

20. Podejmowanie decyzji w warunkach niepewności.

[Wstęp]

Rozpoznawanie obiektów – przypisanie do klasy na podstawie cech (algorytm rozpoznawania)

Schemat: obiekt opisany cechami → pomiar cech → przetwarzanie cech → klasyfikacja

Wyróżniamy 2 podstawowe rodzaje problemów rozpoznawania:

a) Rozpoznawanie z pełną informacją probabilistyczną

Wejścia x i wyjścia j są wartościami pary zmiennych losowych (x, j) . Zmienne losowe przyjmują wartości ze zbiorów wejść i wyjść obiektu, tj.

x – wektor wejść (cech) obiektu

X – przestrzeń wejść,

$x \in X \subseteq \mathbb{R}^n$

j – wynik klasyfikacji obiektu, klasa ze skończonego zbioru $\{1, \dots, J\}$

J – zbiór klas.

i – poprawny wynik klasyfikacji obiektu

Istnieje łączny rozkład prawdopodobieństwa pary zmiennych losowych (x, j) określony na $X \times J$

Funkcję gęstości rozkładu prawdopodobieństwa pary zmiennych losowych (x, j) oznaczmy przez $f(x, j)$.

Warunkową funkcję gęstości rozkładu prawdopodobieństwa zmiennej losowej j , pod warunkiem, że zmienna losowa x przyjmie wartość x , oznaczmy przez $f(j|x)$, a funkcję brzegowego rozkładu prawdopodobieństwa zmiennej losowej x oznaczmy przez $f_x(x)$.

Funkcja gęstości rozkładu prawdopodobieństwa $f(x, j)$ lub funkcja gęstości warunkowego rozkładu prawdopodobieństwa wyjścia, pod warunkiem, że wejście przyjęło określoną wartość $f(j|x)$, oraz funkcja gęstości rozkładu prawdopodobieństwa wejścia $f_x(x)$ są znane.

Przypadek ten nazwiemy „pełną informacją probabilistyczną”.

(W skrócie dalsza część polega na minimalizacji przyjętej funkcji strat – zwykle 0-1 funkcja strat, szukamy takiego $j = \psi(x)$, które ma minimalną wartość funkcji straty - odpowiednie całki znajdują się np. tu: http://www.dbc.wroc.pl/Content/3524/Jerzy_Swiatek.pdf)

b) Rozpoznawanie z niepełną informacją probabilistyczną

Rozkład prawdopodobieństwa pary zmiennych losowych (x, j) istnieje, ale nie jest znany.

W zamian dysponujemy zbiorem obserwacji pary zmiennych losowych (x, j) , czyli zbiorem:

(x_k, j_k) , $k=1, 2, \dots, n$. Przypadek ten nazwiemy „niepełną informacją probabilistyczną”. Możemy wtedy podejść do problemu na 2 sposoby:

- estymować rozkłady prawdopodobieństw na podstawie obserwacji (niestety skończona ilość obserwacji zwykle nie pozwala na wyznaczenie dokładnych rozkładów) i dla wyznaczonego rozkładu wykorzystać identyczne podejście jak przy pełnej informacji probabilistycznej
- zastosowanie inne algorytmy rozpoznawania nie wymagające pełnej informacji probabilistycznej np. odległościowe, drzewa klasyfikacyjne, naiwny Bayes etc.

Rozpoznawanie w oparciu o ciąg uczący – rozpoznawanie z nauczycielem

Algorytmy odległościowe:

- **Algorytm NM** (najbliższej średniej) – obliczamy wartość średnią dla każdej

klasy. Obliczamy odległość x od każdego elementu średniego, bierzemy min odległość i przypisujemy do klasy odpowiadającej tej wartości średniej

- **Algorytm KNN** (k najbliższych sąsiadów) – liczymy odległość x od każdego

elementu ciągu uczącego, sortujemy w porządku niemalejącym, bierzemy k pierwszych pozycji i wybieramy najczęściej występującą klasę

Algorytmy probabilistyczne

- **Naiwny Bayes** – zakładamy, że zmienne losowe – cechy klasy są niezależne. Dla każdej cechy X_k wyznaczamy wartości prawdopodobieństwa $p(J=j|X_k=x_k)$ – czyli podliczamy liczbę obiektów należących do klasy j i mających $X_k=x_k$ w stosunku do wszystkich obiektów klasy j w zbiorze uczącym. Wyznaczamy też prawdopodobieństwo $p(J=j)$. Przyporządkowujemy obiekt do tej klasy, dla której iloczyn powyższych prawdopodobieństw jest największy.

!!!

http://www.statsoft.pl/textbook/stathome_stat.html?http%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstnaiveb.html

Drzewa klasyfikacyjne

algorytm (ogólny) - 2 kroki:

- 1) reguła podziału wierzchołków
- 2) reguła uznania wierzchołka za końcowy oraz wiążąca się z tym reguła przypisania wierzchołka

Algorytm ID.3 (alg. budowy drzewa klasyfikacyjnego)

- 1) Dla każdej cechy liczymy entropię warunkową (w tym celu wyliczamy prawdopodobieństwa, że przykład przybierze daną wartość cechy oraz prawdopodobieństwa warunkowe takie że dla danej wartości cechy przykład będzie należał do klasy oraz dla danej wartości cechy przykład nie będzie należał do klasy)
- 2) Wybieramy cechę o najmniejszej wartości entropii i umieszczamy jako wierzchołek. Od wierzchołka odchodzą krawędzie etykietowane wartościami cechy (teraz będziemy brać pod uwagę tylko przykłady zawierające wartość cechy z tej krawędzi) itd.

Czasami mówi się też o rozpoznawaniu bez nadzoru – gdy nie znany jest zbiór klas I i należy go wyodrębnić. Mamy wtedy do czynienia z problemem i metodami grupowania (klasteryzacji), np. algorytmem k -średnich. Wydaje mi się, że lepiej nie poruszać tego zagadnienia niezapytany;)

Inne istotne zagadnienie powiązane z tematem to redukcja wymiaru/selekcja cech. Pozwala ono na ograniczenie zbioru cech tylko do niezbędnych do rozpoznawania i eliminację cech redundantnych, o niskiej entropii (niewielkiej informacji o klasie obiektu) lub silnie skorelowanych. Usunięcie niepotrzebnych cech pozwala na przyspieszenie procesu nauki, uproszczenie algorytmu i potencjalnie poprawę wyniku rozpoznawania – szczególnie widoczną np. w przypadku algorytmów naiwnego Bayesa.

Jeśli ktoś ma to nie zaszkodzi zajrzeć też do M. Kurzyński – Rozpoznawanie obiektów, metody statystyczne – szczególnie części o pełnej informacji probabilistycznej