подмножества A_4 базы IV – нет пациентов, из подмножества B_4 базы IV оказался 1 пациент. Общее количество пациентов, попавших в группу 0110, – 15. Следует выбрать способ лечения, имеющий наибольшую вероятность успеха. В данном случае вероятность благоприятного исхода первого способа лечения (без терапии) составляет 0,351 (поскольку у пациента неблагоприятный прогноз при применении первого способа лечения), второго способа лечения (только химиотерапия) – 0,5 (у пациента благоприятный прогноз при применении второго способа лечения), третьего способа (только гормональная терапия) – 0,542 (у пациента благоприятный прогноз при применении третьего способа лечения), четвертого способа (и гормо-, и химиотерапия) – 0,414 (так как у пациента неблагоприятный прогноз при применении четвертого способа лечения). Максимальную вероятность успеха имеет третий способ лечения (0,542). Она приведена в столбце p. В столбце T указан рекомендуемый способ лечения (III). В столбце E дается математическое ожидание количества пациентов с благоприятным исходом ($15 \times 0{,}542 = 8{,}13$); в столбце **А** – количество пациентов с благоприятным исходом при применении тех методов лечения, которым они были реально подвергнуты ($4 + 2 + 2 + 0 = 8$). Наконец, в столбце Δ приводится математическое ожидание прироста количества пациентов с благоприятным исходом в случае применения рекомендуемого (III) способа лечения (в данном случае прирост практически равен нулю).

Из табл. 10 (см. строку Σ, в которой приводится сумма чисел по каждому столбцу) следует, что при применении рекомендуемых способов лечения к каждой из 16 групп (с учетом вероятностей благоприятного исхода) количество пациентов с благоприятным прогнозом составит 141,6 человек (вместо 115 в действительности), т. е. на 26,6 человек

больше. Ниже будет показано, что увеличение точности разделения множеств A_i и B_i может привести к увеличению количества пациентов с благоприятным прогнозом (при применении рекомендуемого способа лечения).

4. Разделение баз I–IV с помощью параметров 30, 33 и 35. Теперь исследуем базы I–IV, используя три параметра: наибольшая гладкость (worst smoothness), точка наибольшего изгиба (worst concave point) и наибольшая фрактальность (worst fractal dimension). Выбранные параметры в базе [4] имеют соответственно номера 30, 33 и 35. Эти же параметры изучались в [9].

Для каждой из баз I–IV была найдена гиперплоскость (в данном случае плоскость в трехмерном пространстве), разделяющая (некоторым образом) множества A_i и B_i.

Для базы WCT-113 была построена плоскость $L_1 = \{x \in R^3 \mid h_1(x) = 0\}$, где $h_1(x)$ – линейная функция:

$$h_1(x) = (x, y_1) + d_1,\ x \in R^3,\ y_1 \in R^3,\ d_1 \in R,$$

$$y_1 = (-0{,}8458,\ 0{,}2700,\ 0{,}4602),\ d_1 = 0{,}0325.$$

С помощью этой плоскости проводилась идентификация точек множества WCT-113 следующим образом:

если $h_1(c) \leqslant 0$, то считаем, что $c \in A_1$,

если $h_1(c) > 0$, то считаем, что $c \in B_1$.

Аналогично для базы CT-33 была построена плоскость $L_2 = \{x \in R^3 \mid h_2(x) = 0\}$, где

$$h_2(x) = (x, y_2) + d_2,\ x \in R^3,\ y_2 \in R^3,\ d_2 \in R,$$

$$y_2 = (-0{,}8907,\ -0{,}4300,\ -0{,}1473),\ d_2 = 0{,}1799,$$

с помощью которой проводилась идентификация точек множества CT-33:

если $h_2(c) \leqslant 0$, то принимаем, что $c \in A_2$,

если $h_2(c) > 0$, то принимаем, что $c \in B_2$.

Для базы HT-49 была построена плоскость $L_3 = \{x \in R^3 \mid h_3(x) = 0\}$, где

$$h_3(x) = (x, y_3) + d_3,\ x \in R^3,\ y_3 \in R^3,\ d_3 \in R,$$

$$y_3 = (-0{,}7827,\ -0{,}4550,\ 0{,}4247),\ d_3 = 0{,}1465.$$

С ее помощью проводилась идентификация точек множества HT-49:

если $h_3(c) \leqslant 0$, то считаем, что $c \in A_3$,

если $h_3(c) > 0$, то считаем, что $c \in B_3$.

Для базы CHT-58 была построена плоскость $L_4 = \{x \in R^3 \mid h_4(x) = 0\}$, где

$$h_4(x) = (x, y_4) + d_4,\ x \in R^3,\ y_4 \in R^3,\ d_4 \in R,$$

$$y_4 = (-0{,}1442,\ 0{,}9892,\ 0{,}0269),\ d_4 = 0{,}1974.$$

С помощью этой плоскости проводилась идентификация точек множества CHT-58 следующим образом:

если $h_4(c) \leqslant 0$, то принимаем, что $c \in A_4$;

если $h_4(c) > 0$, то принимаем, что $c \in B_4$.

4.1. Разделение базы WCT-113. В табл. 11 указаны результаты разделения базы WCT-113 с помощью плоскости L_1 и указанного выше идентификационного правила. В результате

из 54 точек множества A_1 значения функции h_1 оказались положительными для 9 и отрицательными – для 45 (т. е. правильно были идентифицированы 45 точек, или 83,3%, а неправильно – 9, или 16,7%);

из 59 точек множества B_1 значения функции h_1 оказались положительными для 35 и отрицательными – для 24 (т. е. правильно были идентифицированы 35 точек, или 59,3%, а неправильно – 24, или 40,7%).

Таблица 11. **Результаты разделения базы WCT-113 с помощью плоскости L_1**

Множество	h_1+	h_1-	Σ	Количество, %
A_1	9	45	54	83,3
B_1	35	24	59	59,3
Σ	**44**	**69**	**113**	

Таблица 12. **Данные идентификации (в %) в группах h_1+ и h_1- базы WCT-113 с помощью плоскости L_1 как точек множеств A_1 и B_1**

Множество	h_1+	h_1-
A_1	20,5	65,2
B_1	79,5	34,8
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_1 и B_1 в группах h_1+ и h_1- для базы WCT-113 с помощью плоскости L_1 приведено в табл. 12.

4.2. Разделение базы CT-33. В табл. 13 указаны результаты разделения базы CT-33 с помощью плоскости L_2 и указанного выше идентификационного правила. Установлено, что

из 6 точек множества A_2 значения функции h_2 оказались положительными для 2 и отрицательными – для 4 (т. е. правильно были идентифицированы 4 точки, или 66,7%, а неправильно – 2, или 33,3%);

из 27 точек множества B_2 значения функции h_2 оказались положительными для 23 и отрицательными – для 4 (т. е. правильно были идентифицированы 23 точки, или 85,2%, а неправильно – 4, или 14,8%).

Таблица 13. **Результаты разделения базы CT-33 с помощью плоскости L_2**

Множество	h_2+	h_2-	Σ	Количество, %
A_2	2	4	6	66,7
B_2	23	4	27	85,2
Σ	**25**	**8**	**33**	

Таблица 14. **Данные идентификации (в %) в группах h_2+ и h_2- базы CT-33 с помощью плоскости L_2 как точек множеств A_2 и B_2**

Множество	h_2+	h_2-
A_2	8	50
B_2	92	50
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_2 и B_2 в группах h_2+ и h_2- для базы CT-33 с помощью плоскости L_2 приведено в табл. 14.

4.3. Разделение базы HT-49. В табл. 15 указаны результаты разделения базы HT-49 с помощью плоскости L_3 и указанного выше идентификационного правила:

из 23 точек множества A_3 значения функции h_3 оказались положительными для 9 и отрицательными – для 14 (т. е. правильно были идентифицированы 14 точек, или 60,9%, а неправильно – 9, или 39,1%);

из 26 точек множества B_3 значения функции h_3 оказались положительными для 17 и отрицательными – для 9 (т. е. правильно были идентифицированы 17 точек, или 65,4%, а неправильно – 9, или 34,6%).

Таблица 15. **Результаты разделения базы НТ-49 с помощью плоскости L_3**

Множество	h_3+	h_3-	Σ	Количество, %
A_3	9	14	23	60,9
B_3	17	9	26	65,4
Σ	**26**	**23**	**49**	

Таблица 16. **Данные идентификации (в %) в группах h_3+ и h_3- базы НТ-49 с помощью плоскости L_3 как точек множеств A_3 и B_3**

Множество	h_3+	h_3-
A_3	34,6	60,9
B_3	65,4	39,1
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_3 и B_3 в группах h_3+ и h_3- для базы НТ-49 с помощью плоскости L_3 приведено в табл. 16.

4.4. Разделение базы СНТ-58. В табл. 17 указаны результаты разделения базы СНТ-58 с помощью плоскости L_4 и указанного выше идентификационного правила. В результате

из 32 точек множества A_4 значения функции h_4 оказались положительными для 9 и отрицательными – для 23 (т. е. правильно были идентифицированы 23 точки, или 71,9%, а неправильно – 9, или 28,1%);

из 26 точек множества B_4 значения функции h_4 оказались положительными для 17 и отрицательными – для 9 (т. е. правильно были идентифицированы 17 точек, или 65,4%, а неправильно – 9, или 34,6%).

Таблица 17. **Результаты разделения базы СНТ-58 с помощью плоскости L_4**

Множество	h_4+	h_4-	Σ	Количество, %
A_4	9	23	32	71,9
B_4	17	9	26	65,4
Σ	**26**	**32**	**58**	

Таблица 18. **Данные идентификации (в %) в группах h_4+ и h_4- базы СНТ-58 с помощью плоскости L_4 как точек множеств A_4 и B_4**

Множество	h_4+	h_4-
A_4	34,6	71,9
B_4	65,4	28,1
Σ	**100**	**100**

Количество правильно идентифицированных точек как точек множеств A_4 и B_4 в группах h_4+ и h_4- для базы СНТ-58 с помощью плоскости L_4 приведено в табл. 18.

4.5. Перекрестное исследование баз I–IV с помощью плоскостей $L_1 - L_4$. Каждую из баз I–IV исследуем теперь с помощью других плоскостей, т. е. выясним, каков прогноз эффективности применения иных способов лечения. В результате все 253 пациента будут разделены на 16 прогностических групп. Описание этих групп (0000, 0001, 0010, 0011,...,1111) см. в п. 3.5.

Результаты разделения баз I–IV плоскостями $L_1 - L_4$ сведены в табл. 19. В последней строке этой таблицы указаны вероятности благоприятного исхода в случае попадания в соответствующее подмножество (A_i или B_i) (см. табл. 12, 14, 16 и 18). Жирным шрифтом дается вероятность благоприятного исхода, если пациент оказался в соответствующем подмножестве с благоприятным исходом (т. е. в подмножестве A_i) при применении данной терапии (i-й).

Для каждой группы (0000, 0001, 0010 и т. д.) в соответствующей ей строке указано количество пациентов из каждого подмножества (A_i и B_i) базы $i, i \in$ {I, II, III, IV},

Таблица 19. **Результаты разделения баз I–IV с помощью плоскостей $L_1 - L_4$**

Группа	A_1	B_1	A_2	B_2	A_3	B_3	A_4	B_4	Σ	p	T	E	**A**	Δ
0000	5	18	1	8	6	10	**6**	**5**	59	0,346	IV	20,42	18	2
0001	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
0010	0	4	0	0	**0**	**0**	0	1	5	0,609	III	3,04	0	3
0011	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
0100	0	0	**0**	**0**	0	0	0	0	0	0,500	II	0	0	0
0101	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
0110	0	0	0	0	**0**	**0**	0	0	0	0,609	III	0	0	0
0111	4	13	1	3	5	4	**2**	**2**	34	0,719	IV	24,44	12	12
1000	**8**	**7**	1	14	3	7	2	6	48	0,652	I	31,3	14	17
1001	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
1010	**5**	**2**	0	1	1	1	1	5	16	0,652	I	10,43	7	3
1011	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
1100	**0**	**0**	0	0	0	0	0	0	0	0,652	I	0	0	
1101	0	0	0	0	0	0	**0**	**0**	0	0,719	IV	0	0	
1110	**0**	**0**	0	0	0	0	0	0	0	0,652	I	0	0	0
1111	32	15	3	1	8	4	**21**	**7**	91	0,719	IV	65,41	64	1
Σ	**54**	**59**	**6**	**27**	**23**	**26**	**32**	**26**	**253**			155,04	115	40
p_i	**0,65**	0,21	**0,5**	0,08	**0,61**	0,35	**0,72**	0,35						

идентифицированных как пациенты этой группы. Комментарий к табл. 19 такой же, как и к табл. 10 (см. п. 3.5).

Из табл. 16 следует, что при применении рекомендуемых способов лечения к каждой из 16 групп (с учетом вероятностей благоприятного исхода) количество пациентов с благоприятным прогнозом составит 155 человек (вместо 115 в действительности), т. е. на 40 человек больше. Это лучше, чем результат, полученный в п. 3.5 (см. табл. 10: там количество пациентов с благоприятным прогнозом увеличилось на 26 человек), что связано с использованием более эффективного разделения множеств A_i и B_i.

Заключение. Представленные результаты являются только иллюстративными. Для реального применения предлагаемой методики необходимо, чтобы используемая база была более репрезентативной и обширнее. Бóльшая точность разделения множеств может привести к улучшению прогнозирования и увеличению числа пациентов с благоприятным прогнозом. Выше множества разделялись с помощью плоскостей в соответствующих пространствах. Разделение можно проводить и более тонкими методами (некоторые из них описаны в [8, 10–13]).

В [3] химио- и гормональная терапии рассматривались как один тип, при этом количество пациентов с благоприятным исходом составило 141 человек. Проведенное нами раздельное исследование химио- и гормональной терапий позволило выяснить, что при рекомендуемом настоящей методикой способе лечения количество пациентов с благоприятным исходом увеличивается до 155 (вместо 115 в действительности).

Описанная выше задача была поставлена В. М. Моисеенко, которому автор выражает свою признательность.

Summary

Demyanova V. V. On prognosing the efficiency of different types of medical treatment.

A methodology of prognosing the efficiency of different ways in the treatment of patients is described. The methodology is illustrated by the database of oncological patients. The following problem is studied: to find a criterion allowing for any patient to give a prognosis for the time of his/her survival in the case chemotherapy or/and hormonal therapy is applied to treat him/her and in the case no therapy is used.

Литература

1. *Demyanova V. V.* The principal expert method in data mining // Applied Comput. Math. 2005. Vol. 4, N 1. P. 70–74.

2. *Демьянова В. В.* Метод главного эксперта в задачах идентификации // Труды Междунар. конференции «Устойчивость и процессы управления» (С.-Петербург, 29.06.2005–01.07.2005) / Ред. Д. А. Овсянников, Л. А. Петросян. СПб.: Изд-во С.-Петерб. ун-та, 2005. Т. 2. С. 815–822.

3. *Демьянов В. Ф., Демьянова В. В., Кокорина А. В., Моисеенко В. М.* Прогнозирование эффективности химиотерапии при лечении онкологических заболеваний // Вестн. С.-Петерб. ун-та. Сер. 10: Прикладная математика, информатика, процессы управления. 2006. Вып. 4. С. 30–36.

4. *Wolberg W. H., Lee Y.-J., Mangasarian O. L.* WPBCC: Wisconsin Prognostic Breast Cancer Chemotherapy Database. Computer Science Dept., University of Wisconsin, Madison. ftp://ftp.cs.wisc.edu/math-prog/epo- dataset/machine-learn/cancer/WPBCC/, 1999.

5. *Lee Y.-J., Mangasarian O. L., Wolberg W. H.* Survival-time classification of breast cancer patients // Computational Optimization and Applications. 2003. Vol. 25. P. 151–166.

6. *Kokorina A. V.* Unsupervised and supervised data classification via nonsmooth and global optimization // TOP (Theory of Optimization). Madrid, Spain, 2003. Vol. 11, N 1. P. 86–89.

7. *Kokorina A. V.* Ranking the parameters in classification databases // Longevity, Aging and Degradation Models. Vol. 2 (Материалы Междунар. конференции LAD'2004). СПб.: Изд-во С.-Петерб. гос. политехн. ун-та. 2004. С. 191–193.

8. *Demyanov V. F.* Mathematical diagnostics via nonsmooth analysis // Optimization Methods and Software. 2005. Vol. 20, N 2–3. P. 191–212.

9. *Григорьева К. В.* Аппроксимация критериального функционала в задачах математической диагностики: Канд. дис. СПб.: С.-Петерб. ун-т, 2006. 191 с.

10. *Lee Y.-J., Mangasarian O. L.* SSVM: A Smooth Support Vector Machine for Classification // Computational Optimization and Applications. 2001. Vol. 20, N 1. P. 5–22.

11. *Advances* in kernel methods. Support vector learning / Eds. B. Schoelkopf, C. J. C. Burges, A. J. Smola. Cambridge, Mass.; London, England: The MIT Press. 1999. 392 p.

12. *Bennett K. P., Mangasarian O. L.* Robust linear programming discrimination of two linearly inseparable sets // Optimization Methods and Software. 1992. Vol. 1, N 1. P. 22–34.

13. *Bagirov A. M., Rubinov A. M., Soukhoroukova N. V., Yerwood J.* Unsupervised and supervised data classification via nonsmooth and global optimization // TOP (Theory of Optimization). Madrid, Spain. 2003. Vol. 11, N 1. P. 1–93.

Статья рекомендована к печати членом редколлегии проф. С. В. Чистяковым.

Статья принята к печати 24 мая 2007 г.