УДК 004

РАЗРАБОТКА И РЕАЛИЗАЦИЯ АЛГОРИТМА ОЦЕНКИ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ ПРИ ДИАГНОСТИКЕ ЗАБОЛЕВАНИЙ

Дашутина Е.В., Блощицкий В.П.
Донецкий национальный технический университет
кафедра автоматизированных систем управления
E-mail: lisaveta_91@mail.ru

Аннотация

Дашутина Е.В., Блощицкий В.П. Разработка и реализация алгоритма оценки информативности признаков при диагностике заболеваний. *Рассмотрен статистический критерий оценки информативности признаков. Определены преимущества оценки меры расхождения между распределениями, соответствующими двум выборкам, по критерию Кульбака. Составлен алгоритм нахождения величины информативности. Разработана база данных для реализации алгоритма.*

Общая постановка проблемы

При изучении объектов, характеризуемых большим числом факторов, часто бывает важно определить, какие из этих факторов в большей степени влияют на интересующие нас свойства объектов. В частности, определение информативности факторов – это один из важных этапов анализа изучаемого объекта. В отличие от других критериев статистической значимости различий, мера Кульбака позволяет оценить не достоверность различий между распределениями, а степень этих различий. Метод анализа признаков путем оценки информативности критерием Кульбака получил широкое применение в медицине, при рассмотрении отдельных факторов, влияющих на постановку диагноза.

Исследования

Под дифференциальной информативностью признака понимают степень различий его распределений при дифференцируемых состояниях А и В. Эти состояния, хранимые в БД в виде статистических данных, являются входными данными.

Первым шагом в разработке алгоритма, производящего вычисление информативности признака, является разбиение интервала статистических данных на диапазоны. Для этого выбираем такие равные между собой диапазоны, чтобы их количество составляло 10. В алгоритме, для получения длины диапазона, производится выбор максимального и минимального значения из всего ряда с последующим делением этого отрезка на 10.

Следующий шаг алгоритма – подсчет числа наблюдений из групп А и В, попавших в данный диапазон. Это частоты данного признака. Затем находим частости путем представления полученных частот в процентах, принимая за 100% сумму частостей А во всех диапазонах и такую же сумму частостей В.

На следующем шаге алгоритма вычисляют сглаженные (средневзвешенные) частости для большинства диапазонов по формуле:

$$\begin{aligned} \bar{y}_3 &= (y_1 + 2y_2 + 4y_3 + 2y_4 + y_5):10 \\ \bar{y}_4 &= (y_2 + 2y_3 + 4y_4 + 2y_5 + y_6):10 \end{aligned} \qquad (1)$$

и т.д.,

где y_1– член выборки, ближайший к любому ее краю;

y_2 – второй от края член выборки;

y_3 – третий от края и т.д.;

$\bar{y}_3$ и $\bar{y}_4$ – «средневзвешенный» или «сглаженный» член выборки.

Для крайних диапазонов № 1, 2 и 9, 10 – по формулам:

$$\bar{y}_3 = (0 + 2y_1 + 4y_2 + 2y_3 + y_4) : 10 \quad (2)$$

$$\bar{y}_3 = (0 + 0 + 4y_1 + 2y_2 + y_3) : 10 \quad (3)$$

Наглядно данный расчет представлен в виде блок-схемы на рис. 1.

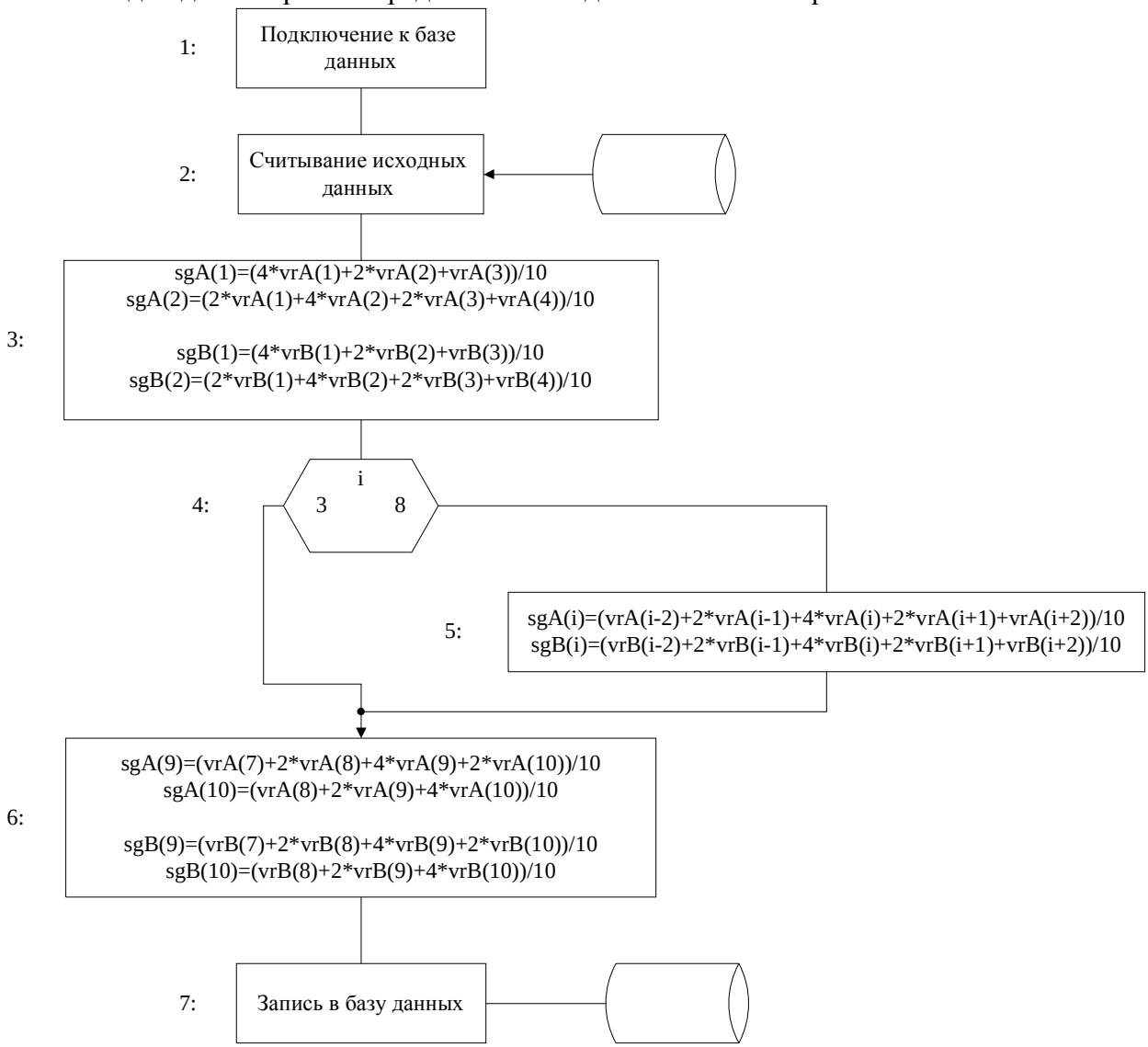

Рис. 1. Блок-схема алгоритма расчета сглаженных частостей для каждого диапазона.

Массивы sgA и sgB предназначены для хранения результатов, а именно сглаженных частостей состояний A и B соответственно. Массивы vrA и vrB получены на предыдущем этапе расчета, они хранят относительные частости попадания в диапазон. В блоках 3 и 6 производится расчет согласно формулам (2) и (3). В цикле в блоке 5 расчет ведется по формуле (1).

Следующим этапом в разработке алгоритма является вычисление отношений сглаженных частостей A и B в каждом диапазоне, полученных на предыдущем этапе.

Теперь переходим к вычислению диагностических коэффициентов по формуле:

$$\overline{ДК}(x_j^i) = 10 \lg \frac{\overline{P}(x_j^i / A)}{\overline{P}(x_j^i / B)}, \quad (4)$$

где $\overline{P}(x_j^i / A)$ и $\overline{P}(x_j^i / B)$ – средневзвешенные частости признака в каждом диапазоне. Все полученные величины диагностических коэффициентов округляют с точностью до единицы.

Последний этап – вычисление информативности каждого диапазона. Согласно формуле Кульбака величина информативности *I* диапазона *i* признака *j* равна:

$$I(x_j^i) = ДК(x_j^i)\frac{1}{2}[P(x_j^i / A) - P(x_j^i / B)] \quad (5)$$

Информативность всего признака x_j равна сумме информативностей его диапазонов:

$$I(x_j) = \sum_i I(x_j^i) \quad (6)$$

Все расчеты, кроме расчета сглаженных частостей, производятся последовательно, учитывая данные текущего диапазона.

Для хранения исходных данных и полученных результатов была разработана база данных, физическая модель которой представлена на рис. 2.

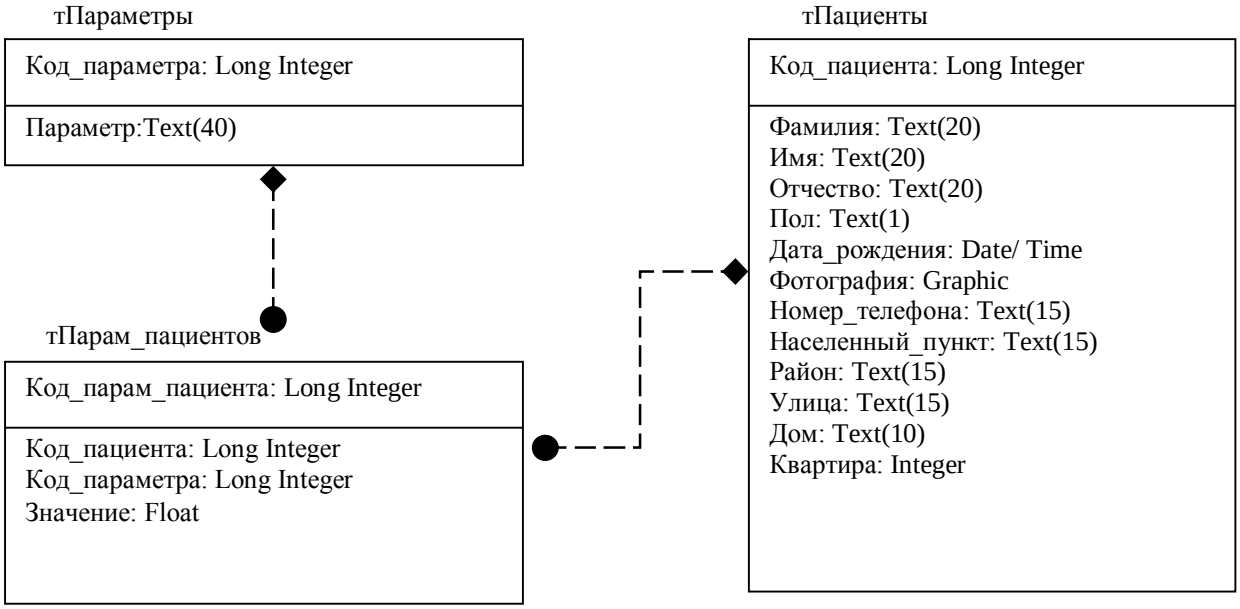

Рис. 2. Физическая модель данных.

Выводы

На основе вероятностных подходов с использованием методики расчета информативности по Кульбаку сформирован алгоритм создания и расчета словаря диагностических признаков.

Алгоритм прошел проверку на данных кардиологического отделения Областной клинической больницы профзаболеваний. Таким образом, реализация рассмотренного алгоритма обеспечивает сокращение трудозатрат при высокой точности расчетов.

Литература

1. Гублер Е.В., Генкин А.А.. Применение критериев непараметрической статистики для оценки различий двух групп наблюдений в медико-биологических исследованиях. М.: Медицина. 1969. 29 с.

2. Генкин А.А. Новая информационная технология анализа медицинских данных; Программный комплекс ОМИС / А. А. Генкин. — СПб. : Политехника, 1999. — 191 с.