Budowa modeli klasyfikacyjnych w oparciu o funkcję odległości w przestrzeni zdarzeń i uogólnienie pojęć probabilistycznych na przestrzenie metryczne

C. Dendek prof nzw. dr hab. J. Mańdziuk

Politechnika Warszawska,
Wydział Matematyki i Nauk Informacyjnych

Outline

1. Cel prac
2. Pojęcia podstawowe
3. Klasyfikacja k–NN

Zagadnienia

Cel pracy

- teoretyczne wyprowadzenie miary odległości w przestrzeniach probabilistycznych
- wykorzystanie teorii w praktycznych zastosowaniach klasyfikacyjnych (*k-NN*)
- znalezienie uogólnień możliwych do zastosowania dla innych klasyfikatorów

Norma trójkątna

Norma trójkątna $T(x, y)$:

- uogólnienie koniunkcji w logice wielowartościowej
- (monotoniczność) $\forall_{a,b,c,d \in [0,1]} a \leq b \wedge c \leq d \Rightarrow T(a, c) \leq T(b, d)$
- (przemienność) $\forall_{x,y \in [0,1]} T(x, y) = T(y, x)$
- (łączność) $\forall_{x,y,z \in [0,1]} T(x, T(y, z)) = T(T(x, y), z)$
- (neutralność względem “prawdy” – liczby 1) $\forall_{x \in [0,1]} T(x, 1) = x$

Rozmyta relacja równoważności

Rozmyta relacja równoważności $R(x, y)$

- $R(1, 1) = R(0, 0) = 1$
- $R(1, 0) = R(0, 1) = 0$

Rozmyta równoważność metryczna

Rozmyta równoważność metryczna R_M

1. $\forall_{x,y \in [0,1]} R_M(x,y) = R_M(y,x)$
2. $\forall_{x,y \in [0,1]} R_M(x,y) = 1 \Leftrightarrow x = y$
3. $R_M(1,0) = 0$
4. $\forall_{x,y,z \in [0,1]} R_M(x,y) \geq T(R_M(x,z), R_M(z,y))$

Rozmyta równoważność metryczna

Rozmyta równoważność metryczna R_M

Dla danej normy trójkątnej można zbudować równoważność metryczną (przy pomocy residuum):

$$\overrightarrow{T}(a,b) = \sup_{z \in [0,1]} \{z | T(a,z) \leq b\}.$$

$$R_M(a,b) = \overleftrightarrow{T}(a,b) = \min(\overrightarrow{T}(a,b), \overrightarrow{T}(b,a)).$$

Odległość rozmytą można wprowadzić jako

$$d_M(x,y) = 1 - R_M(x,y)$$

Wielowymiarowe rozmyte równoważności metryczne

Wielowymiarowe rozmyte równoważności metryczne

Jednowymiarowe równoważności metryczne można łączyć przy wykorzystaniu norm trójkątnych (odległość wielowymiarowa):

$$R_C = T(R_1, \ldots, R_n)$$

Kopuły statystyczne

Kopuły statystyczne

1. (zerowalność)
   $\forall_{\mathbf{u}=(u_1,\ldots,u_n)\in[0,1]^n} \min(u_1,\ldots,u_n) = 0 \Rightarrow C(\mathbf{u}) = 0$
2. (minimalność) $\forall_{\mathbf{u}=(1,\ldots,1,u_k,1,\ldots,1)\in[0,1]^n} C(\mathbf{u}) = u_k$
3. (n–monotoniczność) wybierając dowolne punkty $x_1,\ldots,x_n, y_1,\ldots,y_n$ tak, aby $\forall_{i\in\{1,\ldots,n\}} 0 \leq x_i < y_i \leq 1$ i oznaczając
   - $Q = \{x_1, y_1\} \times \cdots \times \{x_n, y_n\}$,
   - $\mathrm{sgn}_Q : Q \to \{0,1\}, \mathbf{p} = (z_1,\ldots,z_n) \mapsto (-1)^{card\{k\in\{1,\ldots,n\} z_k = x_k\}}$

   spełniony jest warunek

$$\sum_{\mathbf{p}\in Q} \mathrm{sgn}_Q(\mathbf{p}) C(\mathbf{p}) \geq 0$$

Kopuły statystyczne

Kopuły statystyczne

- bliskie pojęciu norm trójkątnych (bez przemienności i monotoniczności)
- rozważamy jedynie kopuły monotoniczne
- określają miarę na zbiorze (tak jak dystrybuanty)
- Twierdzenie Sklara:
  Dla każdego rozkładu prawdopodobieństwa charakteryzowanego dystrybuantą $F(x_1, \dots, x_n)$ istnieje taka kopuła $C(z_1, \dots, z_n)$, że dla rozkładów brzegowych charakteryzowanych dystrybuantami $F_1(x_1), \dots, F_n(x_n)$ zachodzi

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)).$$

Miara odległości pomiędzy zdarzeniem losowym i punktem z przestrzeni obserwacji

Odległość probabilistyczna

$$
\begin{aligned}
d(x;v) &= \int_{F^{-1}(\hat{x}-|\hat{x}-\hat{v}|)}^{F^{-1}(\hat{x}+|\hat{x}-\hat{v}|)} dF(x) = \\
&= \min(1,\hat{x}+|\hat{x}-\hat{v}|) - \max(0,\hat{x}-|\hat{x}-\hat{v}|)
\end{aligned}
$$

Uogólnienie odległości probabilistycznej: kopuły statystyczne

Uogólnienie odległości probabilistycznej

- odległość probabilistyczna jest jednowymiarową miarą
- poprzez liczenie miar hiperprostopadłościanów

Miara odległości pomiędzy 2 zdarzeniami losowymi

Odległość probabilistyczna

Można wykorzystać normę trójkątną w celu symetryzacji:

$$D(u, v) = 1 - T(1 - d(u; v), 1 - d(v; u)).$$

Specyfika klasyfikatora k–NN

Specyfika klasyfikatora k–NN

- najwyższa “rozdzielczość” dla niewielkich wartości miary odległości
- możliwość zastosowania przybliżeń prawdziwych w otoczeniu 0.

Miary odległości dla klasyfikatora k–NN

Miary odległości dla klasyfikatora k–NN

W pracy wyprowadzono poprzez przybliżanie rozwinięć określonych norm trójkątnych następujące typy łączeń:

- łączenie liniowe
- łączenie ważone
- łączenie oparte o miękkie normy trójkątne
- łączenie w oparciu parametryzowaną t–normę Yagera
- łączenie w oparciu o odległość Mahalanobisa.

Miary odległości dla klasyfikatora k–NN

Miary odległości dla klasyfikatora k–NN

W pracy wyprowadzono następujące typy symetryzacji:

- symetryzacja w oparciu o t–normę maksymalną
- symetryzacja w oparciu o t–normę maksymalną zastosowaną bezpośrednio
- symetryzacja w oparciu o parametryzowaną t–normę Yagera
- symetryzacja w oparciu o t–normę multiplikatywną
- symetryzacja w oparciu o t–normę Łukasiewicza
- symetryzacja w oparciu o wartość oczekiwaną argumentów.

Dziękuję za uwagę

Dziękuję za uwagę