FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Databázové systémy a metody zprac.inf.2

KIV/DBM2

Porovnání úspěšnosti dle formy studia
Výtah

Pavel Třeštík
A23N0001P

12. prosince 2023

1 Téma a motivace

Vybraným tématem je „porovnání úspěšnosti dle formy studia“. Hlavním cílem je zjistit úspěšnost každé formy studia (prezenční, kombinované a distanční) a porovnat tyto úspěšnosti. Výsledkem tohoto porovnání by mělo být přibližně stejné procento úspěšnosti.

Termín „úspěšnost“ byl již použit a nejjednodušší statistickou metrikou pro úspěšnost by byl jednoduše podíl úspěšných absolventů ku všem studentům. Tato práce by ovšem neměla moc velký smysl, kdyby pro její dokončení stačilo udělat 3 podíly. Úspěšnost je zde pro nás tedy něco převážně neznámého a cílem práce je vybrat vhodné příznaky, které nám úspěšnost vytvoří. Tato úspěšnost bude založena převážně na známkách. Známky jsou prakticky určeny k tomu, aby ohodnocovaly studentovu úspěšnost. Čistě známka ovšem může být zkreslující údaj sám o sobě, protože předměty mohou mít různé kreditové ohodnocení (tedy váhu), statut (A/B/C) a další. Také může hrát roli např. na kolikátý pokus student známku dostal nebo jestli předmět opakuje.

Počítání úspěšnosti tímto způsobem by mohlo být možné analyticky či některým modelem strojového učení. V případě, že by úspěšnost závisela na pouze malém počtu vlastností, jejichž důležitost je jasně daná, tak by bylo možné úspěšnost pouze vypočítat a výsledkem práce by mohl být pouze SQL skript. Při větším počtu vlastností nebo problémy s určením jejich vah, bude nutné použít jiný přístup, kterým by nejspíše byl nějaký regresní algoritmus strojového učení. Důvodem proč by se mohlo jednat o regresní úlohu je, že práce má vyjádřit úspěšnost pro danou formu studia. Otázkou ale je, jak bude tato úspěšnost počítána.

2 Realizace

Prvním krokem k realizaci práce bylo vybrat vhodné data. Původní myšlenka byla vybrat co nejvíce sloupců a použitím nástroje **weka** postupně odebírat nepodstatné sloupce.

Výběr vhodných dat byl nakonec dělaný manuálně podle názvů sloupců a jejich popisků. Začalo se od tabulky **ZNAMKY**, protože ty jsou hlavním podkladem pro hodnocení úspěšnosti. Ke známkám se navíc přidaly některé sloupce z jiných tabulek, které by mohly mít dopad na předpověď úspěšnosti např. **STUPEN_PRED_VZDELANI**.

Protože vymyslet způsob počítání úspěšnosti bez použití strojového učení je téměř nemožné, tak byl zvolen přístup použití strojového učení. Vybírané sloupce proto byly voleny s ohledem na to, jak by se podle nich dal vytvořit model.

Pro vytvoření regresního modelu, který byl původní myšlenkou zadání by bylo potřeba mít nějaké reálné hodnoty, které úspěšnost určují. V datech takovýto údaj není, takže je potřeba údaj vytvořit a nebo změnit typ modelu.

Vzhledem k datům bylo rozhodnuto, že bude použit klasifikační model, protože s dostupnými daty lze poměrně snadno vytvořit klasifikační třídy. Jako třídy byla použita boolean flaga, určující že student má záznam v tabulce **ABN_ABSOLVENTI** a forma studia. Výsledkem je tedy 6 tříd z kombinace: 1 - absolvoval, 0 - neabsolvoval a P - prezenční, K - kombinovaný, D - distanční. Třída je tedy např. 1P, 0P, 1D,...

Vybraná data byla volena tak, aby se úspěšnost nevztahovala ke studentovi, ale spíše ke známkám. Následkem je, že za každého studenta může v datech být několik desítek záznamů. V době dokončení práce mají výsledná data `2722101` záznamů a z toho `461120` unikátních kombinací. Data povolují duplikáty, protože podle použitého modelu duplikáty mohou ovlivnit váhy modelu a tím model zpřesnit.

Kvůli datům bylo uděláno rozhodnutí, že model bude klasifikačním algoritmus. Jako hlavní model byl použit **Naivní Bayes**. Bayes v této práci porušuje podmínku, že použité atributy by měly být nezávislé, ale i při porušení této podmínky je Bayes stále velmi dobrý klasifikátor. Důvod proč byl primárně použit Bayes je rychlost učení, kdy vytvoření modelu a cross-validace dat s 10ti iteracemi zabere přibližně 1 minutu.

Pro ověření modelu byl pokus použít i **SMO**, což by měla být implementace SVM. Problém s použitím tohoto modelu byl velice dlouhý čas učení modelu, kdy nad všemi daty nestačilo ani půl dne. Model je možné vytvořit během několika minut, pokud je použit 50000 záznamů, což je pouze malý zlomek všech dat. Pro tento účel byl vyroben skript `50k_data_preparation.sql`. Bohužel ani s vytvořeným modelem se nepovedlo SMO použít, protože použit model pro plný dataset je také úloha na minimálně několik desítek hodin. SMO byl proto opuštěn.

3 Problémy

Při výběru dat bylo poměrně těžké vhodně volit data bez znalosti databáze. Některé sloupce a jejich popisky nejsou zrovna nápomocné. Příkladem je zjištění absolventů, kteří jsou použiti k vyrobení tříd v datech. Tabulka `STUDENTI` má sloupec `Absolvent` s popiskem „Absolvent“. Ze sloupce jde poznat, že se jedná o boolean hodnotu, ale už nejde poznat, jestli to značí úspěšného absolventa, studenta v posledním ročníku nebo co informace vlastně znamená. Jako indikace úspěšných absolventů byla použita existence záznamu v tabulce `ABN_ABSOLVENTI`.

Také bylo poměrně těžké z dat vybrat nějaký vhodný způsob, který by sloužil k učení modelu. Původní myšlenka zadaní byl regresní model, ale kvůli datum nešlo získat hodnoty pro učení regrese. Také se objevil nápad pouze shlukování, či kombinace shlukování k postavení regresního modelu. Tento nápad byl zavrhnut, protože realizace by byla složitější než se zdá. Protože už se nabízelo i shlukování, tak byl nakonec vybrán klasifikační model, pro který bylo možné vytvořit správná data a provést tak učení.

Problémem byl také dlouhý čas učení jiných modelů, než Naivní Bayes. Dle dostupných zdrojů je učení Bayese rychlejší oproti jiným modelům, ale SMO model se nepovedlo natrénovat ani po zhruba 20ti hodinách, což je několika 1000 násobek trénování Bayese. Resp. SMO se povedlo natrénovat v rámci minut na 50000 záznamech, což je zhruba 1.8% datasetu. Při zvýšení na 100000 záznamů už učení SMO trvalo několik hodin, po kterých bylo učení přerušeno.

4 Výsledky

Práce měla porovnat úspěšnost dle formy studia. Důležité je zmínit, že výsledky nejsou poměr absolventů a studentů, ale poměr záznamů reprezentující známku a dodatečné informace k ní. Pro porovnání hlavních výsledků lze použít statistický podíl, proti výsledkům modelu.

	Stat.	N. Bayes
prezenční	$\frac{335335}{2406529} = 13.93\%$	$\frac{143168}{2406529} = 5.95\%$
kombinované	$\frac{27727}{296209} = 9.36\%$	$\frac{16371}{296209} = 5.53\%$
distanční	$\frac{2708}{19363} = 13.99\%$	$\frac{1040}{19363} = 5.37\%$

Tabulka 1: Porovnání úspěšnosti modelu a statistiky

Z tabulky 1 vidíme, že výsledky modelu jsou značně horší, než reálné úspěšnosti. Důvodem nejspíš je, že model velmi často zařazuje výsledky do opačné třídy. To naznačuje, že model je pravděpodobně nedostatečně obsáhlý proto, aby dokázal správně predikovat úspěšnost. Na obrázku 1 lze vidět výsledek klasifikace tímto modelem.

```
=== Summary ===

Correctly Classified Instances     2263619               83.1571 %
Incorrectly Classified Instances    458482               16.8429 %
Kappa statistic                          0.584
Mean absolute error                      0.0666
Root mean squared error                  0.1899
Relative absolute error                 50.4718 %
Root relative squared error             73.9342 %
Total Number of Instances          2722101

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.914    0.005    0.956      0.914    0.934      0.928    0.999     0.989     0K
                 0.890    0.295    0.906      0.890    0.898      0.584    0.914     0.971     0P
                 0.427    0.096    0.386      0.427    0.405      0.317    0.855     0.362     1P
                 0.884    0.001    0.874      0.884    0.879      0.878    1.000     0.868     0D
                 0.590    0.008    0.418      0.590    0.489      0.491    0.990     0.396     1K
                 0.384    0.001    0.359      0.384    0.371      0.371    0.999     0.331     1D
Weighted Avg.    0.832    0.237    0.841      0.832    0.836      0.585    0.916     0.890

=== Confusion Matrix ===

      a       b       c       d       e       f   <-- classified as
 245444     132       0     107   22799       0 |       a = 0K
      0 1842874  227981     339       0       0 |       b = 0P
      0  192163  143168       4       0       0 |       c = 1P
     80       0       0   14722       0    1853 |       d = 0D
  11347       2       2       5   16371       0 |       e = 1K
      0       0       0    1668       0    1040 |       f = 1D
```

Obrázek 1: Výstup N. Bayes klasifikátoru

Dále se ukázalo že z vybraných 13ti atributů, jsou některé téměř ire-

levantní a proto mohly být z datasetu odstraněny s minimálním dopadem na výsledky. Tabulka 2 reprezentuje vybrané atributy. Odstraněné atributy jsou přeškrtnuty. Všechny odstraněné atributy jsou z tabulky **STUDENTI** a výsledky klasifikace bez těchto atributů jsou v některých případech i lehce lepší, než s nimi.

Sloupec	Alias
~~NOVE_PRIJATY~~	~~STD_NOVE_PRIJATY~~
~~STUPEN_PRED_VZDELANI~~	~~STD_STUPEN_PRED_VZDELANI~~
~~POCET_ZAPISOVYCH_PROPUSTEK~~	~~STD_POCET_PROPUSTEK~~
FORMA	SP_FORMA
FAKULTA_SP	SP_FAKULTA_PROGRAMU
~~ROK_PLATNOSTI~~	~~SVR_ROK~~
POC_KRED	ZN_KREDITY_ZA_PREDMET
STATUT	ZN_STATUT_PREDMETU
POKUS_CISLO	ZN_POKUS_CISLO
HODNIDNO_ZKZP	ZN_HODNOCENI
TYP_ZK	ZN_TYP_ZKOUSKY
PRAC_ZKR	ZN_PRACOVISTE_ZKRATKA
ZAPOCET_POKUS	ZN_ZAPOCET_POKUS

Tabulka 2: Použité vlastnosti

5 Závěr

Z výsledků porovnání vzniklého modelu ku statistickému podílu je vidět, že model je velmi nepřesný, až skoro nepoužitelný. Při inspekci použitých atributů bylo zjištěno, že některé atributy model téměř zhoršovaly. Je možné, že úspěšnost modelu by dosáhla lepších čísel s větším počtem atributů. Ale také je možné, že zkrátka nelze lépe předpovídat úspěšnost formy studia se známkami jako hlavní zdroj informací.