Shluková analýza hub

Luboš Mátl

***Abstrakt*—Správné rozpoznání jedovatých hub je pro milovníky hub velmi důležité. Tento dokument se snaží nalézt nejlepší shlukovací algoritmus pro tyto data a nastavit ho tak, aby produkoval co nejlepší výsledky. Na základě tohoto algoritmu můžeme houby rozdělit do tříd a z těchto tříd zjišťovat další vlastnosti hub.**

I. Assignment

Vyberte si data, která jsou použitelná pro shlukování. Data předzpracujte, a naimportujte do DM aplikace. Zvolte si shlukovací algoritmus (k-means, hierarchické shlukování, SOM). Najděte nastavení parametrů zvoleného algoritmu tak, aby produkoval co nejlepší výsledky. Interpretujte výsledky.

II. Introduction

Jako shlukovací algoritmy jsem si vybral K-means. Pro tento algoritmus se budu snažit nalézt nejlepší nastavení. Nejprve popíši, jak algoritmus pracuje, následně budu nastavovat parametry a nastavení porovnávat.

Vstupní data měla všechny hodnoty atributů v textové podobě. Bylo nutné je převést na číselné hodnoty. Atributy, které, s jejich hodnotami, můžete vidět v tabulce 1, jsem převedl na číselnou hodnotu podle jejich pořadí (tak jak jsou uvedeny v tabulce 1)

První atribut, je tzv. "učitel". Rozhoduje, zda je tato houba jedlá či nejedlá, tento atribut budu pro shlukování ignorovat, protože se snažím nalézt pouze algoritmus, který produkuje co nejlepší výsledky ve shlukování (učení bez učitele). Pro vytvoření modelu budu používat jejích dalších 21 atributů (které můžete vidět v tabulce 1). Položek v datech je mnoho 8124, použil jsem je všechny, i když testy trvaly velmi dlouho.

název atributu	**hodnoty atributu**
Jedlost	jedlé, nejedlé
Tvar klobouku	zvonovitý, kónický, konvexní, plochý, boulovatý, propadlý
Struktura klobouku	vláknitý, drážkovaný, šipinatý, hladký
barva klobouku	hnědá, žlutohnědá, skořicová, šedá, zelená, růžová, fialová, červená, bílá
Modrá	ano, ne
Zápach	mandlový, anýz, kreozotová, rybí, hnilobný, zatuchlý, žádný, štiplavý, aromatický
Žebrování	připojení, klesající, volné, vrubové
Rozteč žebrování	blízko, mnoho na jednom místě, daleko
Šířka žebrování	široká, úzká
Barva žebrování	černá, hnědá, žlutohnědá, čokoládová, šedá, zelená, oranžová, růžová, fialová, červená, bílá, žlutá
Tvar stonku	rozšiřující, zužující
Kořen stonku	bahňatý, kuželovitý, hrnkovitý, rovný, izomorfní, kořeny, chybí
Povrch stonku nad kloboukem	vláknitý, šupinatý, hedvábný, hladký
Povrch stonku pod kloboukem	vláknitý, šupinatý, hedvábný, hladký
Barva stonku nad kloboukem	hnědá, žlutohnědá, skořicová, šedá oranžová, růžová, červená, bílá, žlutá
Barva stonku pod kloboukem	hnědá, žlutohnědá, skořicová, šedá oranžová, růžová, červená, bílá, žlutá
Barva závoje	hnědá, oranžová, bílá, žlutá
Počet závojů	žádný, jeden, dva
Typ závoje	pavučinový, mizící, zvonový, velký, žádný, visicí, obalený, zónový
Nepravidelně se vyskytující barva	černá, hnědá, žlutohnědá, čokoládový, zelená, oranžová, fialová, bílá, žlutá
Množství na jednom místě	bohatá, chumel, mnoho, rozptýlené, několik, osamocená
Umístění	tráva, listy, louky, cesty, města, odpad, lesy

Tabulka I
Jednotlivé atributy hub

III. Methodology

A. K-means

Tento algoritmus se vyhýbá výpočtu všech vzdáleností dvojic instancí, tak že vytvoří reprezentanty (centroidy). Tyto reprezentanty náhodně nainicializuje. Následně vypočte vzdálenosti od všech nejbližších instancí, k danému centroidu. Z těchto instancí vypočte následující polohu centroidů (každý centroid se posune blíž k instancím). Toto se opakuje, dokud se centroidy nějak pohybují.

Zjednodušeně můžeme algoritmus popsat takto:

Mějme body x_i patřící do R_n a číslo k určující počet shluků.

1) Vytvoř počáteční centroidy c_i .
2) Pro každý bod x_i spočítej příslušnost bodu ke clusteru c_i .
3) přepočítej polohu každého centroidu c_i .
4) Opakuj od bodu 2, dokud se poloha centroidů mění

U tohoto algoritmu můžeme nastavit počet cenroidů a metodu, která bude počítat vzdálenosti instancí od centroidů. Já jsem použil tyto tři metody:

1) **Eukleidovskou vzdálenost**

Dva body v n-rozměrném prostoru:

$$P = (p_1, p_2, ..., p_n) Q = (q_1, q_2, ..., q_n)$$

$$E(PQ) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

počet centroidů	Eukleidovská vzdálenost	Manhattonská vzdálenost	Cosinova vzdálenost
2	0.3968	0.3976	0.4015
3	0.3955	0.3817	0.3972
5	0.5171	0.4847	0.5196
10	**0.5361**	0.3759	**0.5920**
11	0.5311	0.4818	0.5224
12	0.5064	0.4386	0.5388
13	0.4679	**0.5004**	0.5408
14	0.4452	0.3431	0.5117
15	0.3718	0.4375	0.4156
16	0.3975	0.3086	0.4178
20	0.4093	0.204	0.3752
50	0.4182	0.2002	0.409

Tabulka II
TABULKA RŮZNĚ NASTAVENÝCH ALGORITMŮ A JEJICH HODNOT SILHOUETTE

2) **Cosinovu vzdálenost**

$$dist(p,q) = arccos(p.q/|q||p|)$$

3) **Manhattonskou vzdálenost**

$$M(P,Q) = |p_1 - q_1| + |p_2 - q_2| + ... + |p_n - q_n|$$

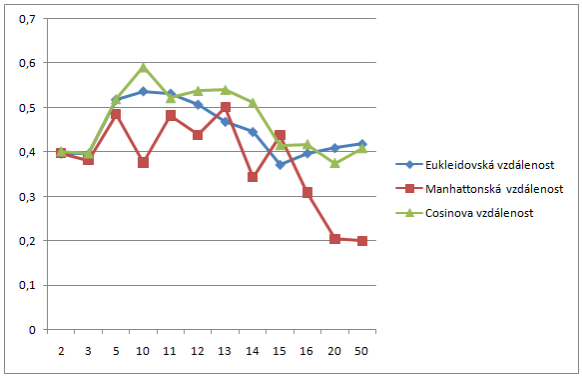
Také jsem měřil s různým počtem shluků a každou metodu jsem vždy (z důvodu náhodnosti nastavení středů) opakoval 5x a vybral nejlepší řešení.

Porovnávání jednotlivých algoritmů se provádí pomocí funkce silhouette. Tato funkce spočítá, jak moc patří jednotlivé body do shluků. Její vzorec je:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i)}$$

kde a(i) je průměrná vzdálenost instance i od instancí shluku, do kterého je zařazena a b(i) je průměrná vzdálenost instance i od instancí nejbližšího shluku

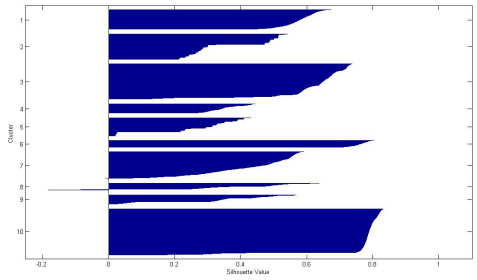
Čím vyšší je hodnota této funkce, tím je výsledný algoritmus lepší. V tabulce 1 můžete vidět výsledky měření (nejlepší výsledky jsou označeny tučně).



Obrázek 1. Graf ůzně nastavených algoritmů a jejich hodnot silhouette

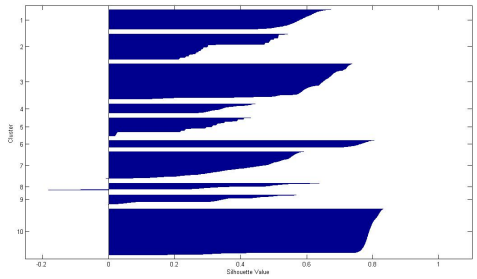
Jak v grafu 1 tak v tabulce 1 můžete vidět, že nejlepší výsledky pro shlukování jsou:

1) **Eukleidovskou vzdálenost** (obrázek 2)
silhouette: 0.5361
počet centroidů: 10



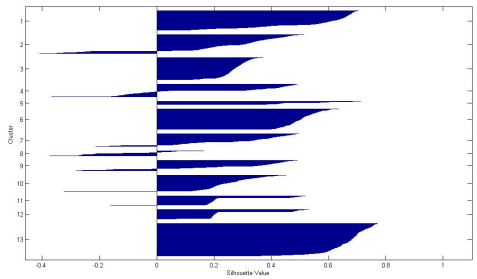
Obrázek 2. Graf silhouette pro Eukleidovskou vzdálenost, který má 10 centroidů

2) **Cosinovu vzdálenost** (obrázek 3)
silhouette: 0.5920
počet centroidů: 10



Obrázek 3. Graf silhouette pro Eukleidovskou vzdálenost, který má 10 centroidů

3) **Manhattonskou vzdálenost** (obrázek 4)
silhouette: 0.5004
počet centroidů: 13



Obrázek 4. Graf silhouette pro Manhattonskou vzdálenost, který má 13 centroidů

Nejlépe si vedla metoda, která určuje vzdálenosti od clussterů pomocí cosinovy vzdálenosti, která jak můžeme vidět na má nejlepší silhouette (až na některé výchylky) v celém rozsahu shluků.

IV. Conclusion

Při měření jsem zjistil, že model bude dávat nejlepší výsledky pro počítání vzdálenosti pomocí Cosinovy vzdálenosti a nastavení počtu centroidů na 10.

Reference

[1] Materiály na stránkách předmětu Y336VD
http://cw.felk.cvut.cz/doku.php/courses/y336vd/start