Úkol B – Hledání podskupin v datech – Forest fires

ZS 2016

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

Lesní požár je událost spojená s velkými riziky pro lesní ekonomii a současně ohrožující lidi, i když z hlediska ekologie a životního prostředí se jedná o přirozenou formu obnovy lesních společenství. Zvládnutelnost lesního požáru a jeho následky jsou ovlivněny počasím v době požáru. Data z portugalské studie nabízí zajímavé informace o vztazích mezi jednotlivými parametry počasí a velikostí lesních požárů.

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 plochou spáleniště.

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 plochou spáleniště [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.

X	odvozená x-ová souřadnice (od 1 do 9)
Y	odvozená y-ová souřadnice (od 1 do 9)
month	měsíc v roce (leden až prosinec)
day	den v týdnu (pondělí až neděle)
FFMC	FFMC kód
DMC	DMC kód
DC	DC kód
ISI	ISI index
temp	venkovní teplota (ve °C)
RH	venkovní relativní vlhkost (v %)
wind	venkovní rychlost větru (v km/h)
rain	venovní déšť (v mm/m^2)
area	celková plocha spáleniště (v ha)

Tabulka 1: Popis proměnných z studie lesních požárů v severovýchodním Portugalsku

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.