Классификация над произведением частичных порядков

Дюкова Елена Всеволодовна
Масляков Глеб Олегович
Прокофьев Пётр Александрович

ВЦ имени Дородницына ФИЦ ИУ РАН
МГУ имени М. В. Ломоносова
ИМАШ имени Благонравова

19-я Всероссийская конференция с международным участием
«Математические методы распознавания образов»

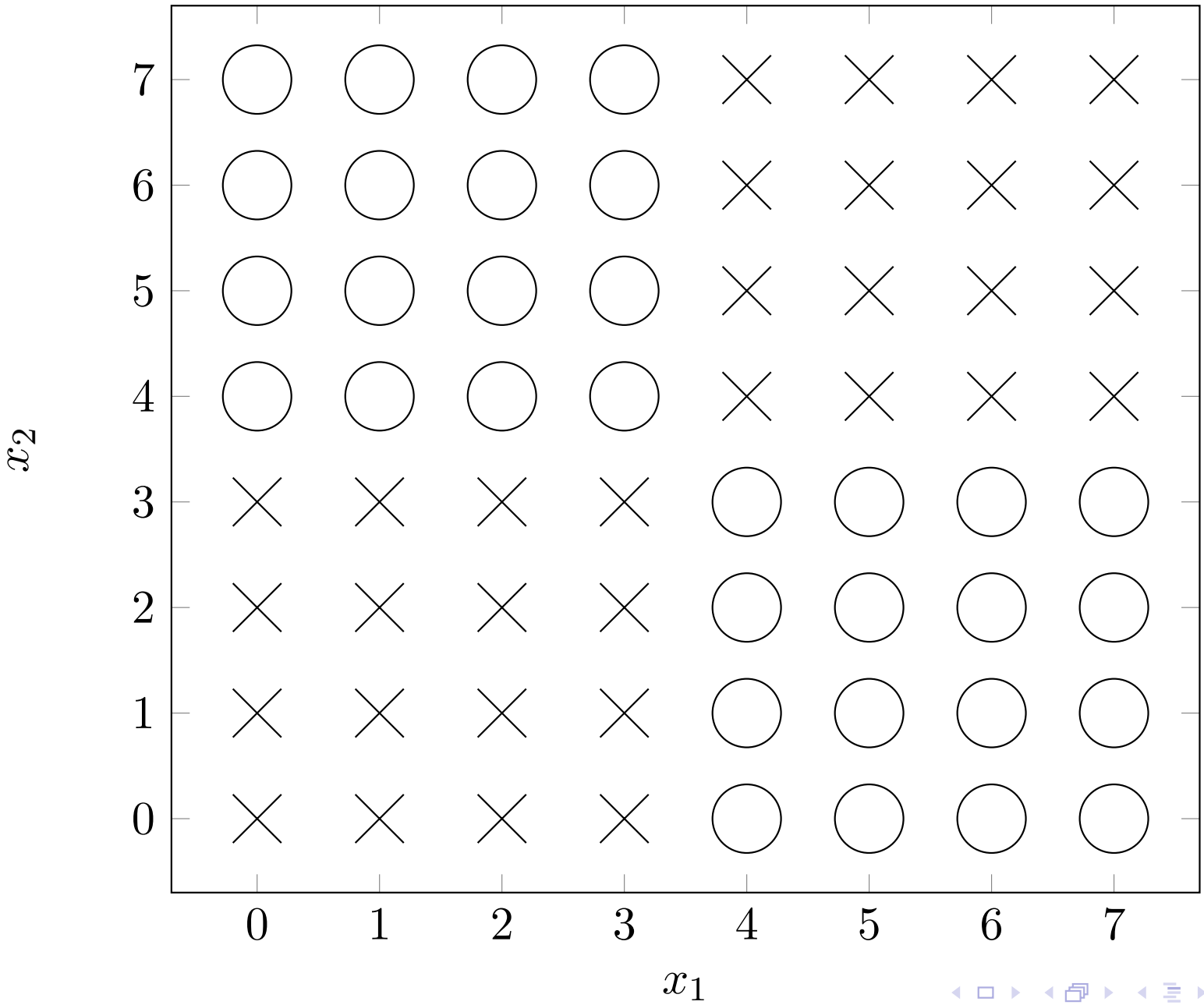
Классическая постановка задачи классификации целочисленных данных

- $M = K_1 \cup \cdots \cup K_l,\ K_i \cap K_j = \varnothing,\ i \neq j$
- $\{x_1, \ldots, x_n\}$ — целочисленные признаки
- N_j — множество допустимых значений признака x_j
- $S_1, \ldots, S_m \in M$ — прецеденты. $S_i \in K_j$
- $S_i = (a_{i1}, \ldots, a_{in})$, a_{ij} — значение целочисленного признака x_j для объекта S_i
- $R(K)$ — множество прецедентов из K
- $R(\overline{K})$ — множество прецедентов не из K
- Пусть $S \in M$
- Требуется определить какому классу принадлежит объект S

Логические процедуры распознавания

- $H = \{x_{j_1}, \ldots, x_{j_r}\}$ — набор различных признаков.
  $\sigma = (\sigma_1, \ldots, \sigma_r)$, $\sigma_i \in N_{j_i}$, $i = \overline{1, r}$
- (σ, H) — элементарный классификатор (эл.кл.)
- $B(S, \sigma, H) = \begin{cases} 1, & \text{если } a_{ji} = \sigma_i, \ \forall i = \overline{1, r}; \\ 0, & \text{иначе.} \end{cases}$
- (σ, H) — корректный для класса $K \in \{K_1, \ldots, K_l\}$, если
  $\forall S' \in R(K)$, $\forall S'' \in R(\overline{K})$, не выполнено
  $B(S', \sigma, H) = B(S'', \sigma, H) = 1$
- корректный эл.кл. (σ, H) — представительный для класса K, если
  $\exists S \in R(K)$: $B(S, \sigma, H) = 1$
- $C^A(K)$ — множество корректных, представительных для K эл.кл
- $P_{(\sigma,H)}$ — число прецедентов из K, содержащих (σ, H)
- $\Gamma(S, K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H) \in C^A(K)} P_{(\sigma,H)} B(S, \sigma, H)$

Модельный пример



Обобщение классического подхода: логическая классификация частично упорядоченных данных

- Пусть $M = N_1 \times \cdots \times N_n$, N_i — конечное частично упорядоченное множество значений признака x_i с наибольшим элементом k_i
- $\tilde{B}(S, \sigma, H) = \begin{cases} 1, & \text{если } a_{ji} \preceq \sigma_i, \ \forall i = \overline{1, r}; \\ 0, & \text{иначе.} \end{cases}$
- Данные ранее определения корректного и представительного элементарного классификатора можно переформулировать путём замены функции B на $\tilde{B}$
- $S(\sigma, H) = (\gamma_1, \ldots, \gamma_n)$, $\gamma_i = \sigma_t, \ i = j_t$, и $\gamma_i = k_i$
- Представительный для класса K элементарный классификатор (σ, H) называется тупиковым, если $\forall (\sigma', H') : \ S(\sigma, H) \prec S(\sigma', H')$ не является представительным для класса K

Поиск тупиковых представительных элементарных классификаторов

- Пусть $R \subseteq M = N_1 \times \cdots \times N_n$
- $R^+ = \{x \in M | \exists a \in R : a \prec x\} \cup R$ — идеал R
- $I(R)$ — множество максимальных независимых от R элементов

Теорема

Элементарный классификатор (σ, H) является тупиковым представительным для класса K тогда и только тогда, когда $S(\sigma, H) \in I(R(\overline{K})) \cap R(K)^+$.

Теорема о существовании представительных элементарных классификаторов

- Для любого прецедента S из произвольного класса K существует представительный элементарный классификатор, порожденный S, если и только если $R(K)$ независим от $R(\overline{K})$
- Пусть $\tilde{P} = P$, и $x \preceq y$ в $\tilde{P} \Leftrightarrow y \preceq x$ в P
- $\tilde{M} = \tilde{N}_1 \times \cdots \times \tilde{N}_n$
- $\phi : M \to M \times \tilde{M}$. $\phi((a_1, \ldots, a_n)) = (a_1, \ldots, a_n, a_1, \ldots, a_n)$

Теорема

Если классы множества M не пересекаются, то любой прецедент из класса $\phi(K)$ порождает тупиковый представительный для класса $\phi(K)$ элементарный классификатор.

Эксперименты

Таблица: Сравнение качества классификации на реальных данных при различном упорядочивании признаков.

Название датасета	Антицепи	Цепи	Смешанные	Дублированные
Car	73%	70%	84%	81%
Ph	43%	10%	53%	61%
Heart	76%	74%	81%	91%
Dermatology	95%	82%	95%	95%

Эксперименты

Таблица: Сравнение качества классификации

Название датасета	Random Forest	Random Sets Of EL.CL.
Car	88%	91%
Ph	78%	76%
Heart	85%	87%
Dermatology	98%	98%

Задача дуализации над произведением частичных порядков

- Пусть $P = P_1 \times \cdots \times P_n$, где $P_1, \ldots, P_n$ — конечные частично упорядоченные множества. Считается, что элемент $y = (y_1, \ldots, y_n) \in P$ следует за элементом $x = (x_1, \ldots, x_n) \in P$, если y_i следует за x_i, $i = \overline{1, n}$
- Запись $x \prec y$ означает, что y следует за x и y не совпадает с x
- $R \subset P$, $R^+ = \{x \in P | \exists a \in R : a \prec x\} \cup R$
- x — независит от R, если $x \in P \setminus R^+$
- $I(R) = \{x \in P \setminus R^+ | \nexists a \in P \setminus R^+ : x \prec a\}$ — двойственное к R множество
- Требуется для заданного R построить двойственное множество $I(R)$

Матричная формулировка задачи дуализации

- $P = P_1 \times \cdots \times P_n$, P_i — частично упорядоченное множество с наибольшим элементом. L_R — матрица, строки которой — наборы из $R \subseteq P$
- $Q_1(x, P) = \{y \in P : x \prec y,\ \forall a \in P : x \prec a \Rightarrow a \not\prec y\}$
- $Q_2(x, y, P) = \{a \in P : a \not\preceq x,\ a \preceq y\}$
- Пусть $\sigma = (\sigma_1, \ldots, \sigma_r),\ \sigma_i \in P_i,\ i = \overline{1, r}$. Упорядоченным тупиковым σ-покрытием матрицы L_R называется набор $H = \{j_1, \ldots, j_r\}$ из r различных столбцов этой матрицы такой, что подматрица L_R^H, образованная столбцами H матрицы L_R, не содержит строк предшествующих σ и $\forall t \in \{1, \ldots, r\}$, $\forall y \in Q_1(\sigma_t, P_{j_t})$ L_R^H содержит каждую из строк $(\beta_1, \ldots, \beta_r)$ такую, что $\beta_t \in Q_2(\sigma_t, y, P_t)$ и $\beta_j \leq \sigma_j$ при $j \in \{1, \ldots, r\} \setminus \{t\}$

Примеры упорядоченных тупиковых покрытий

- упорядоченное $(0,0,0,0)$-покрытие

$$\begin{pmatrix} \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{pmatrix}$$

- упорядоченное $(0,1,2)$-покрытие

$$\begin{pmatrix} \mathbf{1} & 1 & 2 \\ 0 & \mathbf{2} & 2 \\ 0 & 1 & \mathbf{3} \end{pmatrix}$$

Заключение

- На основе обобщения классических понятий предложена схема синтеза корректных логических процедур классификации по прецедентам, ориентированная на задание отношений частичных порядков на множествах значений признаков
- Показано, что в общем случае при построении процедур классификации возникает необходимость рассматривать одну из центральных труднорешаемых дискретных задач, а именно, задачу дуализации над произведением частичных порядков
- Дана матричная формулировка задачи дуализации над произведением частичных порядков
- Эксперименты подтверждают перспективность рассматриваемого подхода