УДК 004.93'1:61

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ РАСПОЗНАВАНИЯ СОСТОЯНИЯ ЭЛЕМЕНТОВ МЕДИКО-БИОЛОГИЧЕСКОЙ СИСТЕМЫ

***Н.С. Бакуменко**, к.т.н., доцент*
НАУ «ХАИ» им. Н.Е. Жуковского
***Р.Р. Абкеримова**, студентка*
НАУ «ХАИ» им. Н.Е. Жуковского
***Е.М. Угрюмова**, к.т.н., научный сотрудник*
НАУ «ХАИ» им. Н.Е. Жуковского

В статье рассматривается алгоритм к-средних, с применением новой метрики, основанной на расстоянии Махалонобиса.

Bakumenko N.S, Abkerimova R.R., Ugryumova. an algorithm to-medium, with new metrics based on the Mahalanobis distance.

Ключовые слова: АЛГОРИТМ К-СРЕДНИХ, МЕТРИКА, КЛАСТЕР.

Keywords: ALGORITHM K-MEANS, METRIC, CLUSTER.

Часто существует проблема составления корректной обучающей выборки и существует ряд медицинских задач, для которых нет смысла затрачивать большое количество времени на сбор подобных данных. Для таких задач и предлагается использовать методы кластеризации данных. Они разнообразны по принципам разделения объектов на классы, но в основном эти методы носят «четкую» природу, то есть не позволяют кластерам пересекаться в пространстве признаков [1].

Исходной информацией является выборка наблюдений, сформированная и N n-мерных векторов признаков $X = \{x(1), x(2),..., x(N)\}$, $x(k) \in X$, k=1,2,…,N. Целевая функция, подлежащая минимизации имеет вид [2-3]:

$$E(w_j(k), c_j) = \sum_{k=1}^{N} \sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j) \to min,$$

при ограничениях:

$$\sum_{j=1}^{m} w_j(k) = 1, k = 1, \dots, n, 0 < \sum_{k=1}^{N} w_j(k) < N, j = 1, \dots, m.$$

Здесь $w_j(k) \in [0,1]$ – уровень принадлежности вектора $x(k)$ к j-му кластеру, c_j - центроид j-го кластера, $d^2(x(k), c_j)$– расстояние между $x(k)$ и c_j в принятой метрике, β – неотрицательный параметр, именуемый «фаззификатором» (в случае использования $d^2(x(k), c_j)$в качестве евклидового расстояния, принимается равным 2).

Начальный набор центров прототипов $\mathbf{c}_j^0$, согласно формуле

$$c_j = \frac{\sum_{k=1}^{N} w_j^{\beta}(k) x(k)}{\sum_{k=1}^{N} w_j^{\beta}(k)}.$$

На основании рассчитанных центров-прототипов $\mathbf{c}_j^0$ далее вычисляется матрица W_1 согласно формуле

$$w_j = \frac{(d^2(x(k), c_j))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (d^2(x(k), c_l))^{\frac{1}{1-\beta}}}.$$

В результате работы алгоритма получим матрицу нечеткого разбиения, в которой пациенты будут разделены на кластеры (диагнозы). Расстояния между $x(k)$ и c_j:

$$d(x(k), c_j) = \sqrt{(x(k) - c_j)^T A_j (x(k) - c_j)},$$

где A_j - матрица, которая может быть определена как обратная нечеткая ковариационная матрица каждого кластера.

Если в качестве матрицы Aj возьмем единичную матрицу, то в результате получим эвклидово расстояние $d(x(k), c_j) = \sqrt{(x(k) - c_j)^T (x(k) - c_j)}$, и форма кластеров будет округлая (гипершары).

Распознавание состояния. Задача распознавания состояния сводится к задаче классификации. Далее для решения задачи классификации состояния объекта использовалась вероятностная нейронная сеть.

Активность элемента слоя образцов определялась зависимостью, соответствующей плотности распределения

вероятностей согласно -закону Стьюдента (что уместно и для ограниченных выборок):

$\rho_{lm} = \rho(\bar{F}_m^*|R_l) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t_{ml}^2}{n}\right)^{\frac{n+1}{2}}$, где Г– гамма-функция Эйлера с *n* степенями свободы

$(n = K_l + K_m - 2;\ \ K_l,\ K_m$ - число прецедентов в соответствующих классах, *l,m=1…M*);

t_{lm} – статистика Стьюдента:

$$t_{lm} = \sqrt{\frac{MD_{lm}^2}{\frac{1}{K_l}+\frac{1}{K_m}}},$$

MD_{lm}^2 - расстояние Махаланобиса от неизвестного прецедента (полагая, что он относится к *l*-тому классу) до *m*-ого образца – $MD_{lm}^2 = \frac{1}{A}(\bar{F}^* - < \bar{F}_m >)^T \Sigma_{pooled}^{-1}(\bar{F}^* - < \bar{F}_m >)$, где $\bar{F}^*$ – значения проекций вектора ГК наблюдаемых симптомов неизвестного прецедента;

$< \bar{F}_m >$– средние значения проекций вектора ГК наблюдаемых симптомов элемента слоя образцов;

Σ_{pooled}– объединенная ковариационная матрица для рассматриваемых сценариев (классов);

В докладе предлагается новый способ нахождения расстояния между кластрами в алгоритме к-средних.

Литература

1. Лбов, Г. С. Метод адаптивного поиска логической решающей функции [Текст] / В. М. Неделько, С. В. Неделько // Сиб. журн.индустр. матем. – 12:3 2009. – С. 66–74

2. Чурюмова, И. Г. Система медицинской диагностики на основе нечеткой логики [Текст] / И. Г. Чурюмова // Восточно-Европейский журнал передовых технологий. – 2006. – 5/2 (23). – С. 89-91.

3. Чурюмова, И. Г. Система донозологической диагностики сердечно-сосудистых заболеваний [Текст] / И. Г. Чурюмова // Восточно-Европейский журнал передовых технологий. – 2007. – № 5/4 (29). – С. 31-33.