

Abstrakt práce

1. Téma práce: Porovnání Bayesovského klasifikátoru založeného na vybraných třídách kopulí s dalšími typy klasifikátorů na základě ROC křivek

2. Meno autora: Tomáš Kucharzyk

3. Ročník a odbor štúdia: 2. rok navazující studium, Manažerská informatika

4. Konzultant/vedúci práce: Ing. Jan GÓRECKI, Ph.D.

5. Obsah:

Klasifikace je proces, při něm klasifikátor analyzuje datovou sadu (například tabulka dat) a rozhoduje, zda zkoumaný znak nebo soubor znaků, patří do určité třídy. Klasifikátory neboli klasifikační metody jsou matematické funkce, které tvoří klasifikační algoritmy. Používají se ke klasifikaci dat. Přiblížit se to dá na jednoduchém příkladu. Naší datovou sadou je seznam pacientů. Každý pacient má svůj soubor znaků, kterými mohou být například různé hodnoty krevních testů, věk, váha a podobně. Podle těchto znaků se dá rozhodnout, do které třídy pacient patří. V našem případě máme třídy dvě. Jedna třída je nemocný a druhá není nemocný. Klasifikátoru se nejdříve učí na všech dostupných datech, jaké znaky vykazuje nemocný pacient a naopak. Poté klasifikátor klasifikuje nová data, u kterých není známo, zda je pacient nemocný a podle znaků, které pacienti vykazují, je zařazuje do tříd nemocný a není nemocný. Třidu může určovat například podle toho, s jakou pravděpodobností pacient spadá do jednotlivé třídy. Vyšší pravděpodobnost u třídy nemocný znamená, že bude pacient klasifikován jako nemocný (patřící do třídy nemocný). Kvalitu klasifikace lze měřit různými způsoby a nástroji. Tato práce porovnává klasifikátory pomocí tzv. receiver operating characteristic (ROC) křivek. Tyto křivky popisují určité vlastnosti klasifikace na datové sadě a jsou vhodným nástrojem pro porovnání kvality klasifikace. Jedním z nejpoužívanějších ukazatelů, který se používá pro porovnání různých ROC křivek mezi sebou je tzv. area under curve (AUC) česky plocha pod křivkou. AUC ohodnotí každou křivku hodnotou od nuly až do jedné. Čím vyšší je tato hodnota, tím se klasifikace považuje za kvalitnější. Dá se tak jasně říct,

který klasifikační model ze všech modelů získaných různými klasifikačními metodami byl na dané datové sadě kvalitnější.

Výzkum je zaměřen na porovnání nových klasifikačních metod navržených v rámci projektu Studentské grantové soutěže na Obchodně podnikatelské fakultě Slezské univerzity v Karviné s nejúspěšnějšími již existujícími metodami, jako jsou Naive bayes, Support vector machines a Random forests na vybraných datových sadách. Postupně jsem srovnal klasifikátory na všech datových sadách a výsledky, které jsem získal, prokázaly, že nově navržené metody v průměru vykazují kvalitu klasifikace srovnatelnou s těmi v současné době nejlepšími klasifikačními metodami a pro určitá data vykazují nejlepší kvalitu ve srovnání s ostatními metodami, což umožní nově navrženým metodám etablovat se v oblasti klasifikačních metod.

6. Klíčové slová: ROC křivky, klasifikace, klasifikátor, AUC, Bayes.