Vysoké učení technické
Fakulta informačních technologií

Získávání znalostí z databází

Databáze restaurací - řešení

Alena Tesařová (xtesar36)
Jan Šorm (xsormj00)

21. prosince 2020

1 Úvod

Cílem projektu bylo vytvoření formulace a vyřešení úlohy pro získávání znalostí nad vybraným vzorkem dat v prostředí RapidMiner. Pro analýzu jsme si vybrali data z databáze restaurací, která byla získána *Národním centrem pro výzkum a technologický rozvoj CENIDET* v Mexiku a jejich popisem se zabývá druhá kapitola s popisem dat.

Zabývali jsme se úlohou predikce spokojenosti zákazníka na základě informací o zákazníkovi a restauraci. Z těchto dat jsme hledali ty nejvhodnější, které vedou k největší přesnosti s odhadem spokojenosti. Samotnou úlohou se pak více zabývá kapitola 3.

2 Popis dat

Pro účely projektu byla vybrána data z databáze restaurací, která byla získána *Národním centrem pro výzkum a technologický rozvoj CENIDET* v Mexiku. Autory jsou *Rafael Ponce Medellín* a *Juan Gabriel González SernaTato*. Data byla použita v roce 2011 při výzkumu účinků příslušných kontextových prvků na výkon systému restaurací, kde bylo úkolem vygenerovat seznam top-restaurací podle zákazníkových preferencí. Zároveň článek zmiňuje i problém vybírání důležitých atributů pro relevantní určení seznamu. [1]

Databáze obsahuje celkem 9 souborů ve formátu CSV. Z nich jsou v 5 souborech uloženy informace o jednotlivých restauracích, které lze spojit pomocí id restaurace, jež je obsaženo v každém z těchto 5 souborů. Dále tyto soubory obsahují následující informace o restauracích:

1. typ kuchyně
2. otevírací doba (hodina a den)
3. podporovaný typ platby
4. typ parkovacího místa
5. zeměpisné souřadnice, název, adresa, město, stát, země, fax, PSČ, prodej alkoholu, kuřácká oblast, dress code, přístupnost, výšku cen, web, atmosféra, franšíza, typ prostor, další služby

V dalších 3 souborech jsou pak obsaženy informace pouze o zákazníkovi, které jsou tentokrát propojeny pomocí id zákazníka. Kromě id zákazníka pak obsahují soubory tyto informace:

1. preferovaná kuchyně
2. preferovaný typ platby
3. zeměpisné souřadnice bydliště, kuřák, úroveň pití, preferovaný styl oblékání, atmosféra, způsob dopravy, rodinný stav, potomci, rok narození, zájmy, osobnost, náboženství, aktivity, oblíbená barva, hmotnost, rozpočet, výška

V posledním souboru jsou pak uloženy informace o hodnocení zákazníka dané restaurace. Kromě id restaurace a id zákazníka jsou zde atributy:

1. celkové hodnocení, hodnocení jídla a hodnocení služeb

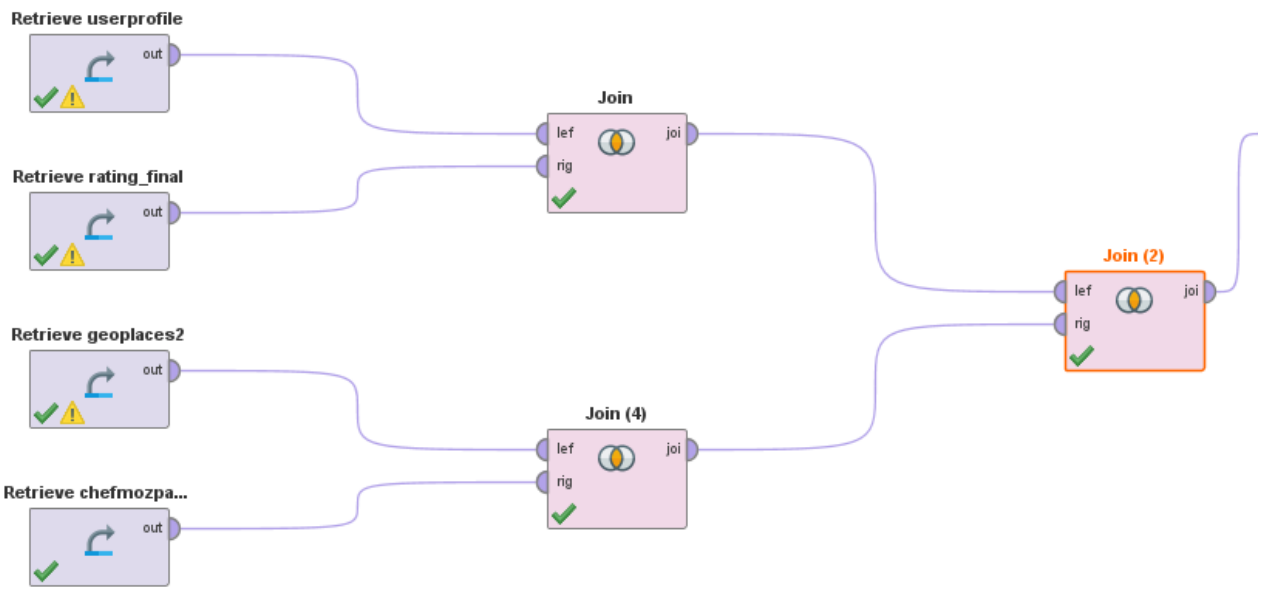
3 Úloha – predikce spokojenosti zákazníka

Jako úlohu projektu jsme si zvolili predikci spokojenosti zákazníka s restaurací na základě informací o restauraci a zákazníkovi. Spokojenost zákazníka bude vycházet z minulých hodnocení restaurací klienty na základě informací od 130 restaurací ležící v Mexiku a 1161 referencí (známkování) na restaurace od zhruba 138 lidí, jelikož taková jsou naše získaná data.

Nejlepší známka odpovídá hodnotě dvě a nejhorší nule, celkem tedy budeme pracovat se třemi třídami hodnocení. Nejprve se v kapitole 3.1 zvolíme základní set atributů pro predikci, který budeme v kapitole 3.2 zvětšovat (případně zmenšovat) na základě výsledků dolovacích metod. V závěru pak zhodnotíme metody a vybereme tu s nejlepšími výsledky. Tato predikce může do budoucna pomoci zákazníkovi ke snadnějšímu vyhledávání restaurace v Mexiku, která by měla nejvíce uspokojit jeho potřeby.

3.1 Příprava dat a výběr atributů

V přípravě je třeba řádně prozkoumat, jaké hodnoty obsahují dodané CSV soubory. Jelikož chceme predikovat spokojenost (tedy údaj v tabulce rating_final), tedy použijeme tento soubor jako základ. Dále nás zajímají informace o uživateli (*userprofile.csv*) a dané restauraci (*geoplaces2.csv*). Zbylo nám 6 nevyužitých souborů, ze kterých si vybereme informaci o možnostech parkování (*chefmozparking.csv*) u restaurace, jelikož z naší osobní zkušenosti se jedná o důležitý faktor při výběru místa. Schéma zapojení je zobrazeno na obrázku 1.



Obrázek 1: Schéma zapojení v RapidMiner

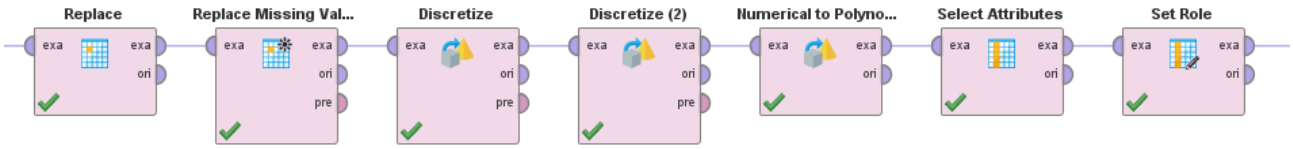
Nyní máme na výběr celkem 42 atributů, ale je jasné, že všechny se nebudou hodit na predikci spokojenosti. Pro první odhad důležitých atributů jsme použili nástroj **Auto Model**, který zobrazí korelaci, stabilitu a další informace u atributů. Dále pak umí vyzkoušet různé dolovací metody a zobrazit jejich úspěšnost. Z vytvořených modelů jednotlivých metod jsme zkoumali, které atributy nejvíce ovlivňují výsledek. Na základě této analýzy z *Auto modelu* jsme vybrali základní set atributů.

1. birth_year – rok narození
2. smoker – jestli daná osoba kouří
3. drink_level – vztah k alkoholu

4. interest – zájmy
5. dress_preference – styl oblékání
6. color – oblíbená barva
7. budget – rozpočet
8. state – stát
9. weight – váha
10. parking_lot – druh parkování
11. religion – náboženství
12. smoking area – místnost vyhrazená pro kuřáky

Po výběru atributů je třeba udělat přípravu dat, jelikož některá data můžou chybět, některé je třeba diskretizovat nebo se zde můžou vyskytovat odlehlé hodnoty. Celý proces přípravy dat lze vidět na obrázku 3.

Replace slouží k nahrazení otazníků (reprezentující chybějící hodnoty v našem datasetu) za prázdné pole. V dalším kroku potom můžeme chybějící hodnoty (pomocí *Replace missing values*) nahradit průměrnou hodnotou. Jelikož váha a datum narození mají velmi rozptýlené hodnoty, rozhodli jsme se tyto dva atributy diskretizovat – rok narození do 2 skupin (mladší do ročníku 1980 a starší), váha rozdělujeme do 3 skupin (hubení do 66 kg, normální do 91 kg a tlustí tvoří zbytek). Numerical to Polynominal děláme kvůli atributu *rating*, který je možné rozdělit do 3 tříd. Nakonec byl tento atribut označen jako *label* pomocí *Set Role*, jelikož ten chceme predikovat.



Obrázek 2: Příprava dat v Rapidminer

3.2 Metody pro dolování dat

Pro naši úlohu jsme vybrali metody pro klasifikaci do 3 tříd – 0, 1, 2, kde hodnota 2 značí největší spokojenost a naopak hodnota 0 největší nespokojenost. Pro odhad výkonnosti modelů jsme nejprve použili *Cross validation* operátor, který rozdělí dataset na k foldů stejné velikosti, ze kterých je jeden použit na testování a zbytek k trénování modelu. Bohužel se tento operátor neosvědčil, jelikož nemáme takové množství testovacích data, proto jsme se místo toho rozhodli použít operátor *Split validation* s poměrem trénovacích a testovacích dat 9 : 1, který náhodně rozdělí datovou sadu. Subrocesy split validace jsou zobrazeny na obrázku 3. Pro testování dat byl model aplikovaný pomocí *Apply model* a přesnost metody byla naměřená pomocí operátoru *Performace*. V trénovací části jsme obměňovali různé algoritmy a jejich zhodnocení bude popsáno v následujících sekcích.

Testovali jsme na následujících metodách:

1. Naive Bayes,
2. Deep learning,
3. Decision tree,

4. Random forest,

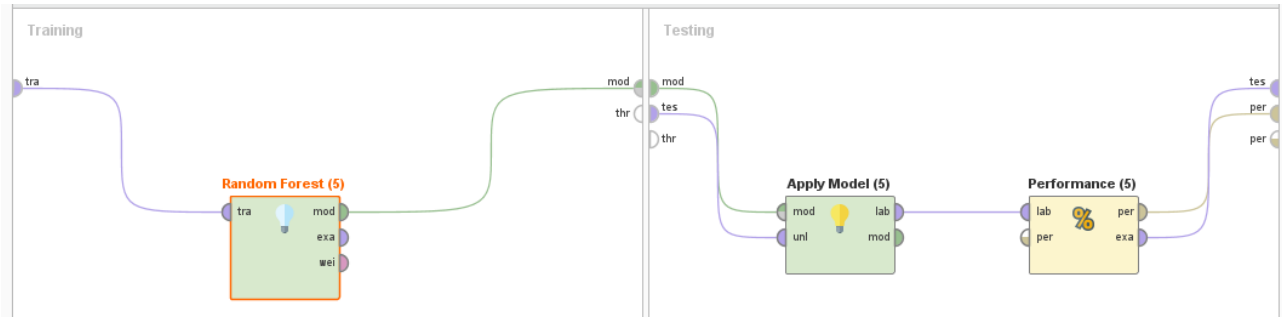
5. Gradient boosted tree,

6. k-Nearest Neighbor (k-NN).

Souhrnné výsledky pro jednotlivé metody lze vidět v tabulce 1. Bližší informace k jednotlivým sadám atributů jsou uvedeny v dalších sekcích.

Metoda	Základní dataset	Optimalizace
Naive Bayes	54,31	65,52
Deep learning	55,17	62,93
Decision tree	58,62	68,97
Random forest	56,33	71,55
Gradient boosted tree	55,17	70,69
k-NN	56,03	68,10

Tabulka 1: Souhrnné výsledky



Obrázek 3: Split validation subprocesy

Naive Bayes

Pro základní dataset *weight, birth_year, smoker, dress_preference, interest, religion, color, state, parking_lot, smoking_area, budget, drink_level* bylo dosaženo přesnosti 54,31 %, což lze pozorovat na obrázku 4.

K dosažení lepší přesnosti byl pak použit operátor *Optimize selection (Evolutionary)*, který vybral soubor nejlepších atributů, kterými jsou *weight, birth_year, smoker, dress_preference, interest, religion, color, state*, se kterými bylo dosaženo úspěšnosti 65,52 % (viz obrázek 5).

accuracy: 54.31%

	true 2	true 1	true 0	class precision
pred. 2	33	16	8	57.89%
pred. 1	13	20	7	50.00%
pred. 0	3	6	10	52.63%
class recall	67.35%	47.62%	40.00%	

Obrázek 4: Přesnost Naive Bayes se zvolením základního datasetu

accuracy: 65.52%

	true 2	true 1	true 0	class precision
pred. 2	40	14	9	63.49%
pred. 1	8	24	4	66.67%
pred. 0	1	4	12	70.59%
class recall	81.63%	57.14%	48.00%	

Obrázek 5: Přesnost Naive Bayes po optimalizaci atributů

Deep learning

Pro metodu *Deep learning* dosáhl základní dataset přesnosti 55,17 %, tedy o něco větší než pro metodu *Naive Bayes* (viz obrázek 6). Po použití operátoru *Optimize selection (Evolutionary)* byly vybrány atributy *weight, birth_year, smoker, dress_preference, interest, religion, color, state*, se kterými se úspěšnost zlepšila na 62,93 % (viz obrázek 7).

accuracy: 55.17%

	true 2	true 1	true 0	class precision
pred. 2	44	27	14	51.76%
pred. 1	5	14	5	58.33%
pred. 0	0	1	6	85.71%
class recall	89.80%	33.33%	24.00%	

Obrázek 6: Přesnost Deep learning se zvolením základního datasetu

accuracy: 62.93%

	true 2	true 1	true 0	class precision
pred. 2	40	19	11	57.14%
pred. 1	7	20	1	71.43%
pred. 0	2	3	13	72.22%
class recall	81.63%	47.62%	52.00%	

Obrázek 7: Přesnost Deep learning po optimalizaci atributů

Decision tree

Metoda *Decision tree* dosáhla se základním datasetem nejlepší přesnosti ze všech použitých metod a to úspěšnosti 58,62 %, což lze vidět na obrázku 8. Po použití operátoru *Optimize selection (Evolutionary)* se přesnost zlepšila, ale už nebyla nejlepší ze všech metod. Tentokrát byly vybrány atributy *smoker, drink_level, dress_preference, religion, color, budget, smoking_area* a úspěšnost byla 68,97 % (viz obrázek 9).

accuracy: 58.62%

	true 2	true 1	true 0	class precision
pred. 2	39	23	8	55.71%
pred. 1	8	17	5	56.67%
pred. 0	2	2	12	75.00%
class recall	79.59%	40.48%	48.00%	

Obrázek 8: Přesnost Decision tree se zvolením základního datasetu

accuracy: 68.97%

	true 2	true 1	true 0	class precision
pred. 2	42	12	8	67.74%
pred. 1	6	25	4	71.43%
pred. 0	1	5	13	68.42%
class recall	85.71%	59.52%	52.00%	

Obrázek 9: Přesnost Decision tree po optimalizaci atributů

Random forest

Základní dataset dosáhl přesnosti $56,33\%$, jak lze vidět na obrázku 10. K dosažení lepší přesnosti jsme použili operátor *Optimize selection (Evolutionary)*, který nám vybral soubor nejlepších atributů, kterými jsou *weight, drink_level, dress_preference, interest, religion, color, budget*. V tomto zastoupení atributů metoda dosáhla úspěšnosti $71,55\%$ (viz obrázek 11).

accuracy: 56.33% +/- 3.72% (micro average: 56.33%)

	true 1	true 2	true 0	class precision
pred. 1	214	136	63	51.82%
pred. 2	161	311	62	58.24%
pred. 0	46	39	129	60.28%
class recall	50.83%	63.99%	50.79%	

Obrázek 10: Přesnost Random forest se zvolením základního datasetu

accuracy: 71.55%

	true 1	true 2	true 0	class precision
pred. 1	26	10	1	70.27%
pred. 2	14	38	5	66.67%
pred. 0	2	1	19	86.36%
class recall	61.90%	77.55%	76.00%	

Obrázek 11: Přesnost Random forest po optimalizaci atributů

Gradient boosted tree

Základní dataset dosáhl přesnosti u této metody $55,17\%$ (obrázek 12).

Po zkoušení různých atributů, dosáhla nejlepších výsledků u této metody následující sada: *birth_year, hijos, interest, color, parking_lot* s přesností 70.69%, jak lze vidět na obrázku 13.

accuracy: 55.17%

	true 1	true 2	true 0	class precision
pred. 1	19	10	8	51.35%
pred. 2	19	36	8	57.14%
pred. 0	4	3	9	56.25%
class recall	45.24%	73.47%	36.00%	

Obrázek 12: Přesnost Gradient boosted tree se základním datasetem

accuracy: 70.69%

	true 1	true 2	true 0	class precision
pred. 1	26	3	6	74.29%
pred. 2	14	46	9	66.67%
pred. 0	2	0	10	83.33%
class recall	61.90%	93.88%	40.00%	

Obrázek 13: Přesnost Gradient boosted tree po optimalizaci atributů

k-NN

Při použití referenčních atributů byla získána přesnost 56,03% (viz obrázek 14). Po optimalizaci atributů jsme získali přesnost 68,10 %, kde byly algoritmem vybrány tyto atributy: *weight, smoker, drink_level, dress_preference, hijos, interest, color, budget, parking_lot.*

accuracy: 56.03%

	true 1	true 2	true 0	class precision
pred. 1	24	18	8	48.00%
pred. 2	14	25	1	62.50%
pred. 0	4	6	16	61.54%
class recall	57.14%	51.02%	64.00%	

Obrázek 14: Přesnost k-NN se základním datasetem

accuracy: 56.03%

	true 1	true 2	true 0	class precision
pred. 1	24	18	8	48.00%
pred. 2	14	25	1	62.50%
pred. 0	4	6	16	61.54%
class recall	57.14%	51.02%	64.00%	

Obrázek 15: Přesnost k-NN po optimalizaci atributů

4 Závěrečné zhodnocení

Ve zkoumání různými metod jsme si ověřili, že výběr nejvíce korelovaných atributů má skutečně vliv na přesnost metody. Náš dataset se ukázal být celkem úspěšný – nejvyšší přesnosti dosáhl Decision Tree (s 58,62 %). Pro získání lepších výsledků jsme použili operátor optimalizace výběru atributů, se kterým se nám podařilo dostat nejlepší výsledky pro následující sadu atributů: *weight, drink_level, dress_preference, interest, religion, color, budget* – algoritmus Random forest (71,55 %).

Na závěr uvádíme i celkové skóre atributů, kde na prvních místech jsou ty, které podle genetického algoritmu nejvíce ovlivňují spokojenost zákazníka.

1. color (6x)
2. dress_preference (5x)
3. interest (5x)
4. religion (5x)
5. weight (4x)
6. birth_year (4x)
7. smoker (4x)

8. drink_level (3x)
9. state (3x)
10. budget (2x)
11. hijos (2x)
12. parking_lot (1x)
13. smoking_area (1x)

Nejpoužívanějším byl tedy atribut *color*, který byl použit ve všech metodách. Atributy *dress_preference, interest, religion* pak byly použity ve všech kromě jedné metody.

Reference

[1] *Blanca Vargas-Govea, Juan Gabriel González-Serna, Rafael Ponce-Medellín. Effects of relevant contextual features in the performance of a restaurant recommender system. In RecSys'11: Workshop on Context Aware Recommender Systems (CARS-2011), Chicago, IL, USA, October 23, 2011, Online:* `http://ceur-ws.org/Vol-791/paper8.pdf`