به نام پروردگار علم و دانش

دانشگاه تهران
(پردیس دانشکده های فنی)

درس داده کاوی

استاد:
دکتر علی فهیم

تهیه کننده :
مینو احمدی ۸۱۰۸۹۷۰۳۲

عنوان: ارائه یک مدل توصیه گر فیلم با استفاده از تجزیه و تحلیل سلایق

فهرست

۱- مقدمه

امروزه یکی از موضوعات مهم کاربردی در فروش آنلاین، توصیه ی صحیح و نمایش موارد مطلوب به خریداران بالقوه است. در این پژوهش با استفاده از تجزیه و تحلیل سلایق[1] که نشان می دهد مشاهدات چه زمانی و چگونه بصورت متناوب باهم رخ می دهند. در واقع هدف ما این است که با توجه متغیرهای در دست بتوانیم علایق و گرایش مردم به فیلم های سینمایی را با توجه به موضوعات آن ها پیشبینی کنیم. مفاهیم اصلی این مقاله به شرح زیر است:

- تجزیه و تحلیل سلایق برای توصیه کردن یک محصول
- استخراج ارتباطات و قواعد انجمنی با استفاده از الگوریتم Apriori

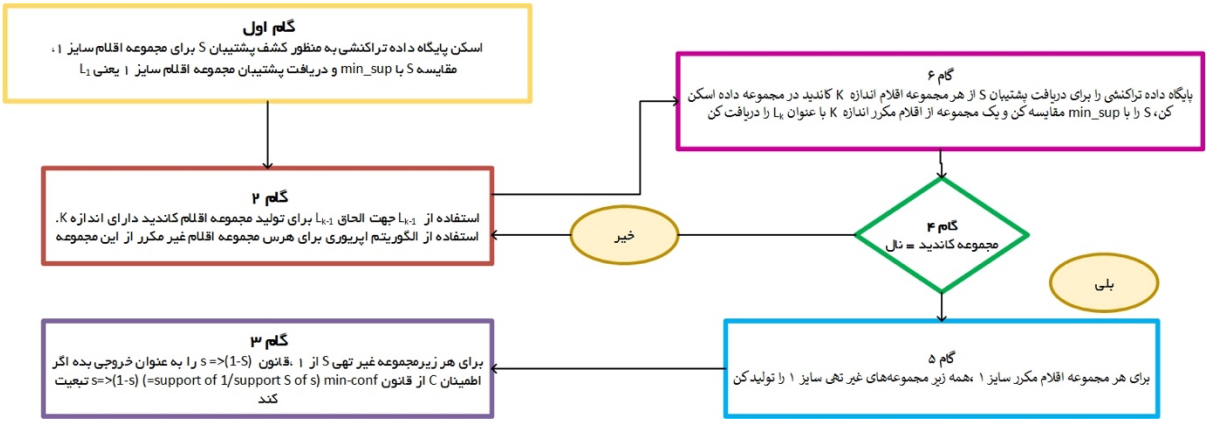
۲- آنالیز سلایق

در تجزیه و تحلیل سلایق هدف ما پیشبینی زمان تکرار وقایع بصورت مشابه است. ماهیت این آنالیز اکتشافی بوده و فراتر از دسته بندی و پیشبینی برچسب داده ها است. بصورت کلاسیک برای این آنالیز از الگوریتم Apriori استفاده می شود. این الگوریتم با تعیین یک حد آستانه، مشکل نمایی بودن تعداد مجموعه هایی از آیتم هایی که باهم تکرار شدند را حل میکند. این الگوریتم میتواند آیتم ست های مکرر (frequent itemset) را استخراج کند. سپس قواعدی که از confidence level معینی بالاتر باشند را انتخاب میکنیم. تعیین این حد آستانه بسیار مهم است چراکه اگر میزان آن بیش از اندازه پایین در نظر گرفته شوند منجر به قواعدی با ساپورت بالا اما صحت کم میشود. در مقابل، بسیار بالا در نظر گرفتن این مقدار نیز که تعداد قواعد خیلی کمی استخراج شوند.

۳- بررسی الگوریتم Apriori

الگوریتم اپریوری(Apriori) ، روشی قابل اعمال روی رکوردهای پایگاه داده و به ویژه پایگاه داده تراکنشی یا رکوردهای حاوی تعداد مشخصی فیلد یا آیتم است. اپریوری یکی از الگوریتم‌های دارای رویکرد «پایین به بالا» است که به تدریج رکوردهای پیچیده را با یکدیگر مقایسه می‌کنند. این الگوریتم یکی از روش‌های کارآمد برای حل مسائل پیچیده کنونی موجود در داده‌کاوی و یادگیری ماشین است. اساسا، الگوریتم اپریوری بخش‌هایی از یک پایگاه داده بزرگ‌تر را دریافت کرده و به آن‌ها «امتیازدهی» کرده و یا آن بخش‌ها را با دیگر مجموعه‌ها به شیوه مرتب شده‌ای مقایسه می‌کند. از نتایج خروجی، برای تولید مجموعه‌هایی استفاده می‌شود که مکررا در پایگاه داده اصلی به وقوع پیوسته‌اند.

فلوچارت الگوریتم اپریوری (Apriori) در ادامه آورده شده است.



شکل۱. فلوچارت الگوریتم Apriori

[1] affinity analysis

۳-۱. مزایا و معایب روش Apriori

این روش دارای مزایا و معایبی است که در ادامه به برخی از آن‌ها اشاره شده است.

مزایا:

1. مصرف کمتر حافظه
2. پیاده‌سازی آسان
3. استفاده از برخی ویژگی‌ها برای هرس کردن که موجب می‌شود مجموعه اقلام باقیمانده برای بررسی نهایی کمتر شوند.

معایب:

1. نیازمند اسکن‌های زیاد از پایگاه داده است.
2. تنها یک آستانه پشتیبان حداقلی منفرد را می‌پذیرد.
3. فقط برای پایگاه داده‌های کوچک مطلوب است.

۴- بکارگیری الگوریتم Apriori

قبل از بکارگیری الگوریتم دیتاست مناسب را آماده کرده و یک متغیر بانری نشان دهنده ی علاقه مندی به فیلم به دیتاست اضافه کردیم به این صورت که فیلم های با امتیاز بالای ۳ در این ستون برچسب True و فیلم های با امتیازات پایینتر برچسب Falseگرفتند. برای اجرای الگوریتم ابتدا باید مجموعه های مکرر را پیدا کنیم. برای این منظور حداقل پشتسیبانی را (support) به مقدار ۵۰ نتظیم میکنیم. سپس قواعد انجمنی را استخراج میکنیم و در پایان ان ها را ارزیابی میکنیم برای هر itemset میتوانیم با تنظیم هر فیلم به عنوان نتیجه و فیلم های باقی مانده تعدادی قواعد ارتباطی ایجاد کنیم در مرحله بعد confidenceهر یک از این قواعد را محاسبه میکنیم.

۵- استخراج قواعد انجمنی

بعد از ان که الگوریتم به طور کامل اجرا شد ما یک لیستی از مجومه های مکرر خواهیم داشت که به راحتی قابل تبدیل به قواعد انجمنی هستند یک مجموعه ی مکررشامل مجموعه ای از ایتم هایی با حداقل supportاست در حالی که یک قاعده انجمنی یک فرضیه و نتیجه دارد حال ما با در نظر گرفتن یکی از فیلم های موجود در itemsetبه عنوان نتیجه و مابقی فیلم ها به عنوان فرضیه میتوانیم قواعد را استخراج کنیم به بیانی دیگر اگر مخاطب تمام فیلم هایی که فرضیه در نظر گرفته شده اند را دوست داشته باشند سیستم ما فیلمی که به عنوان نتیجه تعیین شده بود را به او پیشنهاد میکند پس برای هر itemset ما میتوانیم با در نظر گرفتن هر یک از فیلم ها به عنوان نتیجه و مابقی به عنوان فرض قواعد انجمنی متعدد و متفاوتی را تولید کنیم.

۶- ارزیابی قواعد استخراج شده

در نهایت برای ارزیابی قواعد انحمنی بدست امده از همان روش طبقه بندی استفاده میکنیم به این صورت که داده ها را به دو دسته اموزشی و ازمون تقسیم میکنیم. confidenceقواعد را در data setازمون را محاسبه و با datasetاموزش مقایسه میکنیم. به این منظور ۲۰۰ شناسه کاربری بعنوان دیتاست آموزش و مابقی برای تست یا آزمون درنظر گرفته میشود.

۷- نتیجه‌گیري

در نهایت قوانینی که بهترین عملکرد را در دیتاست تست نیز داشتند به شرح زیر بدست آمد.

```
Rule #1
Rule: If a person recommends Silence of the Lambs, The (1991), Return
of the Jedi (1983) they will also recommend Star Wars (1977)
 - Train Confidence: 1.000
 - Test Confidence: 0.936

Rule #2
```

Rule: If a person recommends Empire Strikes Back, The (1980), Fugitive, The (1993) they will also recommend Raiders of the Lost Ark (1981)
 - Train Confidence: 1.000
 - Test Confidence: 0.876

Rule #3
Rule: If a person recommends Contact (1997), Empire Strikes Back, The (1980) they will also recommend Raiders of the Lost Ark (1981)
 - Train Confidence: 1.000
 - Test Confidence: 0.841

Rule #4
Rule: If a person recommends Toy Story (1995), Return of the Jedi (1983), Twelve Monkeys (1995) they will also recommend Star Wars (1977)
 - Train Confidence: 1.000
 - Test Confidence: 0.932

Rule #5
Rule: If a person recommends Toy Story (1995), Empire Strikes Back, The (1980), Twelve Monkeys (1995) they will also recommend Raiders of the Lost Ark (1981)
 - Train Confidence: 1.000
 - Test Confidence: 0.903

Rule #6
Rule: If a person recommends Pulp Fiction (1994), Toy Story (1995), Star Wars (1977) they will also recommend Raiders of the Lost Ark (1981)
 - Train Confidence: 1.000
 - Test Confidence: 0.816

Rule #7
Rule: If a person recommends Pulp Fiction (1994), Toy Story (1995), Return of the Jedi (1983) they will also recommend Star Wars (1977)
 - Train Confidence: 1.000
 - Test Confidence: 0.970

Rule #8
Rule: If a person recommends Toy Story (1995), Silence of the Lambs, The (1991), Return of the Jedi (1983) they will also recommend Star Wars (1977)
 - Train Confidence: 1.000
 - Test Confidence: 0.933

Rule #9
Rule: If a person recommends Toy Story (1995), Empire Strikes Back, The (1980), Return of the Jedi (1983) they will also recommend Star Wars (1977)
 - Train Confidence: 1.000
 - Test Confidence: 0.971

Rule #10
Rule: If a person recommends Pulp Fiction (1994), Toy Story (1995), Shawshank Redemption, The (1994) they will also recommend Silence of the Lambs, The (1991)
 - Train Confidence: 1.000
 - Test Confidence: 0.794

همانطور که مشاهده می شود در تمام موارد confidenceدر دیتاست آموزش بیشتر از دیتاست آزمون است که امری طبیعی است اما قوانین استخراج شده با confidence بالا و همچنان قابل قبول در دیتاست آزمون بدست آمده اند.

۸- منابع

- فرادرس، https://blog.faradars.org/association-rule-mining-and-apriori-algorithm-using-r/

2. R. Layton. (2017). Learning Data Mining with Python. Packt Publishing Ltd.