دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

پروژه درس یادگیری ماشین آماری
مسأله‌ی اول: پیشنهاد فیلم به کاربر

دانشجو:
سید احمد نقوی نوزاد

استاد:
دکتر نیک‌آبادی

پاییز ۹۴

۱. شرح مسأله:

سیستم‌های سفارش‌گر (Recommendation Systems) این روزها به طور وسیعی توسط وب‌سایت‌های تجارت الکترونیک (e-commerce) مورد استفاده واقع می‌شوند و به نوعی یک جور بازیابی اطلاعات به حساب می‌آیند. اما برخلاف موتورهای جستجو و یا پایگاه‌های داده، به کاربران چیزهائی را پیشنهاد داده و یا ارائه می‌دهند که آن کاربران تا پیش از این چیزی راجع به آن نشنیده‌اند. به همین ترتیب سیستم‌های سفارش‌گر قادر به پیش‌بینی تمایلات ناشناخته‌ی کاربران با توجه به تمایلات شناخته‌شده‌ی آن‌ها می‌باشند. هزاران فیلم موجود هستند که مورد علاقه‌ی کاربران بسیاری می‌باشند و سیستم‌های سفارش‌گر آماده‌ی این هستند که جدای از همه‌ی این فیلم‌های مورد علاقه، بیان کنند کدام فیلم مطلوب شخص شما می‌باشد. اگرچه سیستم‌های سفارش‌گر بسیار کاربردی هستند، اما سیستم‌های فعلی همچنان بهبود بیشتری می‌طلبند، چرا که آن‌ها همیشه یا اقلام بسیار محبوب و یا هم اقلام عجیب که خارج از ذائقه‌ی کاربر می‌باشند را به وی سفارش می‌نمایند. سیستم‌های سفارش‌گر خوب دارای پیش‌بینی صحیح‌تر و پیچیدگی محاسبات کمتری می‌باشند.

در این مسئله نیز ما از مجموعه داده‌ی MovieLens ml-100k استفاده کرده و قصد داریم سیستمی را طراحی نمائیم که برای یک کاربر که اطلاعات سرشماری‌شده‌ی قبلی وی در مجموعه‌ی داده‌ی آموزشی موجود است، پیش‌بینی‌ای را در زمینه‌ی یک فیلم خاص که در مجموعه‌ی داده‌ی آزمایشی موجود است انجام داده و بیان نمائیم که رتبه‌ی کاربر به فیلم مزبور چه مقداری بین ۱ تا ۵ می‌باشد. این مجموعه داده شامل درایه‌هایی برای ۹۴۳ کاربر، ۱۶۸۲ فیلم و ۱۰۰ هزار رتبه‌دهی با مقادیری بین ۱ تا ۵ می‌باشد.

۲. مروری بر کارهای انجام‌شده (Literature Review) به صورت خلاصه

کارهای مربوط به این موضوع تاکنون چندین روش مانند دسته‌بندی‌کننده‌ی K نزدیک‌ترین همسایه (KNN)، دسته‌بندی‌کننده‌ی Naïve Bayesian و خوشه‌بندی K-Medians را برای سفارش‌دهی به کاربران با استفاده از همین مجموعه داده‌ی فعلی به کار برده‌اند و نتایج کار را با روش‌های احتمالاتی و پیش‌بینی مشارکتی (Probabilistic and Collaborative Prediction Techniques) مقایسه نموده‌اند. مؤثرترین روش‌ها برای پیش‌بینی نظرات کاربران در مورد فیلم‌ها در مورد این مجموعه‌ی داده، همان Collaborative Filtering یا

فیلترکردن مشارکتی و دیگر روش‌های Matrix Factorization بوده است. سایت MovieLens خود از روش CF جهت سفارش‌دهی به کاربران استفاده می‌نماید.

همان‌طور که در جدول یک در پایین نشان داده شده است، سیستم‌های سفارش‌گر با استفاده از راه‌های مختلفی قابل پیاده‌سازی هستند. آن‌ها تلاش می‌کنند به کاربر اقلامی را پیشنهاد دهند که احتمالا مورد علاقه‌ی وی می‌باشند و این کار را با استفاده از خصیصه‌هایی که از پروفایل کاربر استخراج شده‌اند انجام می‌دهند. برخی خصیصه‌ها مربوط به محتوای اقلام می‌باشند، و روش مبتنی بر این خصیصه‌ها رویکرد مبتنی بر محتوا (Content-based approach) نامیده می‌شود. به همین ترتیب برخی از خصیصه‌ها برگرفته از محیط اجتماعی کاربر می‌باشند که در نتیجه رویکرد مبتنی بر این خصیصه‌ها رویکرد فیلترکردن مشارکتی (Collaborative Filtering) نامیده می‌شود.

جدول ۱. انواع سیستم‌های سفارش‌گر

Recommendation Systems		
Content-Based	Collaborative Filtering	
	Model-Based	Memory-Based

رویکردهای مبتنی بر محتوا، محتوای اقلام را خوانده و شباهت میان اقلام با استفاده از خصایص استخراج‌شده از محتوا محاسبه می‌گردد. مزایای این رویکرد این است که الگوریتم قادر به مدیریت اقلام جدید بوده و همین‌طور دلیل هر رقم سفارش انجام‌شده توسط الگوریتم قابل توجیه خواهد بود. هرچند که همه‌ی انواع اقلام قابل خوانده‌شدن نیستند، اما سیستم‌های مبتنی بر محتوا به طور عمده بر روی اقلامی تمرکز می‌نمایند که دارای اطلاعات متنی و یا لفظی می‌باشند. اما در مورد سفارش‌دهی و یا به عبارتی پیش‌بینی درباره‌ی فیلم‌ها، رویکرد مبتنی بر محتوا کارکردی نداشته و ازین رو در این مسئله ما رویکرد CF را برمی‌گزینیم.

در مقایسه با رویکرد مبتنی بر محتوا، رویکرد فیلترکردن مشارکتی (CF) اهمیتی به محتوای اقلام نمی‌دهد، بلکه بر ارتباطی که میان کاربران و اقلام وجود دارد تمرکز می‌نماید. همین‌طور است که در این روش اقلامی که کاربران مشابه به آن اقلام علاقه‌مند هستند، مشابه در نظر گرفته می‌شوند.

در اینجا می‌خواهیم به طور ویژه در مورد فیلترکردن مشارکتی صحبت نمائیم. سیستم‌های فیلترکردن مشارکتی (CF) تلاش می‌کنند که برای یک کاربر خاص علاقه‌ی وی به اقلام خاصی

را بر اساس اقلام مورد علاقه‌ی سایر کاربران پیش‌بینی نمایند. تاکنون سیستم‌های فیلترکردن مشارکتی (CF) بسیاری در حیطه‌ی امورات دانشگاهی و البته صنعت توسعه داده شده‌اند. الگوریتم‌هایی که درباره‌ی فیلترکردن مشارکتی (CF) موجود هستند را می‌توان به دو دسته‌ی عمومی مبتنی بر حافظه (Memory-based) (یا مبتنی بر نوع جنس (Item-based)) و دسته‌ی دوم مبتنی بر مدل (Model-based) تقسیم نمود.

الگوریتم‌های مبتنی بر حافظه و یا به عبارتی مبتنی بر نوع جنس، به طور خاص و ضروری ابتکاراتی هستند که پیش‌بینی‌ها را بر اساس کل پایگاه داده انجام می‌دهند. مقادیر تصمیم‌گیرنده در مورد سفارش یک قلم جنس خاص، در قالب مجموعی از اطلاعات سایر کاربران برای همان نوع جنس محاسبه می‌گردد.

بر خلاف روش‌های مبتنی بر حافظه یا نوع جنس، الگوریتم‌های مبتنی بر مدل در ابتدا یک مدل خاص را با توجه به پایگاه داده ساخته و سپس پیش‌بینی‌ها را با توجه به مدل مربوطه انجام می‌دهند. تفاوت اصلی میان الگوریتم‌های مبتنی بر مدل و الگوریتم‌های مبتنی بر حافظه این است که الگوریتم‌های مبتنی بر مدل قوانین ابتکاری را به کار نمی‌گیرند. در عوض این مدل‌ها هستند که از پایگاه داده آموخته‌اند که چگونه به کاربران پیشنهاد دهند.

روش Improved Naïve Bayesian که در ادامه بررسی خواهد شد متعلق به الگوریتم‌های مبتنی بر مدل می‌باشد در حالی که الگوریتم K نزدیک‌ترین همسایه (KNN) متعلق به الگوریتم‌های مبتنی بر حافظه یا نوع جنس می‌باشد.

۳. شرح دقیق روش‌های پیاده‌سازی‌شده

در روش اول که همان روش Improved Naïve Bayesian می‌باشد، برای هر کاربر قصد داریم با توجه به علایق و سلایق شناخته‌شده‌ی وی، علایق ناشناخته‌ی وی را حدس بزنیم. در ابتدا کمی راجع به الگوریتم اصلی بِیزیَن یا همان Original Naïve Bayesian Method صحبت می‌کنیم. در این الگوریتم علاقه‌ی ناشناخته‌ی یک کاربر به صورت زیر بیان می‌گردد:

$$p\left(m_x \middle| m_{u_1}, m_{u_2}, \dots\right) \quad (1)$$

زمانی که علاقه‌ی کاربر به جنس m_x را در نظر می‌گیریم، موارد m_{u_1}، m_{u_2} و الی آخر را به عنوان علایق شناخته‌شده‌ی وی درنظر می‌گیریم. البته که m_x جزء علایق شناخته‌شده‌ی کاربر نمی‌باشد. احتمال شرطی به معنی احتمال آن است که جنس m_x مورد علاقه‌ی کاربری باشد که علایق شناخته‌شده‌ی وی موارد m_{u_1}، m_{u_2} و الی آخر می‌باشند. در الگوریتم ما، اقلام با

احتمال شرطی بالاتر، اولویت بالاتری برای سفارش‌شدن داشته و کار ما در این‌جا این است که احتمال شرطی مربوط به هر قلم جنس برای هر کاربر را محاسبه نمائیم.

$$p(m_x|m_{u_1}, m_{u_2}, \dots) = \frac{p(m_x).p(m_{u_1}, m_{u_2}, \dots|m_x)}{p(m_{u_1}, m_{u_2}, \dots)} \qquad (2)$$

در اینجا فرضیه‌ی استقلال شرطی را داریم:

$$p(m_{u_1}, m_{u_2}, \dots|m_x) = p(m_{u_1}|m_x).p(m_{u_2}|m_x) \dots. \qquad (3)$$

در عمل، مقایسه تنها در میان احتمالات شرطی مربوط به یک کاربر اتفاق افتاد، جائی که مخارج معادله‌ی (۲) ، $p(m_{u_1}, m_{u_2}, \dots)$ همگی یکسان بوده و تأثیری بر نتیجه‌ی نهائی ندارند. بنابراین محاسبات مربوطه به صورت زیر ساده می‌شود:

$$p(m_{u_1}, m_{u_2}, \dots) = p(m_{u_1}).p(m_{u_2}) \dots. \qquad (4)$$

بنابراین احتمال شرطی می‌تواند به صورت زیر محاسبه گردد:

$$p(m_x|m_{u_1}, m_{u_2}, \dots) = p(m_x).q \qquad (5)$$

به طوری که:

$$q = \frac{p(m_{u_1}, m_{u_2}, \dots|m_x)}{p(m_{u_1}, m_{u_2}, \dots)} = \frac{p(m_{u_1}|m_x)}{p(m_{u_1})}.\frac{p(m_{u_2}|m_x)}{p(m_{u_2})} \dots. \qquad (6)$$

و اما در مورد روش Improved Naïve Bayesian، باید گفت که در حقیقت فرضیه‌ی استقلال شرطی برای این مسئله‌ی خاص مناسب نمی‌باشد چرا که ارتباط میان اقلام در واقع بنیان نظریه‌ی الگوریتم ما می‌باشد.

$p(m_x)$ در (۵) نشان می‌دهد که آیا جنس مربوطه به خودی خود جذاب هست یا نه، و q نیز نشان می‌دهد که آیا این قلم مورد سفارش برای دقیقا همین کاربر مناسب می‌باشد یا خیر. در آزمایش‌ها مشخص شده که آخرین مورد اثر بیشتری را نسبت به آنچه تصور می‌شود دارد چرا که در اینجا فقدان استقلال وجود دارد. برای تطبیق بایاس باید داشته باشیم:

$$p(m_x|m_{u_1}, m_{u_2}, \dots) = p(m_x).q^{c_n/n} \qquad (7)$$

به طوری که n تعداد علایق شناخته‌شده‌ی کاربر بوده و c_n نیز یک ثابت بین ۱ و n می‌باشد. تبدیل آخر در (۷) اثر تمام n علاقه‌ی شناخته‌شده‌ی کاربر را برابر اثر تعداد c_n علاقه می‌نماید، که سبب کاهش شدید اثر علایق شناخته‌شده‌ی کاربر می‌گردد. در واقع c_n نشان می‌دهد که اقلام چه قدر از یکدیگر مستقل هستند. مقدار c_n از طریق انجام آزمایشات متعدد حاصل شده و معمولا برای بیشتر nها مقداری حدود ۳ را دارد.

و اما در مورد روش دوم که همان الگوریتم KNN و جزء الگوریتم‌های CF مبتنی بر حافظه یا نوع جنس می‌باشد، به صورت زیر عمل می‌کنیم:

رتبه‌ی r که ما برای فیلم i ارائه‌شده توسط کاربر u را پیش‌بینی می‌کنیم، به صورت مجموعی از رتبه‌ها خواهد بود که توسط N کاربری ارائه شده‌اند که بیشترین شباهت را به کاربر u داشته و به فیلم i رتبه داده‌اند.

$$r_{u,i} = aggr_{u' \in U} r_{u',i} \qquad (8)$$

به طوری که U مجموعه‌ی تمامی کاربران به جز کاربر u می‌باشد.

برای پیداکردن شبیه‌ترین N کاربر، ما از هر دوی Pearson Correlation Similarity میان دو کاربر x و y که به صورت زیر است:

$$simil(x,y) = \frac{\sum_{i \in I_{x,y}} (r_{x,i} - \overline{r_x})(r_{y,i} - \overline{r_y})}{\sqrt{\sum_{i \in I_{x,y}} (r_{x,i} - \overline{r_x})^2 \sum_{i \in I_{x,y}} (r_{y,i} - \overline{r_y})^2}} \qquad (9)$$

و یا Cosine Similarity که به صورت زیر است بهره می‌بریم:

$$simil(x,y) = \frac{\sum_{i \in I_{x,y}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_{x,y}} {r_{x,i}}^2 \sum_{i \in I_{x,y}} {r_{y,i}}^2}} \qquad (10)$$

به طوری که $I_{x,y}$ مجموعه‌ی فیلم‌هائی است که توسط هر دو کاربر x و y رأی داده شده‌اند.

توابع شباهت می‌توانند برای یافتن شباهت میان هر دوی کاربران و فیلم‌ها نیز به کار روند.

و اما سیستم سفارش‌دهی نهائی دو وظیفه‌ی زیر را انجام خواهد داد:

-**سفارش‌دهی:** این‌که آیا کاربر مربوطه فیلم را خواهد پسندید؟ این هم می‌تواند به عنوان یک معیار اطمینان به کار رود و یا هم به صورت یک جواب بله یا خیر ساده با در نظرگرفتن یک حد آستانه بر روی معیار اطمینان. برای همین، ما در ابتدا تعداد N کاربر مشابه به کاربر جدیدمان را با استفاده از خصیصه‌های کاربر پیدا می‌کنیم اگر که کاربر به طور کامل جدید باشد؛ و یا هم برای یافتن این N کاربر مشابه، از خصیصه‌های کاربر و گذشته‌ی وی استفاده می‌بریم اگر که پیش از این به تعدادی فیلم رأی داده باشد. بر این اساس ما یک فیلم را به کاربر سفارش می‌کنیم.

-**پیش‌بینی:** اگر یک کاربر فیلمی را مشاهده نماید، چه رأیی به آن خواهد داد؟ و این جواب عددی مابین ۱ و ۵ خواهد بود. این مطلب سبب می‌گردد تا آرائی که کاربر به فیلم‌های مشابه داده است نیز در نظر گرفته شوند.

۴. ارائه‌ی نتایج به دست‌آمده و مقایسه‌ی آن‌ها

برای هر روش آزمایش‌های مربوطه مطابق الگوریتم‌های گفته‌شده انجام گردیده و نتایج مطابق جدول زیر ارائه می‌گردند:

Approches	Final Errors	
Naïve Bayesian Method	MAE Error	2.4152
	RMSE Error	2.8234
KNN Algorithm	RMSE	1.0700

۵. جمع‌بندی و نتیجه‌گیری

همان‌طور که قابل مشاهده است الگوریتم اول یا همان Naïve Bayesian Method خطای RMSE مطلوبی را در مقایسه با الگوریتم دوم یا همان KNN Algorithm ارائه نمی‌نماید و علت آن نیز می‌تواند همانی باشد که پیشتر قید شد؛ این که در عین این که فرضیه‌ی استقلال شرطی برای این مسئله‌ی خاص مناسب نمی‌باشد و علت آن نیز این است که ارتباط میان اقلام در واقع بنیان نظریه‌ی الگوریتم ما می‌باشد، ما آن را در هر صورت به کار برده‌ایم.

بنابراین در نهایت در میان روش‌های پیاده‌سازی‌شده روش دوم یا همان KNN برنده‌ی مدل‌های انتخابی می‌باشد.

۶. مراجع

[1] Adomavicius, G., Tuzhilin, A.: The next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering (2005)

[2] Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing (2003)

[3] Benjamin Marlin, Collaborative filtering: A machine learning perspective,

Master’s thesis, University of Toronto, Canada, 2004. Intelligence (July 1998)