Нелинейные преобразования признаков и поиск закономерностей на данных больных хроническим лимфолейкозом

Н.А. Игнатьев[1], Е.Н. Згуральская[2], М.В. Марковцева[3]

[1]Национальный Университет Узбекистана, ул. Университетская 4, Ташкент, Узбекистан, 100174

[2]Ульяновский Государственный Технический Университет, проспект Созидателей 13а, Ульяновск, Россия, 432072

[3]Ульяновский Государственный Университет, Льва Толстого 42, Ульяновск, Россия, 432000

Аннотация. Рассматривается поиск логических закономерностей по описаниям объектов в спрямляющем пространстве. При синтезе латентных признаков этого пространства используются правила иерархической агломеративной группировки. Выбор пары признаков для объединения в группу реализуется по максимуму критерия разбиения значений признаков на непересекающиеся интервалы. Приводится аналитический вид арифметических выражений для расчета латентных признаков, используемых для обнаружения скрытых закономерностей на данных больных хроническим лимфолейкозом (ХЛЛ).

1. Введение

Выбор пространства для описания объектов через нелинейные преобразования признаков является средством для поиска скрытых закономерностей в данных. При использовании таких преобразований меняется структура отношений объектов в новом (латентном) признаковом пространстве. Количественные оценки структуры могут выражаться через значения компактности объектов классов и выборки в целом.

Было предложено несколько способов оценки компактности [1, 2]. В [2] через меры компактности объектов классов и выборки в целом показана связь размерности признакового пространства и обобщающей способности алгоритмов распознавания по правилу ближайший сосед. Оценка компактности классов по количественному признаку в [3] вычислялась как экстремум критерия разбиения значений признака на непересекающиеся интервалы.

При анализе данных числовая ось рассматривается как универсальная шкала с отношениями. Универсальная шкала применялась для исследования отношений между объектами классов по результатам нелинейного отображения их описаний по определяемым наборам признаков на

числовую ось [4]. Поскольку состав каждого набора изначально неизвестен, для его поиска было предложено использовать критерий разбиения значений признаков на непересекающиеся интервалы.

Последовательное формирование наборов признаков в [4] для синтеза по ним латентных признаков производилось по правилам иерархической агломеративной группировки. Количество латентных (групп) признаков определялось по результатам группировки. Хорошим свойством для анализа латентных признаков является их упорядоченность по отношению информативности. Свойство упорядоченности даёт определённые преимущества для поиска скрытых закономерностей в данных.

Проблемы построения информационных моделей в медицине чаще всего рассматривается с позиций больших или плохо структурированных систем. В таких системах необходимо учитывать связь различных факторов и их влияние на процессы в организме. В работе исследуется поиск закономерностей для больных ХЛЛ [5]. В качестве факторов, влияющих на продолжительность фактической выживаемости больных, рассматриваются латентные признаки, синтезируемые по правилам иерархической агломеративной группировки. Приводится результаты поиска скрытых закономерностей на данных больных ХЛЛ по латентным признакам и аналитический вид арифметических выражений (формулы) для вычисления их значений.

2. Постановка задачи и метод решения

Пусть задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, содержащее представителей двух непересекающихся классов K_1, K_2. Описание объектов производится с помощью набора из n количественных признаков $X(n)=(x_1,\ldots,x_n)$. Считается, что на E_0 задан оператор A для преобразования описаний объектов из $X(n)$ в $Y(k)$, $k<n$.

Требуется определить:

– количество латентных признаков в $Y(k)$;

– аналитический вид (формулы) арифметических выражений для вычисления латентного признака $y_i \in Y(k)$, $i=1,\ldots,k$.

Аналитический вид арифметических выражений для вычисления латентных признаков формируется на базе алгоритма из [2]. Множество номеров признаков в описании объектов E_0 будем идентифицировать как $I=\{1, \ldots, n\}$. Для вычисления значений латентных признаков используются правила иерархической агломеративной группировки. Латентные признаки, полученные на p–м шаге группировки, обозначаются как x_j^p , $j\in I$, $p\geq 0$. При $p=0$, $|I|=n$. Упорядоченное множество значений признака x_j^p объектов из E_0 разделим на два интервала $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$, каждый из которых рассматривается как градация номинального признака.

Пусть u_i^1, u_i^2 – количество значений признака x_j^p , $j\in I$ класса K_i, $i=1,2$ соответственно в интервалах $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$, $|K_i|>1$, v – порядковый номер элемента упорядоченной по возрастанию последовательности

$$r_{j_1}, \ldots, r_{j_v}, \ldots, r_{j_m} \tag{1}$$

значений x_j^p у объектов из E_0, определяющий границы интервалов как $c_1^{jp} = r_{j_1}, c_2^{jp} = r_{j_v}, c_3^{jp} = r_{j_m}$.

Критерий

$$\left(\frac{\sum_{i=1}^{2} u_i^1(u_i^1-1)+u_i^2(u_i^2-1)}{\sum_{i=1}^{2}|K_i|(|K_i|-1)}\right)\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2} u_i^d(|K_{3-i}|-u_{3-i}^d)}{2|K_1||K_2|}\right) \rightarrow \max_{c_1^{jp}<c_2^{jp}<c_3^{jp}} \tag{2}$$

позволяет вычислять оптимальное значение границы c_2^{jp} для интервалов $[c_1^{jp}, c_2^{jp}]$ и $(c_2^{jp}, c_3^{jp}]$.

Экстремум критерия (2) используется в качестве веса w_j^p $(0 \le w_j^p \le 1)$ признака x_j^p. При $w_j^p = 1$ значения признака x_j^p у объектов из классов K_1 и K_2 не пересекаются между собой.

Значение комбинации b_{rij}^p по паре признаков (x_i^p, x_j^p), $0 \le p < n$, $i, j \in I$, $i \ne j$ объекта $S_r = \{a_{ru}^p\}_{u \in I}$, $S_r \in E_0$ вычисляется как $b_{rij}^p = \eta_{ij}(t_i w_i^p (a_{ri}^p - c_2^{ip})/(c_3^{ip} - c_1^{ip}) + t_j w_j^p (a_{rj}^p - c_2^{jp})/(c_3^{jp} - c_1^{jp})) + (1 - -\eta_{ij}) t_{ij} w_{ij}^p \left(\lambda_{rij}^p - c_2^{ijp}\right)/(c_3^{ijp} - c_1^{ijp}), i, j \in I, t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0;1]$ (3)

где w_i^p, w_j^p, w_{ij}^p – веса признаков, определяемые по (2) соответственно по множеству значений x_i^p, x_j^p и их произведения $\lambda_{rij}^p = \left(a_{ri}^p - c_2^{ip}\right)/\left(c_3^{ip} - c_1^{ip}\right) \times \left(a_{rj}^p - c_2^{jp}\right)/\left(c_3^{jp} - c_1^{jp}\right)$ на E_0, значения $t_{ij}, t_i, t_j \in \{-1,1\}$, $\eta_{ij} \in [0;1]$ выбираются по экстремуму функционала

$$\varphi(p,i,j) = \frac{\min\limits_{S_r \in K_1} b_{rij}^p - \max\limits_{S_r \in K_2} b_{rij}^p}{\max\limits_{S_r \in E_0} b_{rij}^p - \min\limits_{S_r \in E_0} b_{rij}^p} = \max\limits_{t_{ij}, t_i, t_j \in \{-1,1\}, \eta_{ij} \in [0,1]} \quad (4)$$

Экстремум функционала (4) интерпретируется как отступ между объектами классов K_1 и K_2 по множеству значений b_{rij}^p по паре признаков (x_i^p, x_j^p), $0 \le p < n$, $i, j \in I$, $i \ne j$.

В (3) применяется нормирование признаков по границам интервалов, вычисляемых по (2). Из–за нормирования признаков значения b_{rij}^p являются инвариантными к масштабам их измерений.

Обозначим через $\{z_{ij}^p\}_{i,j \in I}$, $p \ge 0$ − квадратную матрицу размера $(n-p) \times (n-p)$, значение элемента z_{ij}^p которой при $p=0$ определяется как

$$z_{ij}^p = \begin{cases} w_i^p, i = j, \\ \text{значению (2) по } \left\{b_{rij}^p\right\}_{r=1}^m, i \ne j, \end{cases} \quad (5)$$

через Γ_η, $\eta > 0$ – подмножество номеров признаков из $X(n)$.

Пошаговая реализация алгоритма иерархической агломеративной группировки будет такой:

1 шаг: $p=0$, $\lambda c = 0$, $\eta = 1$. Выполнять $\Gamma_\eta = \{\eta\}$, $Margin_\eta = -2$, $\eta = \eta + 1$ пока $\eta \le n$;

2 шаг: Вычислить значения элементов матрицы $\{z_{ij}^p\}_{i,j \in I}$ по (5);

3 шаг: Выделить $\Phi = \{z_{uv}^p \mid z_{uv}^p \ge \max(w_u^p, w_v^p)$ and $u \ne v, u, v \in I\}$. Если $\Phi = \varnothing$, то идти 9;

4 шаг: Вычислить $\lambda n = \max\limits_{z_{uv}^p \in \Phi} z_{uv}^p$. Выделить $\Delta = \{(s,t), s,t \in I \mid z_{st}^p = \lambda n$ and $s < t\}$. Определить пару $\{i,j\}$, $i<j$ как

$$\{i,j\} = \begin{cases} \Delta, |\Delta| = 1, \\ \{s,t\}, (s,t) \in \Delta \; and \; \varphi(p,s,t) > \max\limits_{(u,v) \in \Delta \setminus (s,t)} \varphi(p,u,v); \end{cases}$$

5 шаг: Если $\lambda n > \lambda c$ или $\lambda n = \lambda c$ и $Margin_i < \varphi(p, i, j)$, то $\Gamma_i = \Gamma_i \cup \Gamma_j$, $\Gamma_j = \varnothing$, $Margin_i = \varphi(p, i, j)$, идти 7;

6 шаг: Вывод номеров признаков из Γ_i, $\Gamma_i = \varnothing$, $I = I \setminus \{i\}$, идти 3;

7 шаг: $p = p+1$, $I = I \setminus \max(i,j)$, $k = \min(i,j)$, $\lambda c = \lambda n$. Заменить значения признаков в описании объекта $S_r = \{a_{ru}^{p-1}\}_{u \in I}$, $r = 1, \ldots, m$ на

$$a_{ru}^p = \begin{cases} a_{ru}^{p-1}, u \in I \setminus \{k\}, \\ b_{rij}^p, u = k; \end{cases}$$

8 шаг: Для каждой пары (*u, v*), *u, v*∈*I* определить значение

$$z_{uv}^{p} = \begin{cases} z_{uv}^{p-1}, u \in I\{k\}, v \in I, \\ \text{значению (9)на } \{a_{rv}^{p}\}_{r=1}^{m}, u = k, v \in I. \end{cases}$$

Если *n*-*p*> 1, то идти 3;

9 шаг: Конец.

При реализации описанного выше алгоритма значения параметров использовались для формирования аналитического вида арифметических выражений.

Например, в (3) значение η ∈{0,1}. Тогда в аналитическом представлении арифметических выражений при вычислении латентного признака по (3) записывается линейная или нелинейная часть.

Если при описании допустимых объектов используются два или более номинальных признаков, то на основе комбинаций их градаций можно получить латентные количественные признаки [6]. Значения таких латентных признаков вычисляются как обобщённые оценки объектов. Цель использования обобщённых оценок является отказ от бальных оценок объектов, полученных на основе субъективных критериев экспертов.

3.Вычислительный эксперимент

В качестве материала для исследования использовались данные 123 пациента с ХЛЛ А-С стадии по Binet [7, 8] в возрасте от 47 до 82 лет с известными значениями общей выживаемости (ОВ), полученные в гематологическом отделении Ульяновской областной клинической больницы. На момент постановки диагноза регистрировался возраст пациента, рассчитывался индекс коморбидности Charlson, измерялись стандартные биохимические показатели: аланинаминотрансфераза (АЛТ), аспартатаминотрансфераза (АСТ), общий билирубин, непрямой билирубин, глюкоза, креатинин, мочевина, мочевая кислота, лактатдегидрогеназа (ЛДГ), показатель скорости клубочковой фильтрации (СКФ) по MDRD [9]. При прохождении курса лечения регистрировалось количество сеансов химиотерапии и фактический показатель выживаемости (ФПВ) в месяцах. В базу данных не включались больные с ВИЧ-инфекцией и с отличными от ХЛЛ онкологическими состояниями.

По гендерному принципу были сформированы две выборки данных. Выборка данных больных мужского пола состояла из 60 объектов (возраст 64,6 ± 9,0 лет), женского пола из 63 объектов (возраст 67,0 ± 8,4 лет). Объекты каждой из выборок были разделены на два непересекающихся класса K_1 (фактическая выживаемость меньше прогнозируемой ОВ) и K_2 (фактическая выживаемость больше или равна прогнозируемой ОВ).

Наиболее сильные закономерности были получены на данных больных мужского пола. Классы K_1 и K_2 был представлены соответственно 36 и 24 объектами. По результатам вычислительных экспериментов на двух наборах (идентифицируемых как первый и второй) исходных признаков получены нелинейные комбинации, полностью разделяющие объекты двух классов. Во втором наборе отсутствовал признак индекс коморбидности. Полная разделимость классов подтверждается значением критерия (2), равного 1. Приведём последовательность формирования первого латентного признака по каждому из двух наборов исходных признаков.

Последовательность формирования латентного признака из первого набора:

x0=0.142857*(Индекс коморбидности – 4.0);
x1=0.017544*(СКФ по MDRD – 76.0);
x2=1.946341*(x0*x1 + 0.010025);
y1=1.257254*((0.120364*x0 – 0.244247*x1 – 0.533942*x2) – 0.014251).

Последовательность формирования латентного признака из второго набора:

x0=0.024390*(Возраст – 63.0);
x1=0.017544*(СКФ по MDRD – 76.0);
x2=2.379837*(x0*x1 + 0.010270);
y1=2.631118*((–0.098504*x0 + 0.244247*x1 – 0.179067*x2) + 0.033228);
x0=1.069903*y1;
x1=0.008333*(Креатинин – 84.0);
y2=1.153783*((–0.934664*x0 – 0.271799*x1) + 0.046027).

Логическая закономерность в форме полуплоскостей, полученная по второму набору, запишется как φ[y2< – 0.003]=β, β∈{true, false}. При φ[y2< – 0.003]=true пациент проживёт меньше расчётного срока ОВ. Значение порога для полуплоскости определялось по результату (2) разбиения описаний объектов выборки по признаку y2 на два интервала [c_1; c_2](c_2; c_3] как (c_2 + b)/2, где b – ближайшее к c_2 значение из (1) и $b > c_2$.

Существует функциональная зависимость СКФ по MDRD от значений возраста и креатинина в зависимости от гендерной принадлежности [9]. Определение балов для вычисления индекса коморбидности требует от пользователя дополнительных усилий по сбору данных от больных. По практическим соображениям использование второго варианта (без индекса коморбидности) вычисления латентного признака для прогноза выглядит предпочтительнее. Пользователю для прогноза ФПВ больного будет достаточно задать значения измеримых показателей возраст и креатинин.

4.Заключение

Описан поиск скрытых закономерностей по данным больных ХЛЛ. Приводится последовательность вывода аналитического представления арифметических выражений для расчёта значений латентных признаков. Прогноз отклонения фактических сроков выживаемости больных мужского пола в сторону уменьшения или увеличения от сроков ОВ определяется по логической закономерности в форме полуплоскостей. Найденные закономерности могут быть рекомендованы для использования в профильных лечебных заведениях.

5. Литература

[1] Загоруйко, Н.Г. Обучение распознаванию образов без переобучения / Н.Г. Загоруйко, О.А. Кутненко, А.О. Зырянов, Д.А. Леванов // Машинное обучение и анализ данных. – 2014. – Т. 1, № 7. – С. 891-901.

[2] Ignatyev, N.A. Structure Choice for Relations between Objects in Metric Classification Algorithms / N.A. Ignatyev // Pattern Recognition and Image Analysis. – 2018. – Vol. 28. – P. 590-597.

[3] Zguralskaya, E.N. Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness / E.N. Zguralskaya // CEUR Workshop Proceedings. – 2019. –Vol. 2416. – P. 283-289.

[4] Saidov, D.Y. Data visualization and its proof by compactness criterion of objects of classes / D.Y. Saidov // International Journal of Intelligent Systems and Applications. – 2017. – Vol. 9(8). – P. 51-58.

[5] Никитина, А.К. Эффективность лечения и выживаемость больных хроническим лимфолейкозом в зависимости от почечной функции / А.К. Никитина, Н.О. Сараева // Забайкальский медицинский вестник. – 2014. – № 4. – С. 122-127.

[6] Игнатьев, Н.А. Вычисление обобщённых показателей и интеллектуальный анализ данных / Н.А. Игнатьев // Автоматика и телемеханика. – 2011. – №5. – С. 183-190.

[7] Binet, J.L. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis / J.L. Binet, A. Auquier, G. Dighierol // Cancer. – 1981. – Vol. 48. – P. 198-206.
[8] Rai, K.R. Clinical staging of chronic lymphocytic leukemia / K.R. Rai, A. Sawitsky, E.P. Cronkite // Blood. – 1975. – Vol. 46. – P. 219-234.
[9] Medqueen [Электронный ресурс]. – Режим доступа: https://medqueen.com/medicina/diagnostika/diagnostika-statya/1966-skorost-klubochkovoy-filtracii-skf.html (25.06.2018).

Nonlinear transformation of signs and the search for patterns in the data of patients with chronic lymphocytic leukemia

N.A. Ignatyev[1], E.N. Zguralskaya[2], M.V. Markovtseva[3]

[1]National University of Uzbekistan Institute, Universitetskaya 4, Tashkent, Uzbekistan, 100174
[2]Ulyanovsk State Technical University, Sozidateley 13A, Ulyanovsk, Russia, 432072
[3]Ulyanovsk State University, Lev Tolstoy 42, Ulyanovsk, Russia, 432000

Abstract. The search for logical patterns by descriptions of objects in a rectifying space is considered. In the synthesis of latent features of this space, the rules of hierarchical agglomerative grouping are used. The choice of a pair of characteristics for combining into a group is realized according to the maximum criterion for dividing the values of the characteristics into non-intersecting intervals. The analytical form of arithmetic expressions for calculating latent signs used to detect hidden patterns in the data of patients with chronic lymphocytic leukemia (CLL) is given.