

УДК 51-76:616.89-02-07

**АЛГОРИТМ ОБРАБОТКИ МЕДИЦИНСКОЙ БАЗЫ ДАННЫХ НА ОСНОВЕ
РАВНОСИЛЬНЫХ ЧАСТНЫХ МЕТОДИК****О.А. Трояножко***Санкт-Петербургский государственный университет
199034, Санкт-Петербург, Университетская наб., 7-9*

Аннотация. Описывается анализ медицинской базы данных. Применяется алгоритм выявления «шумов» на основе равносильных частных методик. Численно показана эффективность алгоритма. Данная методика применялась при комбинированном использовании с традиционными статистическими методами. Выполнен анализ базы данных по предложенному алгоритму, а также представлены некоторые результаты классификации.

Ключевые слова: Линейный дискриминантный анализ, распознавание, классификация, анализ медицинских баз данных, линейная отделимость классов, шизофрения, пограничное состояние.

Одним из направлений развития математической диагностики является построение разделяющей гиперплоскости на основе предварительного отделения «шумов» - ошибочных измерений, снижающих качество распознавания.

Ниже предлагается методика выявления «шумов», основанная на сопоставлении экспертных оценок. Подобный подход применялся в работах Демьяновой В.В. [1].

В отличие от работы Демьяновой, где выделялись «искусственные» эксперты, в данной статье предлагается использовать равносильные методики.

Основной смысл данного подхода состоит в том, что используются те частные методики, качество разделения которых различается не более, чем на определенную δ (на некоторую фиксированную величину).

Строится модель распознавания n больных по m тестовым показателям с применением алгоритма выявления «шумов». Решается задача классификации пациентов на две подгруппы с разным диагнозом.

Постановка задачи.

Пусть имеется выборка X из n объектов x_i ($i=1, \dots, n$), измеренных по m показателям (признакам).

Рассмотрим множество показателей K , по которым проводились измерения как объединение наборов показателей, и назовем их «частыми методиками». По каждой из этих методик строятся разделяющие гиперплоскости.

Уравнение дискриминантной (разделяющей) гиперплоскости имеет вид линейной регрессии

$$\partial = \sum_{i=1}^n \alpha_i \chi_i = (\alpha, \chi) = \alpha^* \chi, \quad (1)$$

где α - вектор неизвестных «весовых» коэффициентов.

Эти коэффициенты рассчитываются по методу Р. Фишера [2], где критерий построения разделяющей гиперплоскости основан на максимизация отношения межгрупповой дисперсии к общей. Тогда

$$\alpha = \gamma^{-1}(\mu_1 - \mu_2), \quad (2)$$

где γ - матрица рассеяния, μ_1 - математическое ожидание значений первой группы пациентов, а μ_2 - второй группы.

Пусть B - множество булевых значений. $\Phi = \{f : P \rightarrow B\}$ - множество всех отображений из множества объектов в булево пространство. Есть оператор, сопоставляющий набору критериев дискриминантную функцию:

$$M : 2^K \rightarrow \Phi.$$

Рассмотрим функцию диагнозов: $d \in \Phi$ - функция, принимающая истинное значение, если у пациента подтверждается искомый диагноз.

Определение. $Z \subset X : \forall \omega \in 2^K \forall x \in X M(\omega)(x) \neq d(x)$ - множество Z будем называть «шум».

Требуется построить алгоритм выявления «шума».

Метод решения.

Ниже представлен алгоритм поиска ошибочно распознанных пациентов, позволяющий выявлять ошибочные измерения («шумы»).

Выбираем множество объектов, на которых проводятся измерения X , множество частных методик $D = \{d_1, \dots, d_n\}$ и d - подлинную функцию диагнозов.

Множество шумов будет построено за $|X|$ шагов.

Рассмотрим последовательность множеств S_i , $0 \leq i \leq |X|$, $S_0 = \emptyset$ и $S_i \subset X$, которая строится по следующему алгоритму. На каждом шаге мы выполняем последовательно следующие действия:

Действие 1. Выбираем новый объект $x_i \in X$, где i - номер шага.

Действие 2. Выбираем новую частную методику $d_j \in D$

Действие 3. Сравниваем $d_j(x_i)$ и $d(x_i)$. Если $d_j(x_i) = d(x_i)$ то полагаем, $S_i = S_{i-1}$ переходим к следующему шагу алгоритма, если нет, повторяем действие 2.

Действие 4. Полагаем $S_i = S_{i-1} \cup \{x_i\}$ и переходим к следующему шагу.

В результате этого процесса получится расширяющаяся последовательность множеств, последняя из которых и будет множеством шумов.

Пример. Данный алгоритм был испробован на базе, состоящей из 80 пациентов и 13 признаков. Данные в ПНД были собраны д.м.н. Коробовой Е.Л. с помощью различных психологических методик и тестов, что является причиной разной степени точности классификации больных на две подгруппы [3]. Полученные характеристики больных шизофренией крайне близки друг к другу и этим осложняют классификацию. Решается

задача разделения пациентов на две подгруппы – больных с параноидной формой и простой.

Рассматривались три частные методики $\{k_1, k_3\}$, $\{k_3, k_6\}$ и $\{k_5, k_7\}$, где $k_i \in K$. Качество разделения на основе каждой из этих частных методик выбрано одинаковым и в процентном отношении равным 76,3%.

- Одна частная методика опирается на два признака (из тринадцати), ранжированных ранее.
- Существующая база пациентов была разделена дискриминантной гиперплоскостью тремя частными методиками поочередно (рис.1).

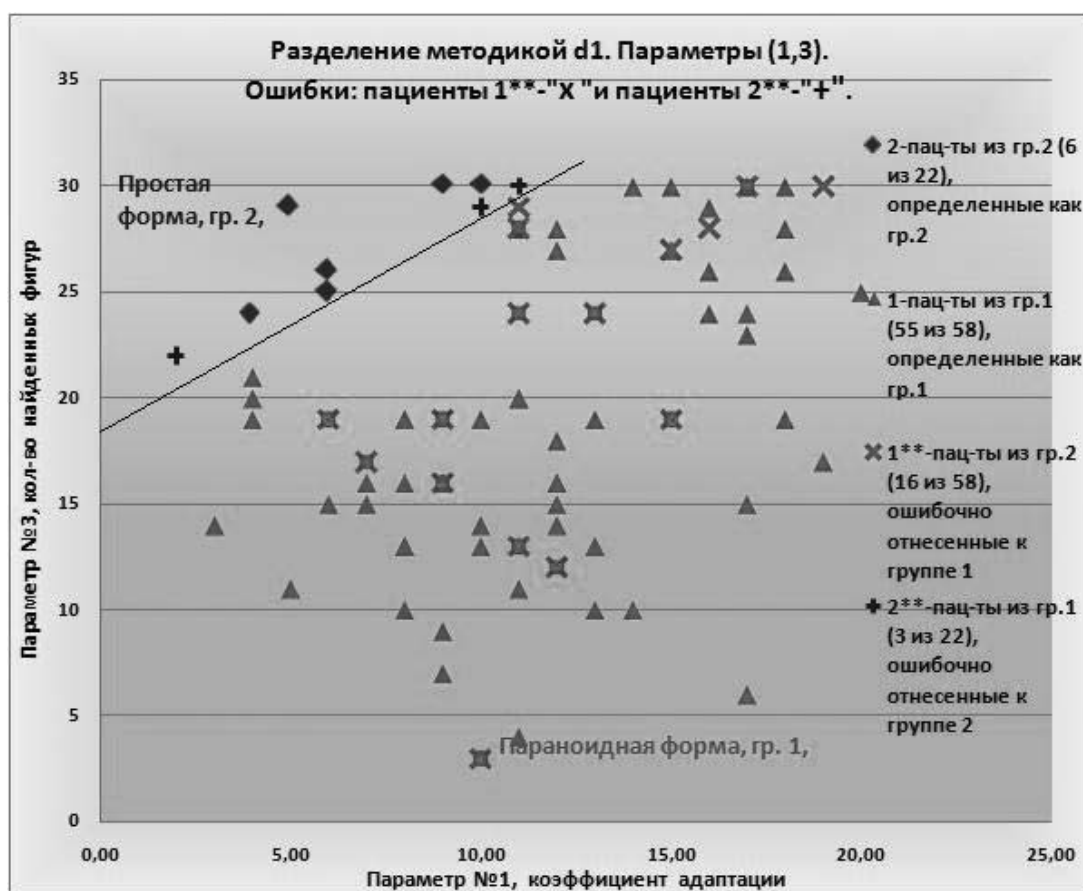


Рис.1. Разделение на основе частной методики №1.

- Было выбрано три равносильные частные методики. Количество ошибок, допущенных при использовании каждой из них, равно 19.
- Далее было составлено множество пересечения ошибочных заключений по всем трем частным методикам.
- Количество неверно распознанных пациентов у каждой частной методики составляет 19. При этом 12 пациентов из этих ошибочно распознанных - общие для трех частных методик, они и составляют множество «шумов». Это множество «шумов» было исключено из общей базы (табл.1).

Таблица 1

Пересечение ошибок по трем частным методикам.

№ пац.	Результат разделения методикой №1 (1,3)	Результат разделения методикой №2 (3,6)	Результат разделения методикой №3 (5,7)
6	1**	1**	1**
13	1**	1**	1**
15	1**	1**	1**
17	1**	1**	1**
34	1**	1**	1**
45	1**	1**	1**
47	1**	1**	1**
48	1**	1**	1**
50	1**	1**	1**
53	1**	1**	1**
66	1**	1**	1**
77	1**	1**	1**

- На следующем шаге были найдены новые дискриминатнтые функции по сокращенной базе. Далее было выполнено разделение тремя гиперплоскостями. Процент разделимости существенно возрос (см. таблицу 2)

Таблица 2

Скачок в качестве распознавания по полной базе и по сокращенной базе после применения методики.

Набор признаков	Полная база	Сокращенная база
(1,3,5,6,7)	73,8%	88,7%
(1,3); (3,6); (5,7)	76,3%	91,2%; 89,7%; 86,8%

Полученные результаты.

При использовании описанного алгоритма выявления «шумов» точность разделения возрастает на 14,9%, т.е. до 85%, что является достаточно хорошим результатом.

Заключение. Данная методика показала эффективное применение алгоритма выявления «шумов». В дальнейшем планируется разработать программный продукт для обработки больших выборок.

Использование алгоритма выявления «шумов» эффективно дополняет традиционные статистические методы, а их совместное использование позволит существенно повысить качество диагностики.

ЛИТЕРАТУРА

1. Демьянова, В. В. Одномерная идентификация методом разделения [Текст]. // Вестн. С.-Петерб. ун-та; Сер. 10. Прикладная математика, информатика, процессы управления. 2006. Вып. 3. С. 28-31.
2. Кендалл, М., Стюарт, А. Многомерный статистический анализ и временные ряды [Текст]. М.: Наука, 1976. 731 с.
3. Холодная, М.А. Когнитивные стили. О природе индивидуального ума [Текст]. 2-е изд. СПб.: Питер, 2004. 384 с.

**MEDICAL DATA BASE PROCESSING ALGORITHM BASED
ON EQUIVALENT SPECIAL METHODICS**

O. Troyanozhko

*Saint-Petersburg state university
7-9, Universitetskaya nab., St. Petersburg, 119034*

Abstract. Analysis of medical database is described. Noise detection algorithm based on equivalent special methodics is used. Effectiveness of algorithms is numerically proven. This methodic is used in combination with traditional statistic methods. Analysis of database using given algorithm is implemented and some classification results are presented.

Key words: Linear discriminant analysis, recognition, classification, medical database analysis, linearly separated sets, schizophrenia, borderline.