Egzamin z Programowania Obiektowego, 2011-O6-O8

Jednym z ważniejszych problemów występujących w sztucznej inteligencji jest problem klasyfikacji. Polega on na tym, by w zadanej dziedzinie, na podstawie szczątkowej wiedzy na temat pojęć odróżniać (klasyfikować) obiekty należące do tych pojęć. Ta szczątkowa wiedza o pojedynczym pojęciu jest zadana poprzez wskazanie tych obiektów dziedziny, o których na pewno wiemy, że należą do tego pojęcia (szczegóły są podane dalej). To nasza cała wiedza o pojęciu - nie znamy jego definicji! Zakładamy, że pojęcia są parami rozłączne.

Dziedzina jest zbiorem obiektów. Każdy obiekt posiada skończony zbiór atrybutów, które go charakteryzują. Wszystkie obiekty z jednej dziedziny są charakteryzowane dokładnie takimi samymi atrybutami. Przez pojęcie rozumiemy jakiś podzbiór dziedziny.

Przykład 1: Dziedziną może być zbiór owoców, z których każdy ma atrybuty: kształt, rozmiar i kolor. Pojęciami w takiej dziedzinie może być bycie pewnym rodzajem owocu (np. bycie jabłkiem albo bycie śliwką).

Przykład 2: Dziedzina to zbiór studentów uczęszczających na przedmiot Programowanie Obiektowe w roku 2010/11. Atrybuty obiektów to imię, nazwisko, wyniki obu klasówek, liczba punktów uzyskanych z laboratorium, wynik egzaminu i procent obecności na ćwiczeniach. Pojęciem są studenci, którzy zaliczyli tę edycję przedmiotu.

W problemie klasyfikacji mamy dany pewien podzbiór obiektów dziedziny. Jest to zbiór treningowy - dla każdego obiektu z tego zbioru znane są wartości jego atrybutów oraz przynależność do odpowiedniego pojęcia. Zadaniem jest skonstruowanie algorytmu (zwanego klasyfikatorem), który będzie umiał z dużą precyzją sklasyfikować nowe (nie występujące w zbiorze treningowym) obiekty z dziedziny, dla których dane są jedynie wartości atrybutów - nie jest dla nich znana decyzja o przynależności do któregoś z pojęć. Wśród najprostszych, często stosowanych algorytmów klasyfikacji wyróżnić można m.in. drzewa decyzyjne, klasyfikatory bayesowskie, algorytm kNN (k najbliższych sąsiadów). .

Niektóre algorytmy wymagają, aby klasyfikowane obiekty spełniały dodatkowe założenia. Np. w algorytmie kNN wymagamy, aby w dziedzinie obiektów była zdefiniowana funkcja odległości między obiektami. Oczywiście dla tej samej dziedziny można wymyślić wiele rozsądnych odległości – stosując algorytm należy zadać tę wybraną. Algorytm kNN działa następująco:

- Dla obiektu, który chcemy sklasyfikować, znajdujemy k najbliższych (w sensie zadanej funkcji odległości) obiektów ze zbioru treningowego. Jeśli ze względu na równe odległości niektórych z obiektów zestaw k najbliższych obiektów nie jest jednoznacznie wyznaczony, to rozstrzygamy w naturalny sposób niejednoznaczności za pomocą losowania, np. dla k=3 i odległości najbliższych obiektów: 1, 5, 6, 6, 6, 9, ... wybieramy obiekty o odległościach 1 i 5, zaś spośród obiektów o odległości 6 losujemy jeden. Jeśli k jest większe od liczności zbioru treningowego, to zgłaszamy wyjątek.

- Sprawdzamy do jakich pojęć należy te znalezione k obiektów i przypisujemy klasyfikowanemu obiektowi pojęcie występujące najczęściej wśród tych k obiektów. Jeśli kilka pojęć występuje największą liczbę razy, to wybieramy jedno z pojęć losowo.

Zadanie

1. Zaprojektuj hierarchię klas dla opisanego problemu klasyfikacji (uwzględniającą algorytm kNN i umożliwiającą łatwe rozszerzenie hierarchii o inne algorytmy).

2. Zaimplementuj metodę *klasyfikuj*, która dla danego zbioru treningowego oraz danego obiektu z dziedziny (ale spoza zbioru treningowego) znajdzie odpowiednie dla tego obiektu pojęcie przy pomocy algorytmu kNN. Metoda *klasyfikuj* powinna mieć sygnaturę:

   Pojęcie<T> <T extends Obiekt> klasyfikuj(Map<T, Pojęcie<T>> z, T o)

3. Zaimplementuj zastosowanie metody *klasyfikuj* dla dziedziny owoców, w której obiekty posiadają atrybuty numeryczne: waga i rozmiar oraz symboliczne: kształt i kolor, a pojęciami mogą być np. banan, jabłko, pomarańcza, mandarynka. Możesz przyjąć dowolną definicję odległości w tej dziedzinie.

Uwaga: *Obiekt* to co innego niż klasa Object z Javy!

Powodzenia!!