Študentská vedecká aktivita

Porovnání Bayesovského klasifikátoru založeného na vybraných třídách kopulí s dalšími typy klasifikátorů na základě ROC křivek

Tomáš Kucharzyk

Sekcia: Kvantitatívne metódy a informatika

Konzultant: Ing. Jan Górecki, Ph.D.

Karviná

2016

ABSTRAKT

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 byly srovnány klasifikátory na všech datových sadách a získané výsledky 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.

OBSAH

Úvod 1
1 Klasifikace 2
2 Vybrané klasifikátory a data sety 3
2.1 Vybrané klasifikátory 3
2.2 Vybrané data sety 5
3 Metody porovnání kvality klasifikace 5
3.1 ROC graf 6
3.2 Ukazatele 7
4 Hodnocení kvality klasifikace na data setu Breast Tissue 9
5 Celkové shrnutí všech datových setů 14
Závěr 16
Seznam použitých pramenů a literatury 17

Úvod

Klasifikace je problém, který, ačkoliv si to neuvědomují, řeší lidé na celém světě každý den. Denně zařazujeme nové poznatky a vjemy do různých kategorií (tříd). Díky rozvoji klasifikačních metod a algoritmů dokážou dnes tento problém řešit za pomoci strojového učení a umělé inteligence výkonné počítače. Klasifikační metoda (klasifikátor) dokáže roztřídit obrovské množství dat, které by člověk nikdy neměl šanci klasifikovat sám.

Klasifikace má velice široké využití v mnoha oborech. V posledních letech se začala velice hojně využívat v oblasti lékařské diagnózy, kde dokáže určit, zda má pacient šanci dostat určitou chorobu. Když se pacient dozví, že je zde vysoké riziko onemocnění, může mu snadno předejít a vyhnout se tak komplikacím zdravotního stavu. Tento způsob využití poukazuje jak důležitý je vývoj klasifikačních metod a jejich zkoumání.

Mnoho klasifikačních metod využíváme každý den a ani o nich nevíme. Například algoritmy vyhledávacích služeb na internetu. Rozpoznávání řeči a obrázků (Google, Microsoft How-Old) nebo například jedno z nejčastějších využití, a to cílený marketing, kde klasifikátory dokáží identifikovat potenciální zákazníky. Velkým úspěchem rozvoje klasifikace se letos v březnu stalo vítězství počítačového programu společnosti Google AlphaGo, který porazil světového mistra ve stolní strategické hře Go. Tento program funguje na principu klasifikace pomoci neuronových sítí. S nástupem sociálních sítí získala klasifikace další pole své působnosti, velké kvanta dat, které tyto sítě generují, jsou zkoumána klasifikačními metodami a využívána na mnoho různých účelů, například pro předvídání jaký obsah uživateli nabídnout.

S rozvojem velkých datových skladů a takzvaných Big data se rozvíjí i výzkum klasifikačních metod. Jsou vyvíjeny nové klasifikační metody a způsoby jak s těmito daty nakládat. Tato práce je zaměřena na porovnání nově vyvinutých klasifikátorů založených na archimedovských a hierarchických archimedovských třídách kopulí s těmi nejlepšími jako jsou Naive Bayes, Support vector machines nebo Decision tree. V této práci jsou, jako poprvé na světě, porovnány archimedovské klasifikátory pomocí ROC křivek. Kvalita klasifikace byla provedena na šesti klasifikačních datových setech a hodnocena metodami jako je plocha pod křivkou (AUC), přesnost predikce a optimální bod operační křivky.

Cílem práce je porovnat nově vzniklé klasifikátory a zhodnotit, zda dokáží konkurovat těm stávajícím a stát se tak vhodnou alternativou.

1 Klasifikace

V dnešní době se v databázích nachází obrovské množství dat, které se stále zvětšuje. Tak vznikla příležitost i potřeba tyto data prozkoumat a získávat z nich znalosti, které by mohly vylepšit rozhodovací proces různých společností. Dobývání znalostí a dolování dat je název často používaný pro multidisciplinární obor, ve kterém se prolínají statistika a strojové učení za účelem získávání užitečných dat z velkých databází. (Freitas, 2002)

Klasifikace je jedna z nejstudovanějších částí procesu dolování dat. Proces klasifikace souvisí s úkoly, se kterými se setkáváme každý den. Například doktoři v nemocnici klasifikují pacienty podle toho, jak moc jsou ohrožení určitou chorobou, nebo učitel rozhodne, jakou známku by měly dostat práce jednotlivých studentů. (Bramer, 2013)

Klasifikace se provádí pomoci klasifikátoru, ten představuje matematická funkce, nebo algoritmus, který odpovídá charakteru analyzovaných dat. (Holčík, 2012).

Při klasifikaci patří každý případ do předem určené třídy, která je určena podle jednotlivých atributů, které případ vykazuje. Klasifikace je poté proces, při němž klasifikátor zkoumá atributy jednotlivých případů a podle jejich hodnot a spojitostí mezi nimi rozhodne, do které třídy případ patří. Atributy, které případ vykazuje, můžou být například výsledky lékařských vyšetření jednoho pacienta, které určují, zda pacient patří do třídy ohrožených nemocí nebo ne. Důležité je, aby zaznamenané atributy takového pacienta byly relevantní vzhledem k zjišťovaným třídám (například jméno pacienta nijak nesouvisí s pacientovým zdravotním stavem). (Freitas, 2002)

Pro správný postup klasifikace je také důležité rozdělit zkoumaná data na testovací a trénovací set. Množinu klasifikátor nejdříve natrénuje klasifikaci na trénovacím setu a vytvoří model, který bude určovat, který případ patří do které třídy. Poté aplikuje natrénovaný model na testovací data, u kterých předpovídá, do které třídy patří (předem nezná třídu testovaných případů). Toto rozdělení dat se nazývá křížová validace (crossvalidation), běžně se na jednom setu dat opakuje například desetkrát. Postupně se tak otestuje deset odlišných testovacích setů, které dohromady obsahují kompletní data. Na základě testování se poté hodnotí výkon klasifikátoru. (Freitas 2002)

2 Vybrané klasifikátory a data sety

2.1 Vybrané klasifikátory

V této práci jsou porovnány nové bayesovské klasifikační metody založené na archimedovských a hierarchických archimedovských kopulích s nejúspěšnějšími dostupnými klasifikačními metodami jako jsou Naive Bayes, Decision tree, Support vector machines. Dále byla přidána ještě metoda K-nearest neighbour a metody založené na studentových t-kopulích a gaussovských kopulích.

Ve zbývající části této kapitoly budou vybrané metody stručně popsány.

NAIVE BAYES

Takzvaný Naivní Bayesův klasifikátor je hojně využívaný klasifikátor, který si získal velkou oblibu především díky své jednoduchosti a efektivitě. Přes svoji jednoduchost dokáže dosahovat podobné výsledky jako sofistikovanější klasifikační metody. Naivní Bayesův klasifikátor dosahuje také vysoké přesnosti a rychlosti i na velkých data setech. Naivní Bayes je založen na Bayesově teorému a patří mezi takzvané pravděpodobnostní klasifikátory (výsledkem pro každý případ je pravděpodobnost, s jakou patří do jednotlivých tříd). (Aggarwal, 2015)

DECISION TREE

Decission tree, neboli rozhodovací strom, je klasifikátor, který má strukturu stromu. Jednotlivá data vstupují u kořenů a postupně postupují přes kmen stromu až do koruny, kde na úplném konci představuje jeden list jednu třídu. Rozhodovací se mu říká proto, že v každém bodu, přes který data procházejí, dochází k rozhodnutí kterým směrem (ke které třídě) budou v daném stromu dále pokračovat. Klasifikátor tak protřídí jednotlivé atributy případů přes rozhodnutí a ta ho dovedou k určité třídě. Rozhodovací strom je využívaný v mnoha odvětvích, a to především díky jednoduchému použití a snadné intepretaci. Klasifikátor Decision tree patří mezi pravděpodobnostní klasifikátory stejně jako Naivní Bayes. (Aggarwal, 2015)

SUPPORT VECTOR MACHINES (SVM)

Klasifikátor SVM je nepravděpodobnostní binární klasifikátor. To znamená, že přiřazuje jednotlivé případy buď do jedné, nebo druhé třídy. SVM pracuje na základě vytváření rovin, které oddělují vektory. Oddělí tak prostor (rovinu), do kterého patří případy z jedné a druhé třídy na základě podobnosti jejich atribut (pro příklad uvažujeme dvě třídy). Pří zkoumání nového případu pak určuje, ke které rovině má zkoumaný případ nejblíže a tam ho přiřadí. Klasifikátor SVM používá takzvanou jádrovou transformaci a dokáže tak převést data, která nejsou lineárně rozdělitelná na data, jež lze lineárně rozdělit (např. dokáže transformovat data, která nejdou vektorem rozdělit do dvou rovin tak, aby šla rozdělit). Klasifikátor se stal velice populárním a je považován za jeden z nejvýkonnějších. (Nisbert, Entezari, 2009)

K-NEAREST NEIGBOURS (KNN)

Klasifikační metoda K-nejbližších sousedů patří k těm jednodušším. Na velkých data setech, ve kterých vykazuje každý případ mnoho atributů, nedosahuje takových výkonů jako předešlé zmíněné klasifikátory. KNN se musí nejdříve naučit, jaké atributy mají jednotlivé případy ze zkoumaných tříd. Tyto případy si zakreslí do N-rozměrného prostoru, kde N je počet atributů každého případu. Při testování pak zakreslí zkoumané případy do téhož prostoru a najde nejbližší sousedy. Případ poté klasifikuje do třídy podle toho, do které třídy spadá většina jeho sousedů. (Thorsten, 2002 a Entezari, 2009)

KLASIFIKÁTOR ZALOŽENÝ NA KOPULÍCH

Celkem byly použity čtyři klasifikátory založené na čtyřech třídách kopulí. Konkrétně to jsou tyto třídy:

- Archimedovské
- Hierarchické archimedovské
- Gaussovské
- Studentovy t-kopule

Klasifikátory založené na archimedovských a hierarchických archimedovských kopulích jsou nové klasifikační metody, které byly vyvinuty na Obchodně podnikatelské

fakultě Slezské univerzity v Opavě v rámci výzkumu v projektu Studentské grantové soutěže. Tyto klasifikátory jsou postaveny na základě klasifikátoru Naive Bayes. Ten pro klasifikaci využívá jednoduchou kopuli, která předpokládá, že jsou na sobě hodnoty atributů nezávislé. Tato kopule byla nahrazena výše zmíněnými třídami kopulí, takže klasifikátor předpokládá závislosti mezi atributy, což by mnělo vést k navýšení kvality klasifikace.

2.2 Vybrané data sety

V práci jsou použity klasifikační data sety Fisher Iris, Wine, Seeds, Vertebral column, Breast Tissues a Appendicitis. Všechny atributy použitých data setů jsou číselné. Tyto data sety byly získány ze stránek UCI Repository[1].

Stručný přehled datových setů je zaznamenán do následující tabulky č 1.

Tabulka 1:Přehled data setů

Název datového setu	Počet případů	Počet atributů	Počet tříd
Appendicitis	106	7	2
Breast Tissue	106	9	4
Fisher Iris	150	4	3
Seeds	210	7	3
Vertebral column	310	6	2
Wine	178	13	3

Zdroj: Vlastní zpracování

Pro účely této práce byly prezentovány výsledky z vybraného datového setu Breast Tissue. Ten obsahuje výsledky měření vyříznuté prsní tkáně za pomocí elektrické impedance na různých frekvencích. Datový set obsahuje čtyři třídy, třídu Adi – 22 případů, Car – 21 případů, Con - 14 a Merged – 49 případů.

3 Metody porovnání kvality klasifikace

Veškeré modelování a výpočty v této práci probíhaly ve vývojovém prostředí aplikace MATLAB. Tato aplikace byla zvolena jako vhodný nástroj, jelikož obsahuje veškeré potřebné funkce a algoritmy, které byly během práce použity.

[1] UCI Machine learning repository. [online]. Dostupné z: https://archive.ics.uci.edu/ml/datasets.html.

3.1 ROC graf

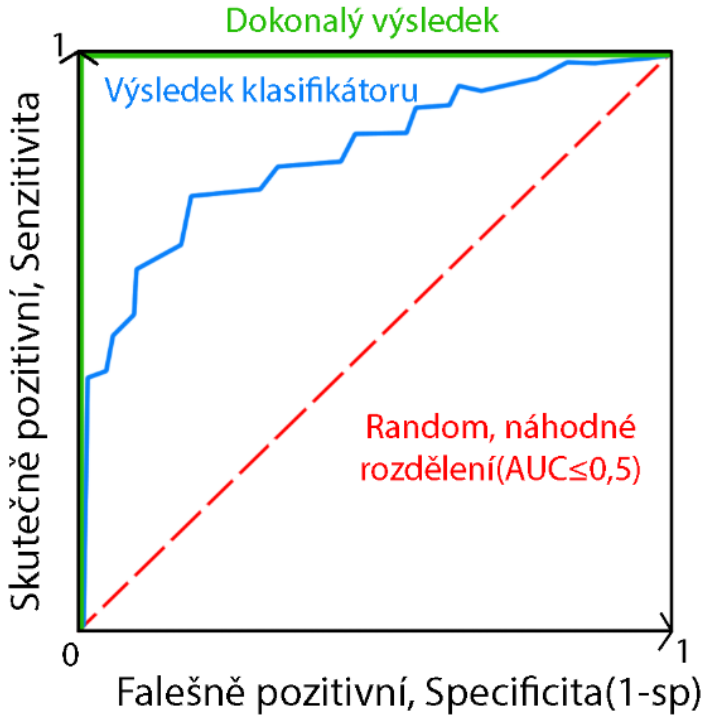
Jako hlavní porovnávací metoda, byly zvoleny ROC (Receiver Operating Characteristic) křivky, které jsou vhodným prostředkem k posuzování kvality klasifikace jednotlivých klasifikátorů.

Pro porovnání klasifikátorů pomocí ROC křivek je nutné rozdělit třídy v data setu na pozitivní a negativní. Tyto třídy se pak při testu postaví proti sobě. Například uvažujeme-li tři třídy, postavíme vždy jednu pozitivní proti dvěma negativním. Pro takový data set lze poté získat tři ROC grafy, a to tak, že vždy zvolím jednu třídu jako pozitivní a zbylé dvě jako negativní.

Klasifikátor v průběhu klasifikace určuje, zda daný případ (například naměřené hodnoty jednoho pacienta s podezřením na zánět slepého střeva) spadá do třídy pozitivní nebo negativní. Pokud pacient má zánět slepého střeva (je nemocný a spadá do třídy jedna, třída jedna je pozitivní) a klasifikátor ho klasifikuje jako pozitivního, udělá tak skutečně pozitivní klasifikaci třídy, označí totiž pacienta, který spadá do pozitivní třídy, jako pozitivního. Pokud je pacient zdravý (jeho hodnoty patří do třídy nula) a klasifikátor ho označí jako pozitivního, vznikne falešně pozitivní klasifikace třídy.

Následující obrázek č 1 zobrazuje ROC graf.

Obrázek 1: ROC graf



Zdroj: Vlastní zpracování

Vodorovná osa zachycuje míru falešně pozitivních případů a specificitu (v tomto případě 1 - specificita). Míra falešné pozitivity vyjadřuje množství negativních případů, které jsou určeny jako pozitivní. Specificita je poměr správně negativních případů ke všem negativním případům. Nabývá hodnot 0 až 1 nebo 0% až 100%. Vyjadřuje množství negativních případů, které jsou jako takové klasifikovány.

$$\text{specificita} = \frac{\text{počet skutečně negativních}}{\text{počet skutečně negativních} + \text{počet falešně pozitivních}}$$

Specificita laicky řečeno ukazuje schopnost klasifikátoru určit případy, které nejsou pozitivní. Nejlepší výkon klasifikátoru vzhledem k falešné pozitivitě a 1- specificita je na vodorovné ose v bodu 0. Nejhorší v bodě 1. Blízko u svislé osy je málo falešně pozitivních případů a specificita je vysoká.

Na svislé ose je zachycena míra skutečně pozitivních případů a senzitivita. Míra skutečně pozitivních tříd je vyjádřením množství pozitivních případů, které jsou jako pozitivní klasifikovány. Senzitivita neboli citlivost ukazuje schopnost klasifikátoru správně označit pozitivní třídy a vyjadřuje úspěšnost, s níž test zachytí pozitivní třídy. Nabývá hodnot 0 až 1 nebo 0% až 100%

$$\text{senzitivita} = \frac{\text{počet skutečně pozitivních}}{\text{počet skutečně pozitivních} + \text{počet falešně negativních}}$$

Obecně je na ROC grafu mnohem úspěšnější klasifikátor, jehož křivka se drží blíže svislé ose. Ideální ROC křivka nejdříve stoupá strmě vzhůru a vykazuje vysokou úspěšnost skutečně pozitivní klasifikace, poté se obvykle začne navyšovat počet falešně pozitivních tříd a křivka pokračuje do pravého horního rohu. Pokud se křivka vydá spíše ve směru úhlopříčky, znamená to, že každé zlepšení senzitivity zapříčiní stejně vysoké snížení specificity.

3.2 Ukazatele

Pro hodnocení ROC křivek se používá například ukazatel zvaný AUC - Area Under Curve (česky plocha pod křivkou). Hodnota AUC byla spočítána pomocí funkcí, které standardně nabízí Matlab. Dále byla použita přesnost predikce, která vyjadřuje, na kolik procent se shodují třídy klasifikované klasifikátorem s třídami původními. Jako třetí míru kvality klasifikace byl použit ukazatel OPTROCPT - Optimal operating point of the ROC

curve (česky optimální operační bod ROC křivky), který je spočítaný také pomocí funkce v Matlabu.

AREA UNDER CURVE

Pro každý klasifikátor je vykreslena jedna ROC křivka, ta jde vytvořit vždy jen pro jednu pozitivní třídu, proto se musí pro jeden data set vytvořit více křivek v závislosti na počtu tříd (počet tříd = počet ROC křivek, vždy jedna pozitivní proti zbytku negativních). Ukazatel AUC se pohybuje v intervalu od nuly do jedné. Jedna je nejlepší, výsledky, které spadají pod hranici 0,5, se dají považovat za čistě náhodné (klasifikátor určí třídu případu náhodně) a snižuje se tak jejich vypovídací hodnota. Ve většině data setů je rozložení tříd nerovnoměrné. Jen v data setu Fisher iris je rozdělení 50:50:50 (150 případů rozděleno stejnoměrně do tří tříd). Proto jsem použil dva upravené ukazatele AUC. A to *průměrný AUC* a *vážený AUC*.

Průměrný AUC je průměr ze všech tříd v data setu bez ohledu na počet případů v každé třídě. Vážený AUC pak tyto počty zohledňuje. Byl použit jednoduchý výpočet, kdy bylo jednotlivé skóre tříd vynásobeno počtem případů v dané třídě, poté byla tyto čísla sečtena a podělena celkovým počtem případů v data setu. Vážený AUC tak zohledňuje, kolik procent celého data setu tvoří daná třída a podle toho bere v úvahu její AUC.

PŘESNOST PREDIKCÍ

Jako další míru kvality klasifikace byla použita přesnost predikce, která je prezentována v procentech.. Přesnost vyjadřuje, na kolik procent se shodují původní třídy s třídami odhadnutými klasifikátorem. Tento ukazatel by se dalo také pojmenovat jako úspěšnost predikce.

OPTROCPT

Tento ukazatel se skládá ze dvou souřadnic, která je tvořena mírou falešné pozitivity a mírou skutečné pozitivity, například [0,3;0,79]. Tyto souřadnice určují polohu bodu, který vyjadřuje ideální poměr navyšování selhání klasifikace pozitivních tříd proti navyšování počtu falešných klasifikací tříd. Bod je tedy v takovém místě, ve kterém je velký počet pozitivních klasifikací, aniž by došlo k velkému nárůstu falešně pozitivních klasifikací tříd.

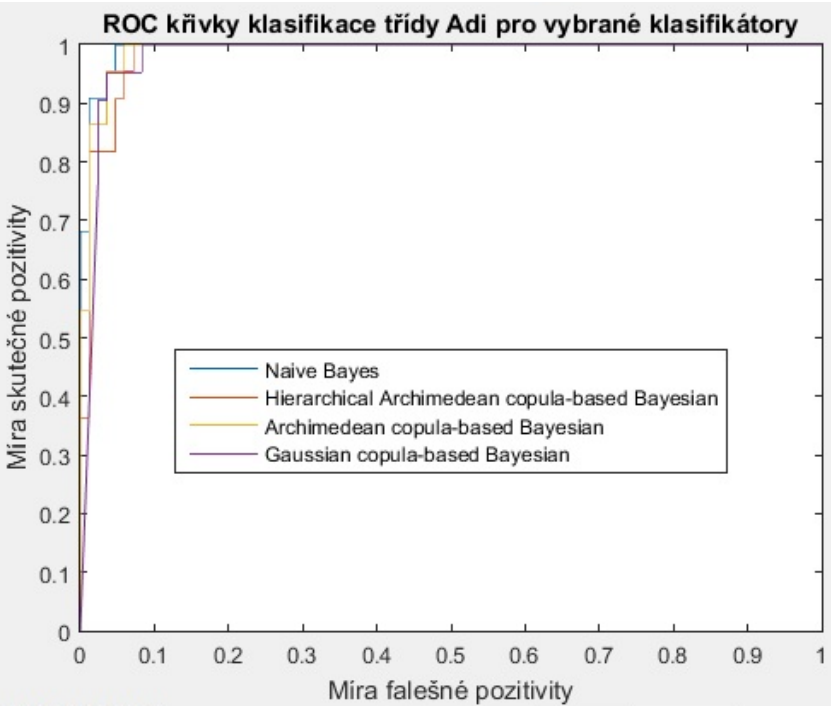
Klasifikátor je považován za kvalitnější čím je tento bod blíže k levému hornímu rohu grafu (0;1) (vyšší AUC a přesnější určování tříd). Tento ukazatel jde využít pro klasické ROC křivky.

4 Hodnocení kvality klasifikace na data setu Breast Tissue

Klasifikace byla prováděna v MATLABu, s využitím desetinásobné křížové validace (rozdělení data setu na testovací a trénovací vzorky). Výstupem jsou čtyři grafy. Pro každou třídu jeden, ve kterém je daná třída uvažována jako pozitivní. Graf vždy zobrazuje ROC křivky odpovídající jednotlivým klasifikátorům. Kvůli zachování přehlednosti při velkému množství klasifikátorů budou na grafu porovnány jen čtyři křivky klasifikátorů, které dosáhly nejvyšší kvality klasifikace. Zbylé ukazatele jako AUC a přesnost predikcí budou uváděny vždy pro všechny klasifikátory. Grafy s ukazatelem OPTROCPT nebyly do této práce zařazeny, jelikož zabírají mnoho prostoru.

Následující obrázek č 2 zobrazuje ROC křivky pro pozitivní třídu Adi pro vybrané klasifikátory na data setu Breast Tissue.

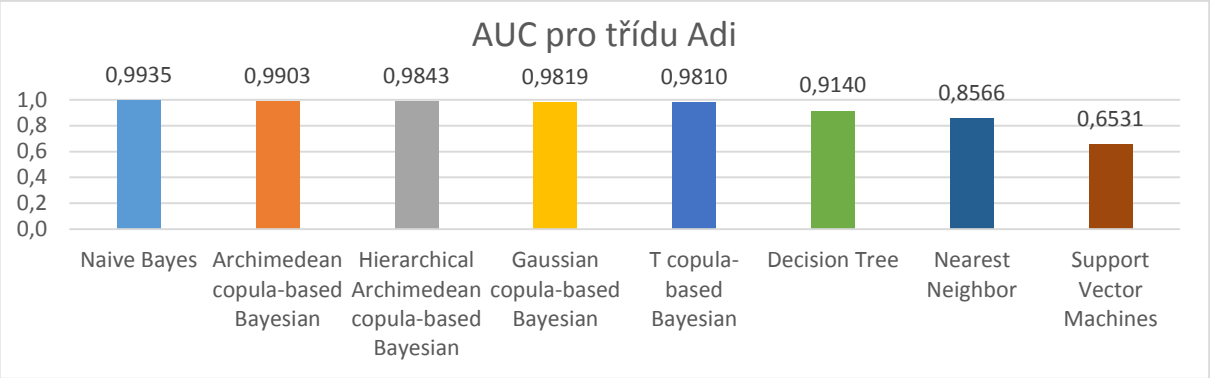
Obrázek 2:ROC křivky pro vybrané klasifikátory na třídě Adi



Zdroj: Vlastní zpracování

Jak lze vidět na grafu, výsledky čtyř nejúspěšnějších klasifikátorů jsou velmi těsné. Při klasifikaci první třídy byl nejvíce úspěšný Naive Bayes (jeho křivka se nejvíce přiblížila levé a horní hraně grafu) těsně následovaný nově vyvinutým Bayesovským klasifikátorem založeným na hierarchických archimedovských kopulích. Podrobnější výsledky v číselném vyjádření AUC jsou vyobrazeny na grafu v obrázku č 3.

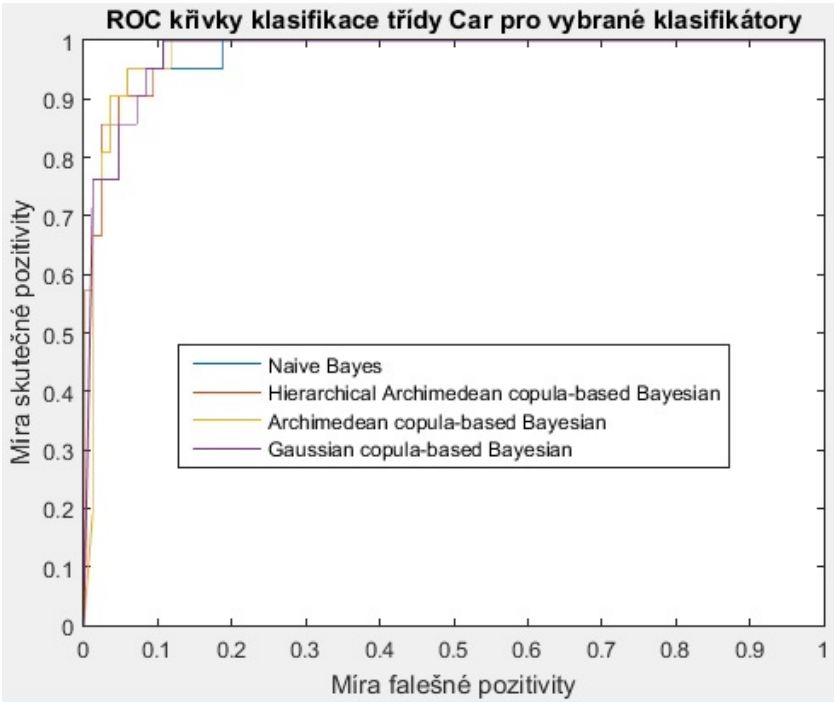
Obrázek 3:AUC pro třídu Adi



Zdroj: Vlastní zpracování

Obrázek č 4 zobrazuje ROC křivky pro pozitivní třídu Car pro vybrané klasifikátory na data setu Breast Tissue.

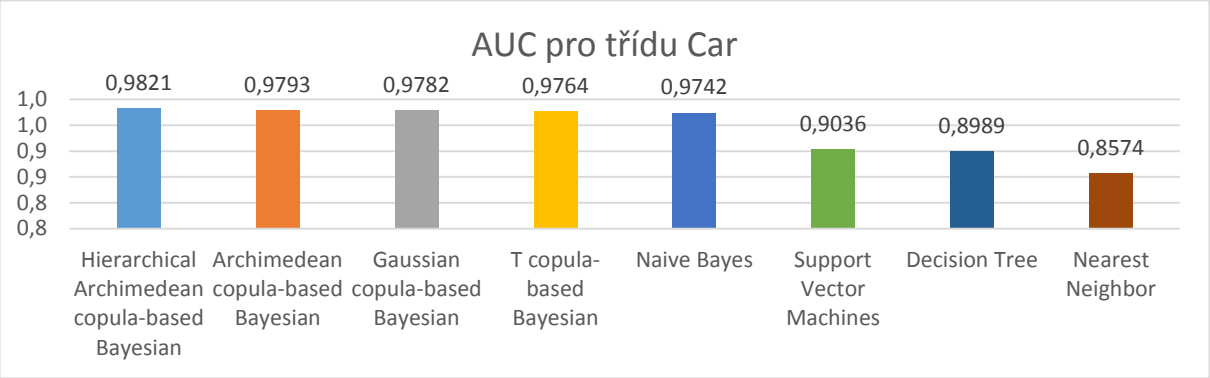
Obrázek 4:ROC křivky pro vybrané klasifikátory na třídě Car



Zdroj: Vlastní zpracování

Klasifikace dopadla opět velmi těsně, nejkvalitnější klasifikace dosáhly oba Bayesovské klasifikátory, založené na archimedovských kopulích. Detailnější výsledky jsou uvedeny v následujícím porovnání AUC v obrázku č 5.

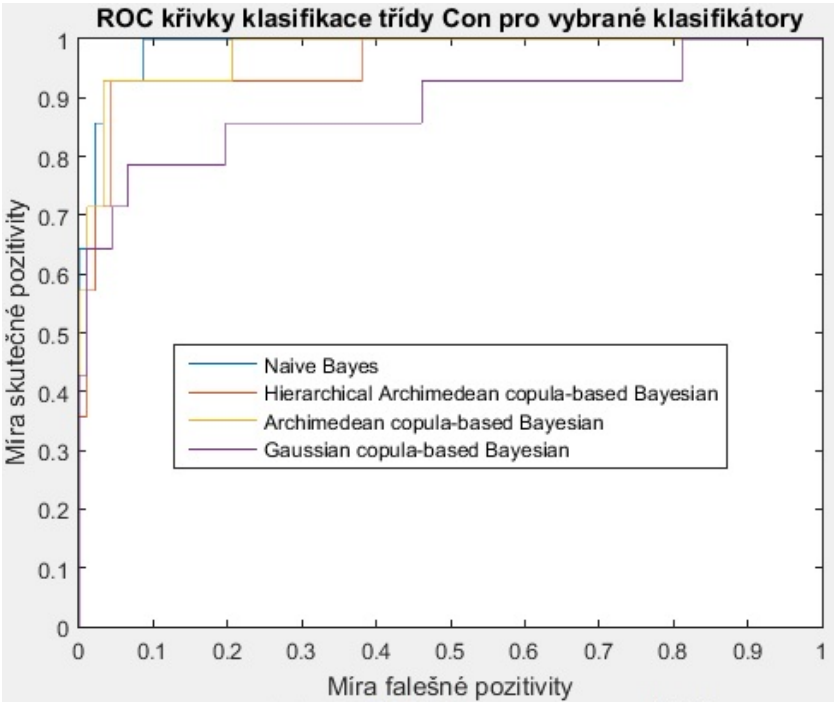
Obrázek 5:AUC pro třídu Car



Zdroj: Vlastní zpracování

Obrázek č 6 zobrazuje ROC křivky pro pozitivní třídu Con pro vybrané klasifikátory na data setu Breast Tissue.

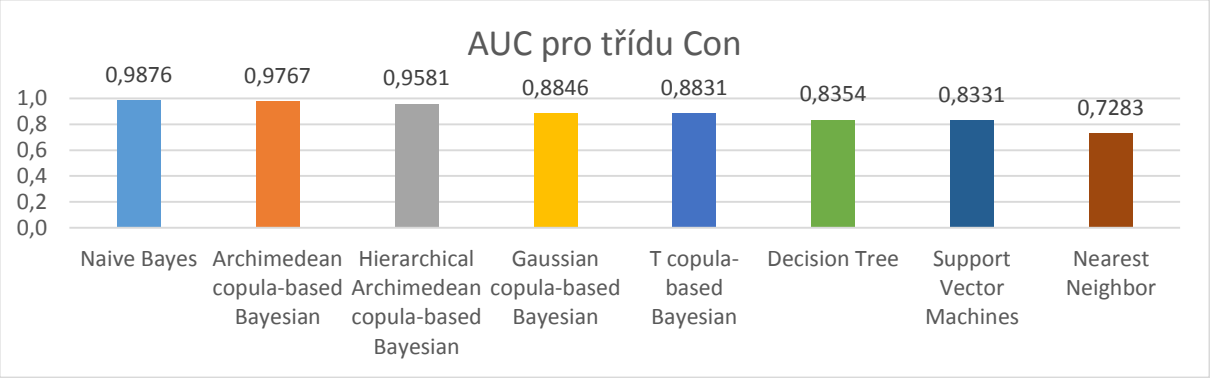
Obrázek 6:ROC křivky pro vybrané klasifikátory na třídě Con



Zdroj: Vlastní zpracování

Na grafu klasifikace pozitivní třídy Con jsou průběhy ROC křivek lépe patrné než na předchozích třídách. Klasifikátory zde nedosáhly tak těsných výsledků. První se umístil Naive Bayes, na druhém a třetím místě opět oba Bayesovské klasifikátory založené na archimedovských kopulích. Porovnání hodnot AUC jsou na obrázku č 7.

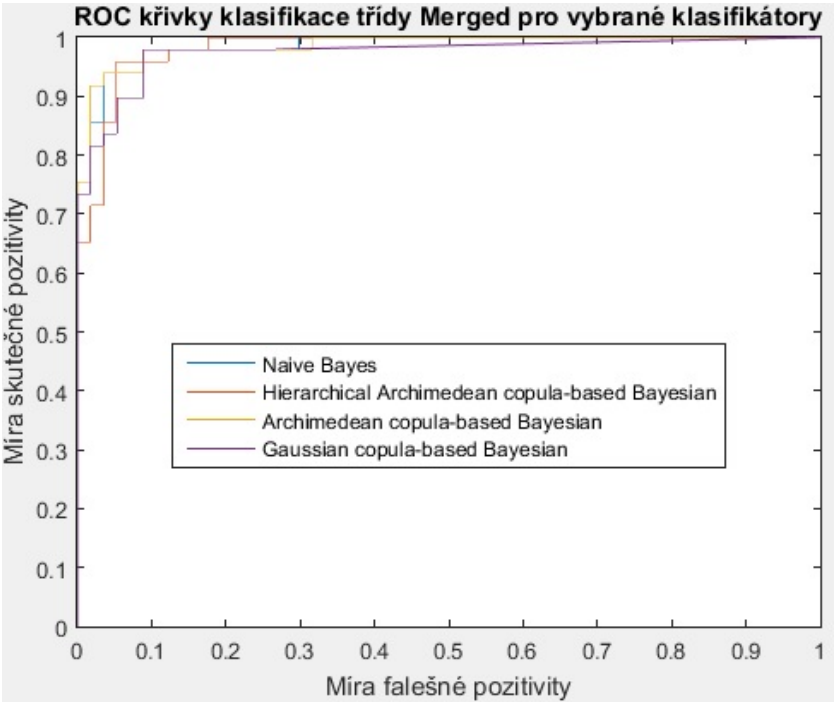
Obrázek 7:AUC pro třídu Con



Zdroj:Vlastní zpracování

Poslední klasifikovanou třídou je pozitivní třída Merged, graf ROC křivek zobrazující tuto klasifikaci je zachycen na obrázku č 8.

Obrázek 8:ROC křivky pro vybrané klasifikátory na třídě Merged

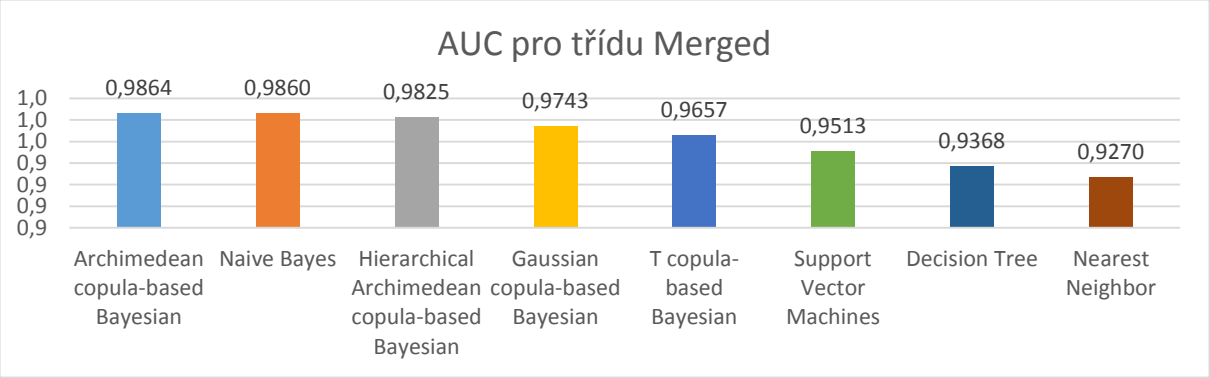


Zdroj: Vlastní zpracování

I na poslední třídě dopadla klasifikace velmi těsně. Vítězem na pozitivní třídě Merged se stal Bayesovský kalsifikátor založený na archimedovských kopulích těsně následovaný klasifikátorem Naive Bayes.

Obrázek číslo 9 zobrazuje graf porovnání hodnot AUC pro třídu Merged.

Obrázek 9:AUC pro třídu Merged

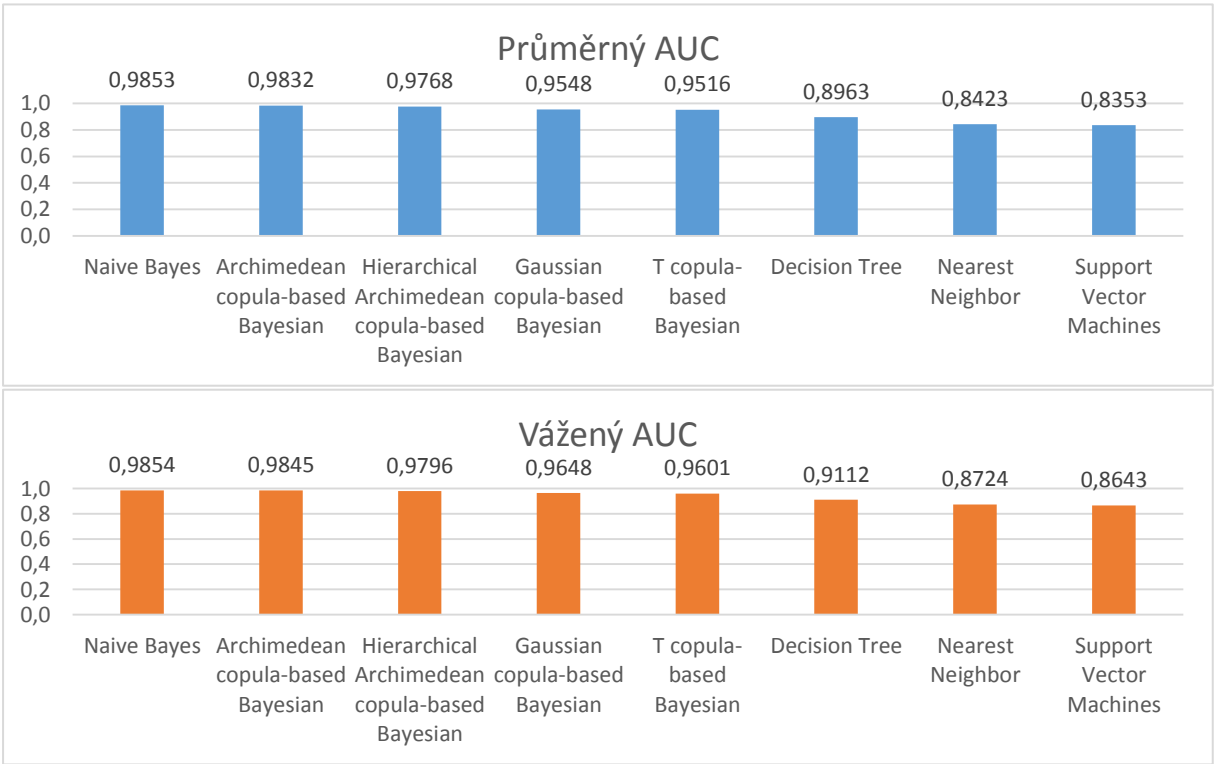


Zdroj: Vlastní zpracování

Nakonec byly porovnány klasifikace všech tříd dohromady v rámci celého data setu Breast tissues za pomocí průměrného a váženého AUC. Toto porovnání zobrazuje následující graf na obrázku č 10.

Obrázek 10:Porovnání průměrného a váženého AUC pro data set Breast tissue

Zdroj: Vlastní zpracování

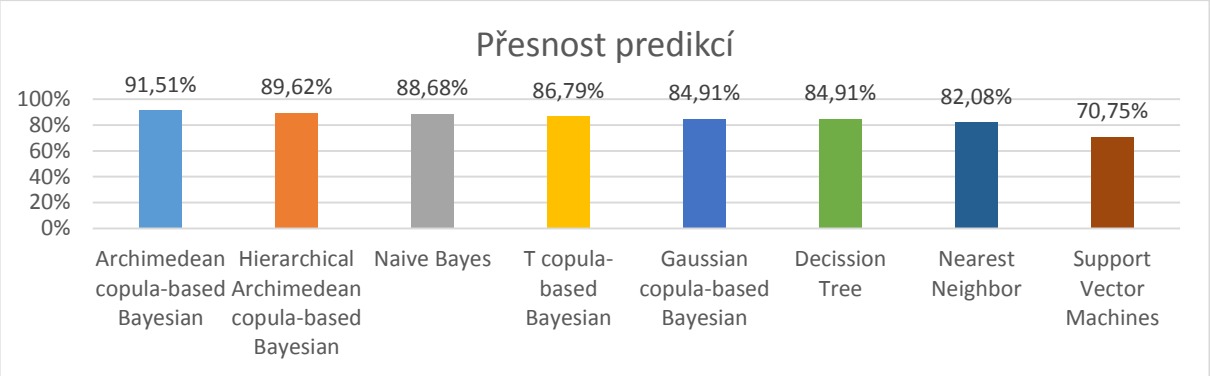


Hodnota váženého AUC je pro každý klasifikátor o něco vyšší než hodnota průměrného AUC. Je to proto, že většina klasifikátorů dosáhla vyšší kvality klasifikace na třídách, které měly vyšší počet případů, takže měly větší váhu.

V celkovém hodnocení na data setu Breast tissue se první umístil klasifikátor Naive Bayes, druhý s opravdu těsným rozdílem s hodnotou AUC menší jen o 0,0009 se umístil Bayesovský klasifikátor založený na archimedovských kopulích. Třetí nejlepší kvality klasifikace dosáhl Bayesovský klasifikátor založený na hierarchických archimedovských kopulích. Nejhůře dopadl překvapivě klasifikátor Support vector machines, který na ostatních data setech dopadl nejhůře na čtvrtém místě z osmi.

Následující obrázek č. 11 zobrazuje porovnání Přesnosti predikcí na datovém setu Breast tissue.

Obrázek 11:Porovnání přesnosti predikce na data setu Breast tissue



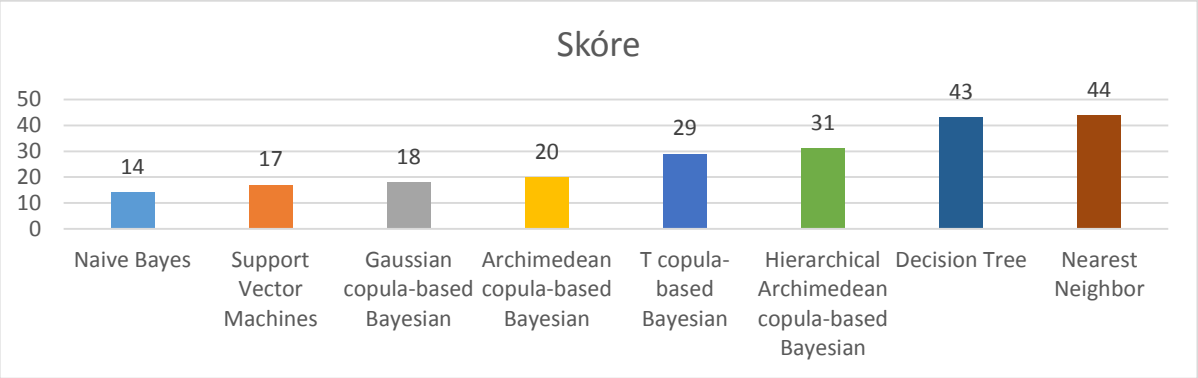
Zdroj: Vlastní zpracování

V přesnosti predikcí dopadly nejlépe oba Bayesovské klasifikátory založené na archimedovských kopulích. Třetí se umístil Naive Bayes. První dva klasifikátory dosáhly vyšší přesnosti i přesto, že je Naive Bayes překonal v hodnotě AUC.

5 Celkové shrnutí všech datových setů

Na základě váženého AUC byly klasifikátory ohodnoceny, podle pořadí na jednotlivých data setech, od prvního do osmého místa. Toto pořadí bylo poté sečteno a vyjadřuje skóre klasifikátoru kdy 8 je nejnižší (nejlepší) a 48 nejvyšší (nejhorší) možné skóre. Graf, který zobrazuje výsledky klasifikátorů podle skóre, je zachycen na obrázku č 12.

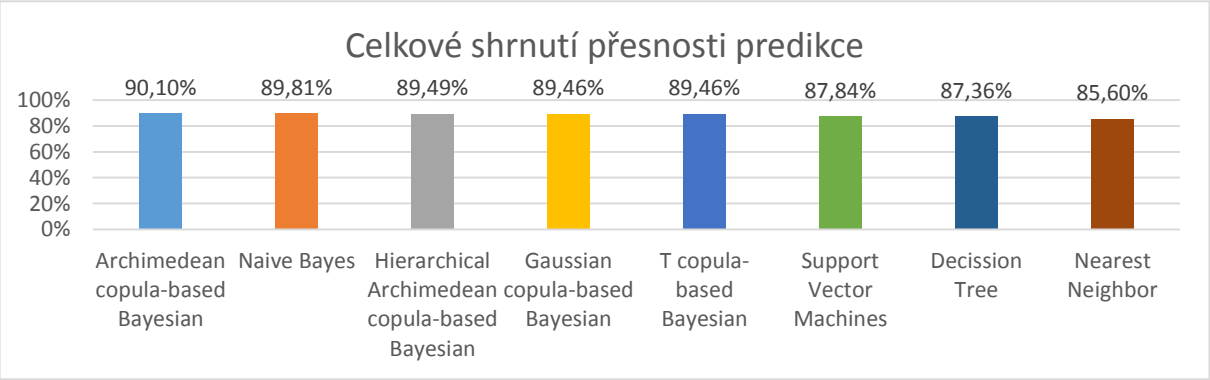
Obrázek 12:Celkové skóre klasifikátorů na všech data stech



Zdroj: Vlastní zpracování

Další srovnání pomocí celkové přesnosti predikce vzniklo zprůměrováním přesnosti predikce na všech datových setech jednotlivých klasifikátorů. Graf celkové přesnosti predikce zobrazuje obrázek č 13.

Obrázek 13:Celkové shrnutí přesnosti predikce na všech data setech



Zdroj: Vlastní zpracování

Jak lze vidět ze shrnutí přesnosti, celkové skóre nelze brát jako směrodatné. Klasifikátor založený na archimedovských kopulích se umístil hned na třech data setech jako druhý a v přesnosti překonal veškeré konkurenční klasifikační metody. I přesto je v celkovém hodnocení podle AUC až na čtvrtém místě. Je důležité brát v úvahu, že pro různé datové sety se hodí různé typy klasifikátorů v závislosti na počtu atributů a celkovém rozdělení tříd v datech. Jednotlivé výsledky klasifikátorů se od sebe liší jen o několik desetin nebo setin, proto je vždy nejdůležitější hodnotit kvalitu klasifikace jen na určitém datovém setu.

Závěr

V práci je porovnána kvalita klasifikace osmi klasifikátorů pomocí ROC křivek na vybraném datovém setu Breast tissue. Postupně byly klasifikátory srovnány pomocí plochy pod křivkou, přesnosti predikce a optimálního operačního bodu křivky.

Na vybraném datovém setu dosáhl největší plochy pod křivkou klasifikátor Naive Bayes, druhé nejvyšší hodnoty, jen s rozdílem necelé jedné tisíciny, dosáhl nově vyvinutý klasifikátor založený na archimedovských třídách kopulí následovaný klasifikátorem založeným na hierarchických archimedovských kopulích, který dosáhl třetí nejvyšší hodnoty.

Z hlediska přesnosti predikce překonaly na daném data setu nově vyvinuté klasifikátory veškerou konkurenci a umístily se na prvním a druhém místě. Klasifikátor založený na archimedovských kopulích překonal všechny konkurenční klasifikační metody dokonce i na celkovém srovnání přesnosti predikce na všech datových setech.

Cílem práce bylo porovnat a zhodnotit, zda jsou nově navržené klasifikátory schopny konkurovat těm nejlepším stávajícím klasifikačním metodám. Výsledky prokázaly, že nové metody vykazují srovnatelnou kvalitu klasifikace s nejlepšími dostupnými klasifikačními metodami a na některých datech je dokonce překonávají. Tyto metody se tak můžou stát alternativou pro stávající metody a etablovat se v oblasti klasifikačních metod.

Seznam použitých pramenů a literatury

[1] AGGARWAL, Ch. C., 2015. Data classification: algorithms and applications. Boca Raton, FL: CRC press, ISBN 978-1-4665-8674-1.

[2] BRAMER, M., 2013. Principles of data mining. 2nd ed. London: Springer, ISBN 9781447148845.

[3] ENTEZARI-Maleki R., 2009. Comparison of Classification Methods Based on the Type of Attributes and Sample Size. Department of Computer Engineering, Iran University of Science & Technology (IUST), Tehran, Iran.

[4] FREITAS A., 2002. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg, ISBN 9783662049235.

[5] HOLČÍK, J., 2012. Analýza a klasifikace dat. Vyd. 1. Brno: Akademické nakladatelství CERM, ISBN 978-80-7204-793-2.

[6] NISBET R., Elder J., 2009. Handbook of statistical analysis and data mining applications. Burlington, MA: Academic Press/Elsevier, ISBN 9780080912035.

[7] THORSTEN J., 2002. Learning to Classify Text Using Support Vector Machines. Boston, MA: Springer US, ISBN 9781461509073.