Министерство образования и науки Российской Федерации
Санкт-Петербургский государственный университет
Национальный проект "Образование"

МЕТОДИКА СОСТАВЛЕНИЯ ПРОГНОЗА ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ РАЗЛИЧНЫХ СПОСОБОВ ЛЕЧЕНИЯ

Выполнено в рамках проекта
"Инновационная образовательная среда в классическом университете"

Санкт-Петербург
2006

УДК 519.3+519.7

Методика составления прогноза эффективности применения различных способов лечения: Учебно-методическое пособие / В.Ф.Демьянов, В.М.Моисеенко, В.Н.Иголкин, В.Т.Приставко, К.В.Григорьева, В.В.Демьянова, А.В.Кокорина. - СПб.: СПБГУ, 2006. - 65с.

Аннотация. В настоящем пособии описывается методика составления прогноза эффективности применения различных способов лечения. Предлагаемая методика разработана в рамках проекта "Медицинская физика и информационные технологии". Ее новизна состоит в том, что эта методика базируется на использовании результатов негладкого дискриминантного анализа (который является дальнейшим развитием и обобщением классического линейного дискриминантного анализа Р. Фишера). Применение современного математического аппарата позволяет строить модели, более адекватные поставленной задаче (в данном случае задаче прогнозирования эффективности лечения). Кроме того, представленная методика ориентирована на использование компьютерных технологий. Несколько демонстрационных программ прилагается (на компакт-диске).

Содержание

Введение

Гл. 1. Задача прогнозирования и метод главного эксперта (В.В.Демьянова)

§1.1 Введение 9

§1.2 Задача идентификации 9

§1.3 Задача прогнозирования 11

§1.4. Исследование множеств методов главного эксперта 12

Литература 15

Гл. 2. Прогнозирование эффективности химиотерапии при лечении онкологических заболеваний (В.Ф.Демьянов, В.В.Демьянова, А.В.Кокорина, В.М.Моисеенко) 16

§2.1 Введение и постановка задачи 16

§2.2 Разделение баз CT-140 и WCT-113. 17

2.2.1 Разделение базы CT-140 18

2.2.2 Разделение базы WCT-113 19

§2.3 Перекрестное разделение баз CT-140 и WCT-113. 21

2.3.1 Исследование базы WCT-113 с помощью плоскости L1 21 2.3.2 Исследование базы CT-140 с помощью плоскости L2 23

§2.4. Заключение и рекомендации 25

Литература 26

Гл. 3. Статистический подход к составлению прогноза эффективности применения различных способов лечения (В.Н.Иголкин).... 28

§3.1 Постановка задачи 28

3.1.1 Байесовский подход 29

3.1.2 Небайесовские решения 29

§3.2 Классификация наблюдений в случае классов, описанных нормальными законами распределения. 30

3.2.1 Случай с равными ковариационными матрицами 30

2.3.2 Случай неравных ковариационных матриц 32

§3.3 Классификация наблюдений, когда классы представлены конечными выборками. 33

3.3.1 Произвольные законы распределения 33

3.3.2 Нормальные законы распределения с параметрами, оцененными по выборкам 33

3.3.3 Нестрогий алгоритм классификации 34

Литература 36

Гл. 4. Липидная оценка эффективности лечения (В.Т.Приставко)..... 37

§4.1 Введение 37

§4.2 Постановка задачи и ее решение 38

§4.3 Методы построения разделяющей гиперплоскости 41

4.3.1 Метод Р.Фишера 43

4.3.2 Метод наименьших квадратов 44

4.3.3 Метод А.Н.Ширяева 44

§4.4 Применение разделяющей гиперплоскости в некоторых медико-биологических задачах 45

§4.5 Липидная оценка эффективности лечения 46

§4.6 Заключение 48

Литература 49

Приложение 1 Ранжирование параметров в задачах обработки данных (А.В.Кокорина)..... 50

§П.1.1 Случай нормально распределенных параметров 50

§П.1.2 Случай дискретных значений 57

Приложение 2 Методика построения разделяющих гиперплоскостей (К.В.Григорьева) 61

Приложение 3 Демонстрационная программа: Ранжирование параметров с помощью t-критерия Стьюдента (на компакт-диске) (В.В.Демьянова)

Приложение 4 Демонстрационная программа: Ранжирование параметров (на компакт-диске) (А.В.Кокорина)

Приложение 5 Демонстрационная программа: Построение разделяющей гиперплоскости с помощью суррогатного функционала (на компакт-диске) (К.В.Григорьева)

Приложение 6 Демонстрационная программа: Построение разделяющей гиперплоскости с помощью модифицированного метода Фишера (на компакт-диске) (А.В.Кокорина)

Введение

Математические методы давно применяются в задачах медицинской диагностики и прогнозирования эффективности лечения. Эти задачи решаются методами математической диагностики. Существуют различные подходы к моделированию и исследованию моделей. Классическим подходом является дискриминантный анализ, разработанный в 30-е годы XX столетия Р.Фишером для решения задач медицинской диагностики и основанный на использовании теории вероятностей и статистики. Другой подход - оптимизационный - начал развиваться в 50-е годы. Вначале он использовал только теорию линейного программирования (линейный дискриминантный анализ). В настоящем пособии описываются элементы негладкого дискриминантного анализа, который применяется к решению задач диагностики и прогнозирования эффективности лечения.

Многие практически важные задачи, например, задачи распознавания образов, классификации, технической и медицинской диагностики, идентификации, обработки экспериментальных данных, спектрального анализа, могут быть описаны математическими моделями, в которых требуется "отделить"два или более множества. Часто указанные множества "неразделимы". Поэтому возникает задача идентификации как можно большего числа точек этих множеств. В задачах медицинской диагностики оптимизационный подход к решению этой задачи был предпринят в 30-е годы Р.Фишером (который разработал линейный дискриминантный анализ). В настоящее время эти задачи занимают ведущее место в теории обучающих машин, в задачах искусственного интеллекта.

Для решения указанных задач существует несколько подходов (чисто статистический подход, метод опорных векторов В.Вапника, метод машинного обучения, метод построения ядра, метод обработки данных Мангасаряна, кластерный анализ). Практическая важность таких задач подчеркивается тем фактом, что в последние годы многие математики, до этого работавшие в других областях, обратили свое внимание на проблемы диагностики. Этот интерес связан с тем, что, с одной стороны, современный уровень развития вычислительной техники позволяет реализовать многие методы и алгоритмы, которые ранее невозможно было использовать ввиду их громоздкости, с другой стороны, состояние теории моделирования позволяет обрабатывать огромные массивы экспериментальных данных в различных областях (в

медицине, биологии, химии) с целью выявления новых закономерностей (возникло даже направление, называемое искусственным интеллектом).

Упомянутые выше задачи и создают направление в теории оптимизации, которое можно назвать математической диагностикой. Задачи математической диагностики представляют хороший полигон для проверки эффективности математических моделей, численных методов решения возникающих оптимизационных задач, отработки и усовершенствования методик идентификации и самих классификаторов (идентификаторов) и критериев. При этом появляются новые математические задачи и возможность практического использования ранее казавшихся абстрактными математических результатов. Хорошие модели существенно упрощают и улучшают процесс проектирования.

Как уже отмечалось, многие из упомянутых задач сводятся к разделению двух или более часто неразделимых множеств. Для этого важно иметь сравнительно простой критерий, с помощью которого можно классифицировать точки. При выбранном идентификаторе (называемом также классификатором) (функционале, по значению которого судят о принадлежности данной точки тому или другому множеству) качество идентификации оценивается естественным критерием - количеством ошибочно идентифицированных точек. Поскольку этот критерий (если его удается записать в виде функции, а не вербально, как происходит при первоначальной содержательной постановке задачи) обычно описывается существенно разрывной функцией, то приходится этот естественный критерий качества заменять некоторым суррогатным функционалом, который может быть изучен методами математического программирования. В основе существующих методов лежат методы линейного и квадратичного программирования, которые существовали и были наиболее эффективны во время создания указанных теорий машинного обучения (60-е - 80-е годы XX века). В отличие от изложенных подходов, в предлагаемом пособии предлагается решать эти задачи, используя негладкие критерии, которые лучше аппроксимируют "естественные"критерии качества. Ожидается (и эти ожидания оправдываются на практике), что негладкий дискриминантный анализ позволит улучшить качество идентификации и распознавания.

Решение задач диагностики заболевания и затем прогнозирования эффективности лечения состоит из следующих этапов:

1. *Ранжирование параметров.* Имеющиеся базы обычно содержат точки в пространстве большой размерности. Поэтому вначале выделяются наиболее информативные параметры (проводится ранжирование параметров). В настоящем пособии описывается два способа ранжирования. Оба они обоснованы в предположении, что случайные величины (значения параметров) подчиняются нормальному закону распределения. Один метод использует t-критерий Стьюдента. Этот критерий дает достоверность различия между выборками из генеральной совокупности по исследуемому параметру (см. Приложение 3, где приведена демонстрационная программа). Второй метод ранжирует параметры по величине площади пересечения подграфиков плотностей распределения (см. Приложение 1, где этот метод описан, и Приложение 4, в котором приводится демонстрационная программа). Методика проверки гипотез описана в главе 3. Использование различных способов ранжирования расширяет возможности выбора параметров для дальнейшего исследования, в особенности учитывая, что вероятностные характеристики случайных величин (параметров) нам не известны.

2. *Построение оптимального идентификатора (решающего правила - РП).* В качестве идентификатора в данной работе выбирается линейный функционал (т.е. множества разделяются с помощью гиперплоскости). Выбирается эта гиперплоскость из условия минимума некоторого критериального функционала. В Приложении 6 эта плоскость строится с помощью модифицированного метода Фишера, а в Приложении 2 - исходя из условия минимума введенного суррогатного функционала (демонстрационная программа представлена в Приложении 5). Использование различных критериальных функций также позволяет провести более точное разделение множеств и на его основе - точность диагностики заболевания.

3. *Построение прогноза эффективности лечения.* Для построения прогноза на основе метода главного эксперта используется методика, описанная в главе 1. Эта методика демонстрируется в главе 2 на примере построения прогноза эффективности применения химиотерапии при лечении онкологических больных (была использована широко доступная база Висконсинского университета).

Отметим, что для удобства пользователя главы 1-4 и Приложения 1 и 2 выполнены в виде законченных самодостаточных частей и могут изучаться и использоваться независимо от других частей. Это же относится и к демонстрационным

программам (Приложения 3-6, они содержатся на компакт-диске, прилагаемом к настоящему Пособию).

Авторами Пособия являются: К.В.Григорьева (Приложения 2 и 5), В.В. Демьянова (главы 1 и 2 и Приложение 3), В.Ф. Демьянов (глава 2), В.Н. Иголкин (глава 3), А.В. Кокорина (глава 2 и Приложения 1, 4 и 6), В.М. Моисеенко (глава 2), В.Т. Приставко (глава 4). Общую редакцию Пособия осуществлял В.Ф.Демьянов.

В.В. Демьянова

Гл. 1. Задача прогнозирования и метод главного эксперта

§ 1.1 *Введение*

Рассматривается задача прогнозирования эффективности применения нескольких методик лечения некоторой болезни (например, различными медикаментами). Предполагается, что для каждой методики известны результаты ее применения, т. е. известен идентификатор, или решающее правило (РП), с помощью которого для любого пациент можно (с некоторой известной точностью) сказать, будет ли данная методик эффективна в отношении его или нет, т. е. в какую группу он попадает: в группу "успешных" пациентов (для которых лечение окажется успешным), или в группу "неуспешных".

В работах автора [1] и [2] был описан "метод главного эксперта"(МГЭ), в котором из нескольких реш ющих правил строилось новое РП, позволяющее более точно проводить идентифицикацию. В настоящей работе МГЭ распространяется на случай задачи прогнозирования эффективности нескольких методик (ниже подробно рассматривается случай двух методик). В результате строится несколько прогностических групп, для каждой из которых дается прогноз оценки эффективности той или иной методики.

Предлагаемый подход описывается на примере задачи прогнозирования эффективности лечения некоторого заболевания с помощью двух методик лечения.

Проведенная проверка на некоторых известных базах данных дает обнадеживающие результаты. Ниже эта методика применяется к решению задачи прогнозирования эффективности химиотерапии при лечении онкологических заболеваний.

Литература к гл. 1

1. Демьянова В.В. Метод главного эксперта в задачах идентификации. // Труды Международной конференции "Устойчивость и процессы управления"(С.-Петербург, 29.06.2005-01.07.2005). Редакторы Д.А.Овсянников, Л.А.Петросян. Изд-во СПбГУ. 2005. Том 2. С. 815-822.

2. Demyanova V.V. The Principal Expert Method in Data Mining. // Applied Comput. Math. 2005. V. 4, No. 1. P. 70-74.

3. Bagirov A.M., Rubinov A.M., Soukhoroukova N.V., Yerwood J. Unsupervised and Supervised Data Classification Via Nonsmooth and Global Optimization. // Top. 2003. V. 11, N 1. P. 1-93.

4. Bennett K.P., O.L. Mangasarian O.L. Robust Linear Programming Discrimination of two linearly inseparable sets. *Optimization Methods and Software*. 1992. V. 1, N 1. P. 22-34.

5. A.V.Kokorina. Ranking the parameters in the mathematical diagnostics problems. // Comments to the paper [3]. 2002. P. 86–89.

6. Yuh-Jye Lee, O.L.Mangasarian O.L. SSVM: A Smooth Support Vector Machine for Classification. *Computational Optimization and Applications*. 2001. V. 20, No. 1, 5-22.

7. Vapnik V. (2000) *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, N.Y. 2000. xix+314 p.

В.Ф. Демьянов, В.В. Демьянова, А.В. Кокорина, В.М. Моисеенко

Гл. 2. Прогнозирование эффективности химиотерапии при лечении онкологических заболеваний

§ 2.1 *Введение и постановка задачи*

В работе описывается методика прогнозирования эффективности применения химиотерапии при лечении онкологических заболеваний. Имеются две базы данных: о пациентах, подвергшихся хирургической операции и прошедших курс химиотерапии, и о пациентах, которым была сделана хирургическая операция, но они не прошли курс химиотерапии. Предполагается, что эти базы представляют собой репрезентативные выборки из множества пациентов, подвергшихся хирургической операции.

Для обеих баз данных известны результаты применения (или неприменения) химиотерапии. Для каждой из них (являющейся *обучающей выборкой*) строится (методами *математической диагностики*) критерий (решающее правило), позволяющий предсказать результат лечения. При этом получаются и вероятности данных предсказаний. Такой критерий затем применяется к другой базе (служащей *контрольной выборкой*). В результате все пациенты делятся на четыре прогностические группы. Для первой группы прогноз и применения, и неприменения химиотерапии неблагоприятный; для второй – прогноз благоприятный в случае применения химиотерапии и неблагоприятный – в случае ее неприменения; для третьей – прогноз неблагоприятный при применении химиотерапии и благоприятный – в случае ее неприменения; наконец, для четвертой – прогноз и применения, и неприменения химиотерапии благоприятный. Для каждой группы даются вероятности благоприятного исхода в случае и применения, и неприменения химиотерапии.

Предлагаемая методика демонстрируется на примере базы CHEMO-253 ("WPBCC: Wisconsin Prognostic Breast Cancer Chemotherapy Database"), хранящейся в репозитории Висконсинского университета и широко доступной.

В настоящей работе приводятся результаты исследования базы данных CHEMO-253 (см. [1]). Она содержит сведения о 253 пациентах, больных раком молочной железы, которым была сделана хирургическая операция. 140 из них были подвергнуты химиотерапии (будем их называть пациентами с х/т), а 114 – нет (пациенты без х/т). Каждый из пациентов описан точкой в 39-мерном пространстве,

представляющей информацию о данных анализов (30 параметров), а также сведения о продолжительности жизни после операции в месяцах (наблюдения велись 13 лет), размере опухоли, наличии и количестве метастаз.

В [2] рассматривалась задача идентификации тех пациентов, для которых применение химиотерапии может увеличить продолжительность жизни. В [2] использовался математический аппарат, основанный на методе гладких опорных векторов (SSVM – Smooth Support Vector Machine) (см. [3–5]). Другие методы разделения можно, например, найти в [6].

Будем считать, что операция (с химиотерапией или без нее) прошла успешно, если пациент жил после операции не менее 5 лет, и неудачно – если срок жизни после операции был меньше 5 лет.

Из пациентов с химиотерапией 61 человек жил не менее 5 лет (множество этих пациентов обозначим A_1), а 79 – менее 5 лет (их множество – B_1). Из пациентов без химиотерапии 54 человека жили не менее 5 лет (их множество обозначим A_2), а 59 – менее 5 лет (их множество – B_2).

В работе изучается следующая задача: найти критерий, с помощью которого для каждого пациента можно определить, следует ли ему рекомендовать химиотерапию или она ему противопоказана (т. е. дать прогноз о продолжительности жизни в случае, если химиотерапия будет назначена, и в случае, если химиотерапия назначена не будет).

§ 2.4 *Заключение и рекомендации*

Таким образом, *выполнение рекомендаций, полученных с помощью описанной методики, позволило бы перевести в группу с благоприятным прогнозом на* $7 + 20 = 27$ *чел. больше, чем оказалось в действительности.* Всего в базе CHEMO-253 из 253 пациентов с неблагоприятным исходом оказались $79 + 59 = 138$ чел., т. е. предлагаемая методика *позволила бы уменьшить количество пациентов с неблагоприятным исходом на 27 чел.* (111 вместо 138, или на 19.5%).

Литература к гл. 2

1. *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).

2. *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.

3. *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.

4. *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.

5. *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.

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

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

8. *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. политехн. ун-та. 2004. С. 191–193.

9. *Демьянова В. В.* Метод главного эксперта в задачах идентификации // Труды Междунар. конференции "Устойчивость и процессы управления"(С.-Петербург, 29 июня – 1 июля 2005). Ред. Д. А. Овсянников, Л. А. Петросян. СПб: Изд-во С.-Петерб. ун-та, 2005. Т. 2. С. 815–822.

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

В.Н. Иголкин

Гл. 3. Статистический подход к составлению прогноза эффективности применения различных способов лечения

§ 3.1 *Постановка задачи*

Имеются две группы больных, к которым применялось данное лечение. Для одной группы лечение оказалось эффективным, для другой - нет. Используя эту информацию, нужно принять решение об эффективности применения данного способа лечения к вновь появившемуся пациенту. Эту задачу можно поставить, как задачу статистической классификации (различения гипотез).

Каждому больному сопоставляется упорядоченный набор (вектор) X признаков (показатели крови, температура и.т.д.). В силу индивидуальных особенностей больных вектор признаков можно считать случайным. Будем считать, что больным , для которых лечение оказалось эффективным, соответствует одно распределение признаков $F_1(X)$, для тех, для которых данное лечение оказалось не эффективным, распределение $F_2(X)$. Оценки этих распределений можно получить из имеющейся статистики двух групп, (их называют обучающими последовательностями), считая их выборками из генеральных совокупностей. Имеется обширная литература, посвященная нахождению выборочных распределений [1-3]. Если $F_1(X)$ и $F_2(X)$ найдены (точнее найдены их оценки), рассматривается задача классификации: к какому распределению отнести предъявленный вектор X, то-есть соответствующему больному рекомендовать данное лечение или нет.

Решение задачи классификации состоит в разбиении выборочного пространства Ω на непересекающиеся подмножества Ω_1 и Ω_2 такие, что $\Omega_1 \cup \Omega_2 = \Omega$ и, если $X \in \Omega_1$, то больному рекомендуется лечение, если $X \in \Omega_2$, то не рекомендуется. При этом возможны следующие ошибки: (1)может оказаться, что мы рекомендуем лечение больному, которому не следует применять это лечение (ошибка первого рода); (2) мы не рекомендуем лечение больному, которому следовало рекомендовать это лечение (ошибка второго рода). Вероятности совершения этих ошибок соответственно равны

$$\alpha = \int_{X_1} f_2(X)dX, \qquad \beta = \int_{X_2} f_1(X)dX$$

Хотелось бы, чтобы решающее правило (разбиение $\Omega_1 \cup \Omega_2$) минимизировало бы α и β одновременно. Но обычно уменьшение одной величины ведёт к увеличению другой. Наша задача имеет два критерия и возможны различные подходы к решению этой задачи оптимизации.

3.1.1. Байесовский подход. Назначаются потери от неправильной классификации: $L_{1,2}$ при ошибке первого рода и $L_{2,1}$ при ошибке второго рода. На самом деле важно их соотношение, а не абсолютная величина. Пусть q_1 и q_2 априорные вероятности гипотез. О них, например, можно судить по количеству больных в группах

$$q_1 = \frac{n_1}{n_1 + n_2}, \qquad q_2 = \frac{n_2}{n_1 + n_2},$$

где n_1 и n_2 число больных в первой и второй группах.

Тогда величину $\rho = q_2 L_{1,2}\alpha + q_1 L_{2,1}\beta$ можно назвать средним риском (потерями) и решающее правило ищут из условия $\min \rho$. Оно называется байесовским и множество Ω_1 получается таким

$$\Omega_1 = (X \in \Omega | L_{2,1} q_1 f_1(X) > L_{1,2} q_2 f_2(X)) \tag{1}$$

Заметим, любое небайесовское решающее правило можно найти, как байесовское при подходящих значениях q_1 и q_2, если фиксированы $L_{1,2}$ и $L_{2,1}$.

3.1.2. Небайесовские решения. В подходе Неймана - Пирсона фиксируется одна из ошибок, например β и среди всех решающих правил, обеспечивающих данную величину β, ищется правило, дающее $\min \alpha$. Критерий Неймана - Пирсона приводит к такому решающему правилу

$$\Omega_1 = (X \in \Omega | \frac{f_1(X)}{f_2(X)} > k), \tag{2}$$

где величина k определяется из уравнения

$$L_{1,2} \int_{X_1} f_2(X) dX = \alpha. \tag{3}$$

При минимаксном критерии $\min_d max(L_{2,1}\beta,\, L_{1,2}\alpha)$ решающую функцию d находят из условия

$$L_{2,1} \int_{X_2} f_1(X) dX = L_{1,2} \int_{X_1} f_2(X) dX. \tag{4}$$

Литература к гл. 3

[1] Кендалл М, Стюарт А., Теория распределений. М. "Наука"1966г. 587 с.

[2] Надарая Э.А. О параметрических оценках плотности вероятностей и регрессии. журнал: Теория вероятносей и её применения. Вып.1 М. 1965. с. 199-203.

[3] Иголкин В.Н.,Ковригин А.Б., Старшинов А.И., Хохлов В.А. Статистическая классификация, основанная на выборочных распределениях. Изд. ЛГУ, Ленинград, 1978, 102 с.

[4] Robbins H., Monro S. A stochastic approximation method. J. Ann. Math. Stat. 1951. vol. 22. 1. p. 400-407.

[5] Цыпкин Я.З. Основы теории обучающих систем. М. "Наука"1970. 251 с.

В.Т. Приставко

Гл. 4. Липидная оценка эффективности лечения

§ 4.1 *Введение*

Задача прогноза эффективности способа лечения того или иного заболевания является актуальной проблемой современной медицины. Она неразрывно связана с математической теорией распознавания образов.

Сложность решения поставленной задачи определяется тем, что, как правило, каждому заболеванию того или иного вида сопутствует патология другого типа. При этом описание заболевания характеризуются достаточно большим числом параметров, что, естественно, приводит к тому, что даже группа опытных врачей допускает ошибки. Развиваемые методы математической диагностики позволят в будущем избежать возможных врачебных ошибок и являются необходимым инструментом на этапе диагностики заболевания и принятия решения о способе его лечения. Необходимость развития методов математической диагностики следует из перечня основных групп терапевтических заболеваний:

- болезни органов дыхания,
- болезни сердечно-сосудистой системы,
- ревматические и системные заболевания соединительной ткани,
- болезни органов пищеварения,
- болезни почек,
- болезни системы крови,
- болезни эндокринной системы.

Каждая группа заболеваний делится на болезни. Так, например, в медицине классифицируют следующие болезни сердечно-сосудистой системы [1]: аневризма аорты, аритмии сердца, атеросклероз, гипертензия артериальная пограничная, гипертензия легочная артериальная первичная, гипертензия артериальная симптоматическая, гипертоническая болезнь, гипертонический криз, гипотензия артериальная, гипотензия артериальная ортостатическая, дистония нейроциркуляторная, инфекционный эндокардит, ишемическая болезнь сердца, внезапная коронарная смерть, стенокардия, инфаркт миокарда, кардиомиопатия, миокардиодистрофии, миокардиодистрофия алкогольная, миокардиты неревматические, миокардит идиопатический

Абрамова - Фидлера, недостаточность кровообращения острая, недостаточность кровообращения хроническая, пороки сердца приобретенные, перикардиты, тромбоэмболический синдром, тромбоэмболия легочной артерии.

Последние достижения в медицине показывют тенденцию развития медико-биологических исследований в направлении комплексного лечения заболеваний. Необходимость такого подхода очевидна и следует из необходимости строго избирательного действия лекарств. В настоящее время данного эффекта не всегда удается достичь. Так, например, на фоне приема β-блокаторов, ингибиторов моноаминоксидазы, резерпина и других лекарств иногда возникает токсический бронхоспастический синдром, прием аспирина провоцирует язвенную болезнь и т.д. Суть этой проблемы состоит в индивидуальности каждого пациента как сложной живой системы клеток.

Внимательное изучение более 100 болезней [1] показывает, что большинство из них в достаточной мере описывается одними и теми же биохимическими показателями и лабораторными данными гематологических, цитологических и других исследований. Приведенный список болезней сердечно-сосудистой системы человека и заметно усиливающаяся тенденция “узкой” специализации врачей наглядно показывают насущную потребность здравоохранения в разработке компъютерной диагностики основного и сопутствующих заболеваний, а также и министерства образования для подготовки специалистов, достаточно полно владеющих методами компьютерной диагностики заболеваний и оценок эффективности их лечения.

Данная работа является продолжением исследований применения методов разделяющей гиперплоскости, изложенных в монографии [1].

§ 4.2 *Постановка задачи и ее решение*

Задача оценки эффективности лечения заболевания и ее решение чрезвычайно важна для населения Российской Федерации. Практически все населения в условиях перехода на платную основу здравоохранения сталкивается с проблемой, которая давно известна в Европе и США и состоящей в том, что врачу (желает он того или не желает, отдает себе в этом отчет или нет) на подсознательном уровне выгодно экономически "содержать" пациента в не здоровом состоянии здоровья, но и в не безнадежном, в так называемом "пограничном слое — не здоровый, но и не

больной". Это вполне очевидный факт, определяемый еще и тем, что зарплата высококвалифицированного врача в государственном медицинском учреждении равна зарплате уборщицы в гипермаркете. В противоположной медицине, когда участковый врач практически не зависит от "кошелька" пациента, ему выгодно "содержать" всех в состоянии абсолютного здоровья, а остальных направлять в санатории и к специалистам, тогда у него освобождается масса свободного времени, которое он может использовать в своих личных целях. Знание того, что пациент может проконтролировать эффективность лечения заболевания, возможно уменьшит остроту указанной проблемы нынешнего здравоохранения.

Исходя из принятой концепции [1] о том, что только врач имеет право на диагностику заболевания оценку его лечения, математическая модель прогноза эффективности способа лечения того или иного заболевания в недетерминированной среде строится следующим образом [1].

На **первом этапе** врач (или врачи) высокой квалификации (назначаемый приказом главврача лечебного учреждения) определяет тестовый набор пациентов (так называемая "обучающая"выборка из не менее, чем 60-ти пациентов), по истории болезни которых известна достоверная классификация данного k-го терапевтического заболевания, и набор его диагностических признаков в соответствии с программой обследования в виде базы данных (матрицы значений исходных переменных $\mathbf{X}$):

Номер пациента	X_1	X_2	...	X_g	Y_1	Y_2	...	Y_p
1	x_{11}	x_{12}	...	x_{1g}	y_{11}	y_{12}	...	y_{1p}
2	x_{21}	x_{22}	...	x_{2g}	y_{21}	y_{22}	...	y_{2p}
3	x_{31}	x_{32}	...	x_{3g}	y_{31}	y_{32}	...	y_{3p}
...	...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{ng}	y_{n1}	y_{n2}	...	y_{np}

База данных k-го терапевтического заболевания. Таблица **4.1.**

Где X_1, X_2, ..., X_g — биохимические показатели и лабораторные данные гематологических, цитологических и других исследований, которые образуют группу переменных факторов заболевания до начала лечения;

$Y_1, Y_2, ..., Y_p$ — результативные качественные показатели психо-эмоционального и физического состояния пациента после проведенного курса лечения.

Итак, общее число факторов $m = p + g$. При этом, естественно, что $p \leq g$.

На **втором этапе** посредством алгоритма распознавания заболевания тот же врач формирует набор (базу данных, матрицу) стандартных значений: $S = || \ s_{ij} \ ||_{n \times k}$, где i – номер заболевания, а j – номер стандартных значений вектора X_j. Он также определяет значения разделяющих[1] гиперплоскостей $A = \{[a_j]\}_{i=1}^{n}$ и делит "шкалу жизни"[2] данного i-го заболевания на отрезки (матрица L), характеризуемые величинами: $\mathbf{Lp}_0, \ \mathbf{Lp}_1, \ \mathbf{Lp}_2, \ \mathbf{Lp}_3, \ \mathbf{Lp}_4$.

Таким образом, данное i-ое заболевание полностью определяется множеством диагностических признаков, значениями "весовых" коэффициентов и параметрами шкалы жизни. Назовем это множество классом ("образом") i-го заболевания. Тогда $\mathbf{\Omega}_i = \{s_i\} \cup \{a_i\} \cup \{L_i\}$.

В результате выполнения этих двух этапов формируется база видов терапевтических заболеваний, характерная для данной местности и данного лечебного учреждения $\mathbf{\Omega} = \bigcup\limits_{i=1}^{n} \mathbf{\Omega}_i$.

Как следует из теории вероятностей система подмножества $F = \{A : \ A \subseteq \mathbf{\Omega}\}$ является σ – алгеброй, а пространство $(\mathbf{\Omega}, \ \mathbf{F})$ – измеримым. Задавая вероятностную[3] меру $\mathbf{P} = \mathbf{P}(A)$, получаем вероятностное пространство $(\mathbf{\Omega}, \ \mathbf{F}, \ \mathbf{P})$.

Зная достоверно "наихудшее" сочетание факторов риска, характеризующее гибель объекта, и минимально значимое, характеризующее его рождение[4], определим $Y_m = \inf \ \mathcal{Y}(X)$ и $Y_M = \sup \ \mathcal{Y}(X)$. Тогда линейным преобразованием

$$\mathbf{Lp}(X) = \frac{Y(x) - Y_m}{Y_M - Y_m}$$

определяется липидность каждого объекта X в фиксированный момент времени $t, \ t \in [t_0, \ T]$. По значению $\mathbf{Lp}(A)$ можно дать вероятностную оценку риска заболеваемости той или иной болезнью конкретного индивидуума A.

В работе [1] принята концепция, состоящая в том, что только врач имеет право на диагностику заболевания и методы его лечения. Исходя из данной концепции, каждый врач, на основе собственного опыта, формирует набор стандартных

[1]). См. п.° 4.3. Методы построения разделяюшей гиперплоскости.

[2]). $\mathbf{Lp}(X), \ \ 0 \leq \mathbf{Lp}(X) \leq 1$. См. [1].

[3]). В учебном пособии [3] приведены (таблица 1 с. 170) наиболее употребительные типы дискретных вероятностных мер.

[4]). Не обязательно присущие какому-либо индивидууму.

значений s_j, базу видов заболеваний и "делит" шкалу жизни каждого заболевания по своему усмотрению на следующие отрезки:

абсолютное здоровье – $(\mathbf{Lp}_0 \leq \mathbf{Lp}(A) < \mathbf{Lp}_1)$,

здоров – $(\mathbf{Lp}_1 \leq \mathbf{Lp}(A) < \mathbf{Lp}_2)$,

периодическое врачебное наблюдение – $(\mathbf{Lp}_2 \leq \mathbf{Lp}(A) < \mathbf{Lp}_3)$,

исследование в стационаре – $(\mathbf{Lp}_3 \leq \mathbf{Lp}(A) < \mathbf{Lp}_4)$,

срочная госпитализация – $(\mathbf{Lp}_4 \leq \mathbf{Lp}(A) \leq 1)$.

На **третьем этапе** врач определяет вероятностную меру $\mathbf{P} = \mathbf{P}(A)$ и оптимальное решающее правило (ОПР) из предлагаемого набора допустимых мер и ОПР. В соответствии с ОПР компъютерный анализ значений биохимических и лабораторных показателей факторов риска каждого вновь поступившего пациента определяет основной вид его заболевания и сопутствующие ему. При этом имеется возможность просмотра историй болезни наиболее близких по вероятности пациентов из базы видов терапевтических заболеваний. Известными методами математической статистики определяется возможная эффективность проведения того или иного способа лечения. Вполне очевидно, что окончательная оценка эффективности проведенного курса лечения определяется в каждом случае по Y_1, Y_2, ..., Y_p результативным качественным показателям психо-эмоционального и физического состояния пациента отдельно после прохождения курса лечения и реабилитационнго периода восстановления параметров жизнидеятельности организма.

§ 4.6 *Заключение*

Поставленная задача прогноза эффективности различных способов лечения является триединой задачей, состоящей из: синтеза базы (таблица 4.1) данного заболевания и способов его лечения, совокупности "разделяющих" гиперплоскостей и надежности статистических выводов.

Все три задачи достаточно широко изучены и поэтому допустимо считать их тривиальными. Автор данного раздела сознательно не указывает все нюансы решения данной задачи, так как имеется достаточное количество лжеученых и программистов, которые могут безответственно реализовать любой опубликованный в открытой печати алгоритм. А речь здесь идет о жизни и смерти живого человеческого организма. Так по заказу одной американской фирмы в 1994 году автором вместе

с врачами и программистами был разработан алгоритм оценки риска заболевания сердечно-сосудистой системы человека "Cardiovaskular"(название принадлежит американской фирме), который сейчас широко распространен в Интернете. Однако, в виду известных американских принципов ведения бизнеса, опирающихся на плагиат и воровство идей, код программы имел небольшие отклонения, приводящие к завышению количественной оценки риска заболевания. Данный недостаток ими не устранен до сих пор. Это завышение не имеет последствий к развитию паталогии, но заставляет пациента внимательно относится к своему холестерину, потреблению алкоголя и курению.

Трудности решения задачи заключаются в том, что практически все базы данных являются закрытыми, так как принадлежат конкретному лечебному учреждению и имеют сведения о пациентах. Известные в открытой печати базы данных **X** носят искусственный характер, чтобы показать преимущество одного из методов над другим. С точки зрения практикующих врачей, эти базы данных не выдерживают никакой критики.

Данная задача, несмотря на свою тривиальность, обладает сложностью своей многоплановости, зависимости от местности и от опыта практикующих врачей. Рассматривая перспективу решения данной задачи, видно, что *необходимо* создавать компьютерную диалоговую среду при непосредственном участии известных, практикующих врачей, посредством которой каждый врач будет получать необходимый инструментарий для решения задачи выбора метода лечения того или иного заболевания с контролем эффективности принимаего им и только им решения. Задача является актуальнейной, перспективной, но требующей достаточно большого времени и капиталовложений на ее техническую реализацию.

Литература к гл. 4

1. Приставко В.Т. Матричные модели управления. СПб: НИИ Химии СПбГУ, 2001.– 255 с.

2. Кендалл М., Стюарт А. Многомерный статистический анализ и временные ряды.– М.: Наука, 1976.– 731 с.

3. Ширяев А.Н. Вероятность.– М.: Наука, 1980.– 575 с.

А.В.Кокорина

Приложение 1. Ранжирование параметров в задачах обработки данных

В задачах идентификации, возникающих в математической диагностике, и в задачах обработки данных мы имеем дело с точками многомерного пространства, что серьёзно затрудняет процедуру идентификации. Но успешное отделение множеств не требует использования всех существующих координат (параметров или свойств), достаточно рассматривать только очень ограниченное число из данных координат, делая, таким образом, численные расчеты действительно выполнимыми. Однако остается открытым вопрос, как определить наиболее важные параметры. Ниже изложены некоторые простые приёмы для ранжирования параметров баз данных.

К.В.Григорьева

Приложение 2. Методика построения разделяющих плоскостей

Задачи математической диагностики (в частности, задачи диагностирования пациентов) можно свести к исследованию модели, в которой требуется разделить в пространстве два множества точек. Это можно сделать и различными способами, в том числе с помощью гиперплоскости. Для практического решения такой задачи идентификации требуется: - построить правило идентификации; - сформулировать для данной задачи естественный критерий (натуральный функционал, обычно это - количество ошибочно классифицируемых точек). Критерий часто не является непрерывным; - выбрать соответствующую модель (суррогатный функционал); - создать алгоритмы и программное обеспечение для получения численных результатов. Иными словами, требуется выбрать сравнительно простой критерий, по значению которого можно судить о принадлежности данной точки тому или другому множеству, и построить алгоритмы получения разделяющих гиперплоскостей.

Литература к Приложению 2

1. Григорьева К.В. Метод проектирования в одной задаче идентификации. Процессы управления и устойчивость. Труды 34 науч. конф. - СПбГУ, 2003, с.268-271.

2. Григорьева К.В. Идентификация множеств с помощью негладкой модели. Процессы управления и устойчивость. Труды 35 науч. конф. - СПбГУ, 2004, с.294-296.

3. Григорьева К.В. Суррогатные функционалы в математической диагностике. Актуальные проблемы строительства. 63-я науч. конф. -СПбГАСУ, 2006, с.90-92.