Klasifikace hub

Luboš Mátl

***Abstrakt*—Správné rozpoznání jedovatých hub je pro milovníky hub velmi důležité. Tento dokument se snaží popsat, zda lze rozpoznat houby podle specifických vlastností a s jakou pravděpodobností se uživatel splete.**

I. Assignment

Vyberte si data, která jsou použitelná pro klasifikaci nebo regresi. Data předzpracujte, a naimportujte do DM aplikace. Zvolte si klasifikátor (kNN, lineární a polynomiální separace, rozhodovací strom) nebo regresní algoritmus (kNN, lineární a polynomiální regrese, regresní strom). Najděte nastavení parametrů zvoleného algoritmu tak, aby produkoval co nejlepší modely.

II. Introduction

Pro vytvoření modelu jsem vybral dvě metody lineární a polynomiální separaci a rozhodovací stromy. Nejdříve se pokusím nalézt nejlepší nastavení parametrů pro obě metody a následně zvolím, která z metod je pro tento případ lepší. 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). Po převedení hodnot atributů na číselnou podobu, je bylo nutné normalizovat.
První atribut, je tzv. ”učitel”. Rozhoduje zda je tato houba jedlá či nejedlá. Budu se snažit vytvořit model, který nejlépe pozná, zda je houba jedovatá nebo jedlá podle 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.

III. Methodology

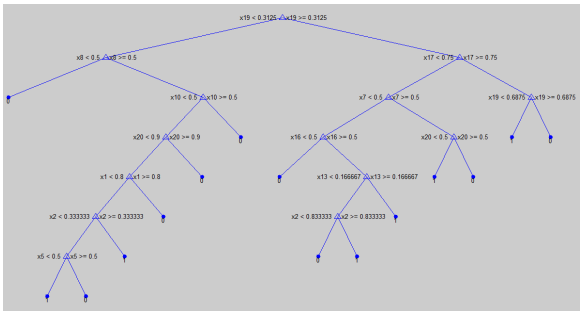
A. Rozhodovací stromy

Při vytváření rozhodovacího stromů vytváříme podmínky, pro jednotlivé uzly stromu. Každý uzel stromu (kromě listů) reprezentuje podmínku na hodnotu určitého příznaku. Každý list reprezentuje jednotlivou výstupní třídu. Na obrázku 1 můžete vidět výstupní model rozhodovacího stromu. Který lze přepsat na podmínky:

1) if $x19 < 0.31250.3125$ then node 2 else node 3
2) if $x8 < 0.5$ then node 4 else node 5
3) if $x17 < 0.75$ then node 6 else node 7
4) class = 0
5) if $x2 < 0.333333$ then node 8 else node 9
6) if $x16 < 0.5$ then node 10 else node 11
7) if $x19 < 0.6875$ then node 12 else node 13
8) if $x21 < 0.833333$ then node 14 else node 15
9) if $x5 < 0.1875$ then node 16 else node 17
10) class = 0
11) if $x7 < 0.5$ then node 18 else node 19
12) class = 1
13) class = 0
14) class = 0
15) if $x4 < 0.5$ then node 20 else node 21
16) class = 0
17) class = 1
18) if $x13 < 0.166667$ then node 22 else node 23
19) if $x20 < 0.5$ then node 24 else node 25
20) class = 1

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



Obrázek 1. Klasifikační strom

21) class = 0
22) if $x2 < 0.833333$ then node 26 else node 27
23) class = 1
24) class = 1
25) class = 0
26) class = 0
27) class = 1

Atribut, podle kterého budeme v daném uzlu rozhodovat, můžeme vybrat podle Entropie množiny instancí:

$$H(D) = \sum_{i=1}^{t} -p_i \log p_i$$

$p_1, p_2, ..., p_t$ poměrné velikosti tříd

$p_i, =$ počet instancí třídy i v D / počet všech instancí v D

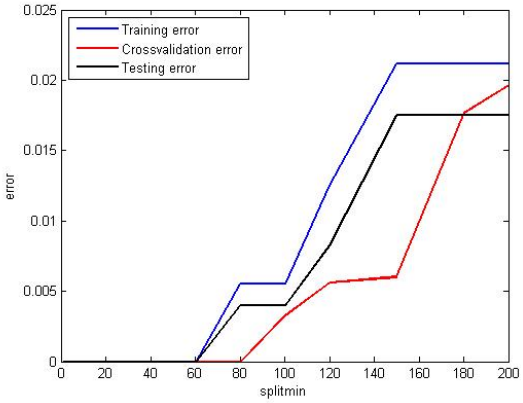
Po výpočtu entropií všech atributů, vybereme ten atribut, který má nejmenší entropii. Takto postupujeme dokud není celý strom rozdělený nebo dokud není splněna nějaká další podmínka (např. minimální počet dat v listu).

Na obrázku 2 je graf, který popisuje chyby klasifikační metody, při různé složitosti vyhodnocovacího stromu (minimálního počtu trénovacích případů v uzlu). Graf využívá 60% dat na trénování, ostatní data používá pro testování. Z tohoto grafu lze vyčíst, že z atributů hub zle velmi přesně zjistit zda je houba jedovatá nebo ne. Pro rozdělení od 2 do 60 trénovacích případů v uzlu. Jsou všechny chyby nulové (trénovací, crossvalidační i testovací). Dokonce ani nedochází k žádnému přeučení.

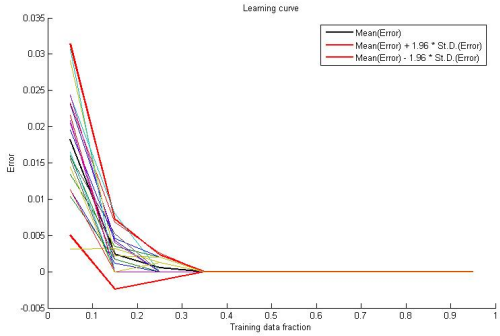
Graf na obrázku 3 zachycuje přesnosti modelů, které jsou vytvořeny podle určité velikosti trénovacích dat. Minimální počet trénovacích případů v uzlu je pro každý model nastaven na 10. Na grafu je vidět, že při použití více než 35% dat jako trénovacích, nedochází k žádné chybě. Opět se neprojevuje přeučení.

B. Lineární a polynomiální separace

Tato metoda hledá model ve formě separačních přímek. Atributů je více, proto hledáme tzv. nadrovinu (hyperrovinu). Třídy jsou pouze dvě (houby jsou jedlé nebo nejedlé), takže stačí nalézt jednu nadrovinu a nebude docházet ke vzniku oblastem nejednoznačnosti. Snažíme se nalézt rovinu, která nejlépe rozdělí houby na jedlé a nejedlé. Pro vytvoření této



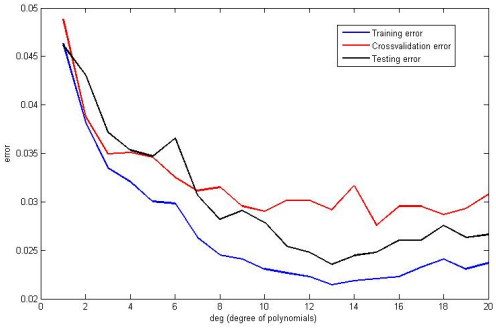
Obrázek 2. Chyba v závislosti na ohebnosti modelu



Obrázek 3. Přesnost modelu v závislosti na velikosti trénovacích dat

roviny použijeme perceptronový algoritmus. Protože lineární separace určitě nebude nejlepším modelem, budeme postupně rozšiřovat bázi (převádět lineární separaci na nelineární) a zjišťovat pro jaký stupeň polynomu je model nejlepší.

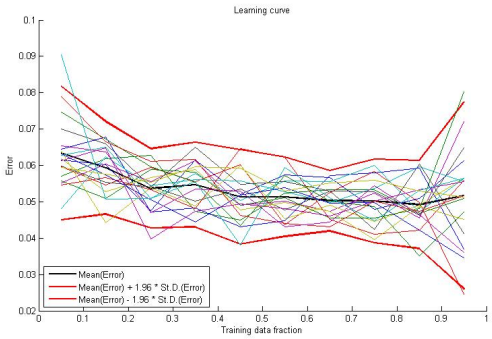
Graf na obrázku 4 porovnává chyby modelů s různým stupněm polynomu. Můžeme vidět, že u této metody jsou chyby daleko větší a nedostaneme se k nulové hodnotě chyb jako u rozhodovacích stromů. Při nastavení stupně polynomu na hodnotu 13 jsou chyby nejmenší.



Obrázek 4. Chyba v závislosti na ohebnosti modelu

Graf na obrázku 5 zachycuje přesnosti modelů, které jsou

vytvořeny podle určité velikosti trénovacích dat. Pro nejvyšší přesnost metod musíme stanovit velikost trénovacích dat mezi 60% a 70%.



Obrázek 5. Přesnost modelu v závislosti na velikosti trénovacích dat

IV. Conclusion

Při měření jsem zjistil, že metoda rozhodovacích stromů vytváří modely, které mají velikou přesnost při nastavení minimálního počtu trénovacích případů v uzlu mezi 2 a 60, dokonce je tato chyba nulová. Metoda lineární a polynomiální separace nedosahuje takových výsledků, jako metoda rozhodovacích stromů. Její nejlepší nastavení stupně polynomu je 13. Skutečná chyba tohoto modelu, při nejlepším nastavení, je 0,024. Proto bych pro tyto data raději volil metodu rozhodovacích stromů.

Reference

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