Úkol B – Hledání podskupin v datech – Heart disease

ZS 2017

Cíl

Cílem druhého úkolu je seznámit se s praktickým použitím shlukovacích metod. Shlukovací metody lze využít k odhalení vnitřní struktury dat. Poznání vnitřní struktury dat může pomoci k lepšímu porozumění datům a jevům, které zachycují, stejně tak ke konstrukci lepších klasifikátorů.

Data

Heart disease data set je soubor dat, který se stal standardním testovacím souborem pro algoritmy strojového učení. Data byla naměřena v Cleveland Clinic Foundation v roce 1988 a od té doby jsou stále objektem zájmu mnoha výzkumníků. Pacienti kardiologické kliniky byli podrobeni sérii testů, přičemž z původních 76 naměřených příznaků se používá pouze 14 pro klasifikaci pacientů na zdravé a nemocné.

Hledání podskupin v datech

Pokuste se v datech nalézt vnitřní strukturu (podskupiny) na základě hodnot příznaků. Zvolte vhodnou metodu. Zamyslete se nad tím, zda je vhodné data normalizovat. Výsledky vizualizujte a pokuste se je interpretovat, např. pomocí typických (průměrných) reprezentantů. Prozkoumejte, které z příznaků se nejvíce podílí na rozdělení pozorování do jednotlivých shluků. Uvažte, jestli je možné dát nalezenou strukturu do souvislosti s klasifikací srdečních onemocnění.

Požadované kroky analýzy [10 bodů]

- upravte data pro shlukování [1 b]
- zvolte vhodnou metodu shlukování [1 b]
- zvolte vhodné parametry shlukovací metody [1 b]
- prezentujte výsledky shlukování pomocí typických reprezentantů shluků [3 b]
- zjistěte v jakých příznacích a jak se jednotlivé shluky liší [3 b]
- porovnejte výsledky shlukování s klasifikací srdečních onemocnění [1 b]

Výsledky upravte do formy zprávy, která bude obsahovat stručný **úvod**, popis metod, které jste použili, v sekci **metody,** výsledky jejich aplikace na data v sekci **výsledky** a závěry, které jste zjistili interpretací výsledků v sekci **závěr**. Maximální délka zprávy je 3 stránky. Zprávu ve formátu pdf odevzdejte pomocí UploadSystemu.

age	Věk v letech
sex	Pohlaví 1 = muž, 0 = žena
cp	Typ bolesti na hrudi: 1 = typická angina, 2 = atypická angina, 3 = neanginální bolest, 4 = asymptomatická
trestbps	Klidový krevní tlak (v mmHg)
chol	Sérový cholesterol v mg/dl
fbs	Krevní cukr nalačno > 120 mg/dl: 1 = true, 0 = false
restecg	Výsledek klidového EKG: 0 = normal, 1 = abnormalita ST-T vlny, 2 = ventrikulární hypertrofie
thalach	Maximální tepová frekvence
exang	Bolest na hrudi vyvolaná cvičením: 1 = yes, 0 = no
oldpeak	ST deprese vyvolaná cvičením vztažená ke klidovému stavu
slope	Strmost peaku: 1 = stoupající, 2 = neměmnící se, 3 = klesající
ca	Počet hlavních cév označených pomocí fluoroskopie
thal	3 = normální, 6 = opravený defekt, 7 = reversibilní defect
dis	Diagnóza: 0 = zdravý, 1 = nemocný

Tabulka 1: Stručný popis příznaků

Bonusová úloha – simulovaná data

Na rozdíl od hlavního úkolu je bonusová úloha zaměřená na simulovaná data. Data k bonusové úloze představují závislosti, která jsou těžká pro standardní algoritmy shlukování jako je k-means. Vaším úkolem je v tomto případě vyzkoušet na data metody shlukování prezentované v rámci DVZ a pokusit se najít takovou kombinaci parametrů metod (metriky, způsob linkování pozorování) a transformace příznaků (polynomiální příznaky, kernelová metoda), které by umožnili správně klasifikovat data v souborech *jain.csv*, *spiral.csv* a *pathbased.csv*. Všechna data obsahují správné řešení v proměnné *class*. Volbu metody zdůvodněte a vysvětlete, proč Váš přístup data správně shlukuje. Za bonusovou úlohu je možné získat až 3 body.