پایان‌نامه دوره کارشناسی کامپیوتر

گرایش مهندسی نرم‌افزار

موضوع:

طراحی و پیاده‌سازی سامانه توصیه گر و صفحات وب

استاد راهنما:

دکتر جواد حمید زاده

نام دانشجویان:

دانیال جاهد-دانیال وفادار راد

بهمن ۱۳۹۵


بسم الله الرحمن الرحيم

پایان‌نامه دوره کارشناسی ناپیوسته کامپیوتر

گرایش مهندسی نرم‌افزار

موضوع:

طراحی و پیاده‌سازی سامانه توصیه گر و صفحات وب

استاد راهنما:

دکتر جواد حمیدزاده

استاد داور:

دکتر محمدمهدی سالخورده حقیقی – مهندس احمد شکرانی بایگی

نام دانشجویان:

دانیال جاهد - دانیال وفادار راد

بهمن ۱۳۹۵

تقديم به:

تمامي ره پويان راه علم و معرفت

سپاسگزاري

حضرت علي (ع) فرمودند:

«مَن عَلَّمَني حَرفاً فَقَد سَيَرني عَبداً»

از تمامي معلمين، اساتيد، دوستان و خانوادهايمان در كل دوران تحصيل سپاسگزاريم.

چکیده

سامانه توصیه گر و صفحات وب

سیستم‌های توصیه‌گر[1] در سال‌های اخیر در مواجهه با مسئله‌ی سرریز اطلاعات از طریق پیشنهاد رایج‌ترین محصولات از بین حجم عظیم داده به کاربران، گسترش فراوانی یافته‌اند. برای محصولات چندرسانه‌ای[2]. سامانه توصیه گر گروهی[3] در تلاش است برای کمک به کاربران جهت دسترسی به فیلم‌های موردعلاقه از طریق پیدا کردن همسایه‌های مشابه در بین کاربران یا محصولات بر اساس امتیازات گذشته می‌باشد. بااین‌وجود، به دلیل پراکندگی داده‌ها، انتخاب همسایه با افزایش سریع محصولات و کاربران، سخت‌تر شده است. در این پروژه، یک سیستم توصیه‌گر مبتنی بر مدل ترکیبی پیش نهاد می‌شود که از الگوریتم خوشه‌بندی[4] بهبودیافته به همراه الگوریتم ژنتیک جهت تقسیم فضای کاربر استفاده می‌کند. این سیستم تکنیک کاهش داده‌های تحلیل مؤلفه‌های اساسی[5] را برای متراکم کردن فضای جمعیت محصولات به کار می‌گیرد که می‌تواند پیچیدگی محاسباتی را در سامانه توصیه‌گر هوشمند نیز کاهش دهد.

کلمات کلیدی

سامانه توصیه‌گر، پلایش گروهی[6]، خوشه‌بندی، ژنتیک، دادهای پراکنده.

---

[1]Recommendation System

[2] Multi Media

[3] Collaborative Recommendation System

[4] Clustring

[5] Principle Component Analysis

[6] Collaborative Filtering

فهرست مطالب

عنوان صفحه

فهرست جداول

عنوان صفحه

فهرست شکل‌ها

عنوان صفحه

مقدمه

سیستم‌های توصیه‌گر معمولاً بر اساس چگونگی تولید توصیه‌ها، به دسته‌های زیر طبقه‌بندی می‌شوند:

- پالایش مبتنی بر همکاری[7] (پالایش گروهی): به کاربر اقلامی توصیه خواهد شد که دیگران درگذشته با تمایلات و ترجیحات مشابه او، این اقلام را پسندیده‌اند.
- پالایش مبتنی بر محتوا[8]: به کاربر اقلامی توصیه خواهد شد که او قبلاً آن‌ها را ترجیح داده است.
- رویکردهای ترکیبی[9]: در این رویکردها، از ترکیبی از رویکردهای دیگر استفاده می‌شود.

- پالایش مبتنی بر دانش[10]: متدهای مبتنی بر دانش یا رویکرد پالایش مبتنی بر دانش با استدلال در مورد آن اقلامی که نیازمندی‌های کاربر را رفع می‌کنند، توصیه‌هایی ایجاد می‌کند. معمولاً زمانی که امکان اعمال پالایش گروهی و یا پالایش مبتنی بر محتوا وجود ندارد (به دلیل، رتبه دار نبودن اقلام، یا مشکلات شروع سرد)، از پالایش مبتنی بر دانش استفاده می‌شود. برای مثال، در یک سیستم توصیه‌گر آپارتمان یا اتومبیل، به دلیل اینکه تعداد خریدها و بازخوردها (رتبه‌دهی) به‌ندرت انجام می‌گیرد، درنتیجه، پالایش مبتنی بر محتوا و مبتنی بر همکاری، به دلیل کافی نبودن امتیازهای موجود، قابل‌اعمال نمی‌باشند. درنتیجه، متدهای مبتنی بر دانش در زمینه‌های زیر اعمال می‌شود:
  - ❖ کاربران یا مشتریان، نیازمندی‌های خود را به‌صورت صریح در سیستم وارد می‌کنند (برای مثال، اتومبیل باقیمت کمتر از ۱۵ میلیون تومان)
  - ❖ نیازمندی‌های کاربران یا محدودیت‌های توصیه، از طریق سؤال و پرسش کسب می‌شود.

---

[7] Collaborative filtering

[8] Content-based filtering

[9] Hybrid

[10] Knowledge-based filtering

- ❖ زمانی که در سیستم امکان نقد پیشنهاد برای کاربر، لحاظ شده باشد. برای مثال، پس از پیشنهاد اتومبیلی به کاربر، کاربر با مشاهده ویژگی‌های اتومبیل، بگوید اتومبیل ارزان‌تری را دوست دارد.

۱-فصل یکم بررسی سامانه‌ی توصیه گر

رشد سریع تکنولوژی اینترنت منجر به رشد روزافزون اطلاعات موجود در دهه اخیر شده است. سیستم‌های توصیه گر، به‌عنوان یکی از کاربردهای موفق فیلترینگ اطلاعات، روش مؤثری برای حل مسئله سرریز اطلاعات به شمار می‌رود. هدف سیستم‌های توصیه گر، تولید اتوماتیک هدف‌های پیش نهادی (فیلم‌ها، کتاب‌ها، اخبار، موسیقی، سی‌دی‌ها، دی‌وی‌دی‌ها، وب پیج‌ها) برای کاربران با توجه به ترجیحات تاریخی‌شان مفید می‌باشد.

۱-۱-انواع سامانه‌های توصیه گر

۱-۱-۱-پالایش مبتنی بر همکاری

در سیستم‌های توصیه‌گر همکاری محور (سیستم‌های پالایش مبتنی بر همکاری) بهره‌وری اقلام برای یک کاربر خاص، بر اساس اقلامی که قبلاً به‌وسیله‌ی کاربران مشابه، رتبه‌دهی شده‌است، پیش‌بینی می‌شود. برای مثال، در یک سامانه توصیه‌ی فیلم، برای توصیه‌ی فیلم‌ها به کاربر u، سیستم توصیه‌گر مبتنی بر همکاری سعی می‌کند تا کاربران متناظر با کاربر u را پیدا کند، برای نمونه، کاربران دیگری که دارای سلایق مشابه در فیلم‌ها هستند (به همین فیلم‌ها به‌صورت مشابه رتبه می‌دهند). سپس، تنها فیلم‌هایی که بیشتر، موردپسند کاربران متناظر با کاربر هدف قرارگرفته‌اند، توصیه می‌شود.

رویکردهای گوناگونی برای محاسبه‌ی مشابهت بین دو کاربر در سیستم‌های توصیه‌گر مبتنی بر همکاری استفاده‌شده است. در بسیاری از این رویکردها، مشابهت بین دو کاربر بر اساس رتبه‌هایی است که هر دو کاربر به اقلام داده‌اند. نتیجه‌ی اندازه‌گیری مشابهت، بین +۱, -۱ است؛ +۱ مشابهت کامل را نشان می‌دهد و مقادیر مثبت بیان‌گر وجود شباهت بین دو کاربر است و نتیجه‌ی منفی، بیان‌گر عدم مشابهت در امتیازات دو کاربر به اقلام، می‌باشد.

درروش توصیه‌گری پالایش گروهی، ابتدا باید اجازه داد تا کاربران در سیستم مشارکت نمایند و به آیتم‌های مختلف موجود در سیستم امتیاز دهند. البته این امتیاز دادن‌ها، می‌تواند به‌صورت ضمنی نیز اتفاق افتد و توسط سیستم تشخیص داده شوند. به‌عنوان‌مثال، یک نوع امتیاز دادن ضمنی می‌تواند به این شکل باشد که آیتم‌هایی که بیشتر دانلود شده‌اند، احتمالاً از محبوبیت بیشتری برخوردار بوده‌اند و درنتیجه امتیاز بیشتری نسبت به بقیه به آن‌ها داده می‌شود. فلسفه این روش، استفاده از نظرات دیگران در تصمیم‌گیری است که برای قرن‌ها مورداستفاده انسان‌ها بوده است. برای مثال اگر دوستان شما از فیلمی تعریف کنند، شما راغب به دیدن آن فیلم خواهید شد؛ و یا برعکس اگر از فیلمی تعریف نکنند، احتمال کمی دارد که آن فیلم را ببینید.

به‌علاوه بعد از مدتی شما خواهید فهمید که نظرات کدام‌یک از دوستانتان به نظرات شما نزدیک‌تر است و به‌تدریج فقط به آن دسته از دوستان که به شما شباهت دارند، توجه خواهید کرد.

یک سیستم پالایش مشارکتی برای برطرف کردن نیازهای کاربرانی با ویژگی‌های زیر مناسب‌تر است. اگرچه برای هر حالت دیگری می‌توان از این رویکرد استفاده کرد، اما در موارد زیر، از سایر موارد بهتر جوابگو می‌باشد و پاسخ‌های مناسب‌تری تولید می‌نماید:

- نیاز به راهنمايي دريافتن يک آيتم موردعلاقه جديد: سيستم بايد بتواند از بين حجم انبوهي از آيتم‌ها، آن دسته را که احتمالاً کاربر به آن‌ها متمايل است را پالايش کرده و ارائه دهد.
- نیاز به راهنمايي در مورد يک آيتم مشخص: اين نياز در شرايطي ظاهر مي‌شود که کاربر، آيتمي را در ذهن دارد و نياز دارد نظر افرادي که مشابه وي هستند را بداند تا راحت‌تر و مطمئن‌تر در مورد آن تصميم‌گيري نمايد.
- نیاز يک کاربر به يافتن کاربري که علايق مشابه وي را دارد: گاهي اوقات، نياز کاربران اين است که بدانند، بايد به نظر چه کساني توجه کنند. با مطابقت دادن و مرتبط کردن کاربران مشابه با يکديگر، مي‌توان به اين نياز پاسخ داد.
- نیاز يک گروه از کاربران دريافتن آيتم‌هاي موردعلاقه: به کمک تکنيک پالايش گروهی مي‌توان به گروه‌هايي از کاربران، آيتم‌هايي را پيشنهاد کرد که موردپسند اعضاي آن‌ها واقع شود.

مزایا و معایب درروش پالایش مبتنی بر همکاری

- مزایا:
  - ❖ در این روش نیاز نیست اطلاعاتی درباره ویژگی‌های آیتم‌ها داشته باشیم.
- معایب:
  - ❖ شروع سرد[11]: آیتم‌هایی که توسط کاربران کمی امتیازدهی شده‌اند به‌سختی می‌توانند پیشنهاد شوند.

[11] Cold start

- ❖ مشکل اسپارس بودن داده‌ها[12]: اگر تعداد کاربران و آیتم‌ها زیاد باشد، ماتریس کاربر/ امتیاز پراکنده (sparse) است.
- ❖ مشکل تعصب محبوبیتی[13]: سیستم در این روش، بیشتر آیتم‌هایی را پیشنهاد می‌دهد که محبوب و شناخته‌شده‌اند.

۱-۱-۲-پالایش مبتنی بر محتوا

در برخی از سیستم‌های توصیه‌گر، ممکن است از پالایش مبتنی بر محتوا استفاده شود که پیشنهادها را بر اساس انتخاب‌های گذشته‌ی کاربر، ارائه می‌دهد (برای مثال، در برنامه کاربردی تحت وب توصیه‌ی فیلم، اگر کاربر چند فیلم تخیلی را درگذشته خریداری کرده باشد، سیستم توصیه ممکن است فیلم‌های تخیلی اخیری را که او تابحال نخریده است، پیشنهاد دهد). همچنین، پالایش مبتنی بر محتوا، پیشنهادها را با استفاده از محتوای اشیاء در نظر گرفته‌شده برای پیشنهاد، ارائه می‌دهد؛ بنابراین، محتوای خاصی ممکن است تحلیل شود، مانند متن، تصویر، صدا. برای مثال، در کاربرد توصیه فیلم، برای توصیه فیلم‌ها به کاربر c ، سیستم توصیه‌گر مبتنی بر محتوا، سعی می‌کند تا نقاط اشتراک میان فیلم‌هایی که کاربر c درگذشته برای آن‌ها رتبه‌ی بالایی را در نظر گرفته را پیدا کند (بازیگران خاص، کارگردانان، ژانر، موضوع و غیره). سپس، تنها فیلم‌هایی را که دارای میزان بالای شباهت به هر یک از ترجیحات کاربر هستند را توصیه می‌کند.

به‌صورت عمومی، مراحل پالایش مبتنی بر محتوا به شرح زیر است:

- استخراج صفات[14] مربوط به آیتم‌ها: برای آن‌که یک سیستم توصیه‌گر مبتنی بر محتوا به‌خوبی عمل نماید، ابتدا می‌بایست صفات مربوط به آیتم‌ها استخراج شوند. عموماً صفات به‌طور صریح همراه با آیتم‌ها در سیستم درج می‌شوند؛ بنابراین استخراج این‌گونه صفات با مشکل خاصی مواجه نمی‌باشد؛

---

12 Data Sparsity

13 Popularity bias

14 attributes

اما گروهی دیگر از صفات هستند که بر اساس دامنه[15] سیستم، برای استخراج آن‌ها باید از تکنیک-های خاصی استفاده نمود. به‌عنوان‌مثال در سیستم‌هایی که آیتم‌ها اسناد متنی هستند، می‌بایست از روش‌های کلاسیک بازیابی اطلاعات[16] استفاده نمود تا بتوان به صفاتی از قبیل طول اسناد دست پیدا کرد.

- مقایسه صفات آیتم‌ها با سلایق کاربر: پس از مشخص شدن صفات آیتم‌ها، باید تحلیل‌هایی صورت پذیرد که نشان دهد آیتم‌های موجود در سیستم تا چه اندازه با علایق و سلایق کاربران همخوانی دارند که این کار عموماً با استفاده از روش‌هایی از قبیل روش‌های اکتشافی[17] و یا الگوریتم‌های انجام می‌شود.
- پیشنهاد دادن آیتم‌هایی که شباهت بیش‌تری به سلایق کاربر دارند.

در پالایش مبتنی بر محتوا، بایستی که از ابتدای کار، علایق و نیازهای کاربر را بدانیم و از محتوا و ویژگی آیتم‌های موجود، اطلاعات کافی داشته باشیم. البته به‌تنهایی، دانستن این موارد کافی نیست، بلکه باید توانایی این را داشته باشیم که به‌طور مداوم علایق و نیازهای کاربر را دنبال کنیم و در صورت تغییر، بتوانیم اطلاعات خود را در مورد کاربر هدف، به‌روزرسانی کنیم.

مزایا و معایب درروش پالایش مبتنی بر محتوا

- مزایا:
  - ❖ نیازی به اطلاعات دیگر کاربران نیست.
  - ❖ سیستم در این رویکرد به کاربران با علایق خاص هم می‌تواند پیشنهادهای مناسب دهد.
  - ❖ می‌تواند آیتم‌های جدید و یا کمتر شناخته‌شده را هم به‌درستی پیشنهاد دهد.

---

[15] domain

[16] information retrieval

[17] heuristic

- معایب:
  - ❖ از قضاوت دیگر کاربران استفاده نمی‌کند.
  - ❖ به دست آوردن ویژگی‌های آیتم‌ها دشوار است.

۱-۱-۳-پالایش مبتنی بر دانش

متدهای مبتنی بر دانش یا رویکرد پالایش مبتنی بر دانش با استدلال در مورد آن اقلامی که نیازمندی‌های کاربر را رفع می‌کنند، توصیه‌هایی ایجاد می‌کند (برای نمونه، توصیه‌ای برای اتومبیلی که موارد وابستگی به اقتصاد سوخت یا راحتی، اهمیت بیشتری برای کاربر دارد). دانش از طریق ثبت ترجیحات انتخاب‌های کاربر یا از طریق درخواست از کاربر جهت ارائه‌ی نیازمندی‌ها، ایجاد می‌شود. تابع مشابهت، نشان‌دهنده حدی است که نیازهای کاربر با گزینه‌های اقلام موجود، مرتبط می‌باشند. مقدار تابع مشابهت معمولاً برای نشان دادن مفید بودن هر توصیه بکار می‌رود. معمولاً زمانی که امکان اعمال پالایش همکاری و محتوا محور وجود ندارد (به دلیل، رتبه دار نبودن اقلام، یا مشکلات شروع سرد) از پالایش مبتنی بر دانش استفاده می‌شود. برای مثال، در یک سیستم توصیه گر آپارتمان یا اتومبیل، به دلیل اینکه تعداد خریدها و بازخوردها (رتبه دهی) به‌ندرت انجام می‌گیرد، درنتیجه، پالایش مبتنی بر محتوا و مبتنی بر همکاری، به دلیل کافی نبودن امتیازهای موجود، قابل‌اعمال نمی‌باشند. متدهای مبتنی بر دانش درزمینهٔ مای زیر اعمال می‌شود:

کاربران یا مشتریان، نیازمندی‌های خود را به‌صورت صریح در سیستم وارد می‌کنند (اتومبیل باقیمت کمتر از ۱۵ میلیون تومان) نیازمندی‌های کاربران یا محدودیت‌های توصیه، از طریق سؤال و پرسش کسب می‌شود.

زمانی که در سیستم امکان نقد پیشنهاد برای کاربر، لحاظ شده باشد. برای مثال، پس از پیشنهاد رستورانی به کاربر، کاربر با مشاهده ویژگی‌های رستوران، بگوید «رستوران ارزان‌تری را دوست دارد»

معایب استفاده از متدهای پالایش مبتنی بر دانش:

هزینه‌ی کسب دانش از کاربر (برای مثال، درزمینهٔ مایی که واردکردن ترجیحات به‌صورت صریح توسط کاربر، هزینه‌بر هست).

توصیه‌های جزئی‌تر، نیازمند دانش جزئی‌تر است که نیازمند چرخه‌های تعامل بیشتری می‌باشد.

با توجه به اینکه شخصی‌سازی از مهم‌ترین اهداف سیستم‌های توصیه گر می‌باشد، در این سیستم‌ها توسط تکنیک‌هایی، حساسیت و واکنش به پروفایل‌ها و زمینه‌های کاربر ایجاد می‌شود. پروفایل کاربر در سامانه‌های توصیه گر، ساختاری است شامل اطلاعاتی که به‌طور مستقیم یا غیرمستقیم، ترجیحات کاربر، رفتار و زمینه را نگهداری می‌کند. یک پروفایل معنایی کاربر، توصیفی از علایق و غیر علایق کاربر است. پروفایل‌های کاربر باید بیشتر از لیست ساده‌ای از کلمات کلیدی علایق باشند. آنتولوژی، مکانیزم های استنتاج که می‌تواند توصیه را بهبود ببخشد، پشتیبانی می‌کند. درواقع، آنتولوژی مدل معنایی است که برای توصیف یک دامنه اعمال می‌شود و علایق در یک دامنه را به‌عنوان مجموعه‌ای از مفاهیم و ارتباطات مدل می‌کند. درواقع آنتالوژی های دامنه، تطبیق معنایی اقلام و پروفایل‌ها را ممکن می‌سازند (به‌جای انطباق کلمه کلیدی). برای مثال کلمه پایتون در مارِ پایتون و زبان برنامه‌نویسی پایتون ازنظر معنایی باهم متفاوت‌اند.

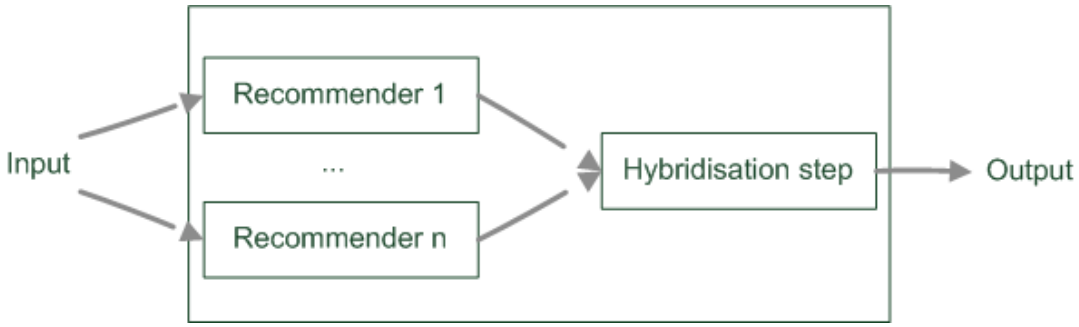
۱-۱-۴-رویکردهای ترکیبی

سیستم‌هایی هستند که از ترکیب متدهای گفته‌شده استفاده می‌کنند و از مزیت‌های یک تکنیک برای رفع مشکلات تکنیک دیگر بهره می‌برند و درنتیجه بهره‌وری را افزایش می‌دهند. ترکیب متدها ممکن است به روش‌های مختلف انجام گیرد: برای مثال، با ایجاد پیش‌بینی‌های مبتنی بر محتوا و مبتنی بر همکاری به‌صورت مجزا و سپس ترکیب آن‌ها؛ و یا با اضافه کردن قابلیت‌های محتوا محور به رویکرد همکاری محور و غیره. رویکردهای ترکیبی خود به طراحی‌های مختلف طبقه‌بندی می‌شوند. دو مورد از طراحی‌های رایج در این رویکرد عبارت‌اند از:

- طراحی هیبریداسیون موازی[۱۸]: در این روش، خروجی مربوط به رویکردهای مختلف سیستم‌های توصیه‌گر باهم ترکیب می‌شوند و در میان پیشنهادهای متفاوت، بهترین پیشنهادها انتخاب می‌شوند. مثلاً اگر سیستم توصیه‌گر ترکیبی ما، متشکل از رویکردهای مبتنی بر محتوا و گروه‌محور باشد، به‌این‌ترتیب کار می‌کند که سیستم مبتنی بر محتوا چندین پیشنهاد تولید می‌کند، سپس سیستم

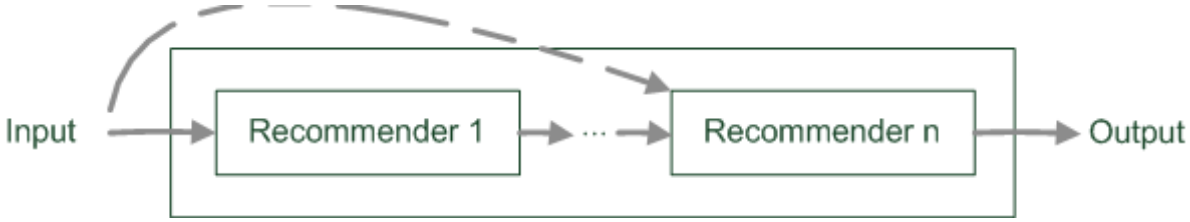
[18] Parallelized hybridization design

پالایش گروهی نیز پیشنهادهایی تولید می‌کند، حال از بین تمامی پیشنهادهای موجود، یک یا چند پیشنهاد که می‌توانند بهترین پیشنهادهای موجود باشند، به کاربر هدف ارائه می‌شوند.



شکل ۱-۱ مثالی از سیستم‌های توصیه‌گر موازی ترکیبی

- طراحی هیبریداسیون خط لوله: در این روش، دو یا چند سیستم توصیه‌گر با رویکردهای مختلف، به‌طور سری قرارگرفته‌اند. حال این سیستم ترکیبی به این روش عمل می‌کند که سیستم توصیه‌گر اول، چندین پیشنهاد تولید می‌کند، پیشنهادهای این سیستم توصیه‌گر، به‌عنوان ورودی به سیستم توصیه‌گر بعدی می‌رود و مجدد پالایش می‌شود. این کار آن‌قدر اجرا می‌شود که درنهایت، تعدادی پیشنهاد از آخرین سیستم توصیه‌گر موجود خارج می‌شود و به کاربر موردنظرمان ارائه می‌گردد.



شکل ۱-۲ مثالی از سیستم‌های توصیه گر خط لوله ترکیبی

۱-۱-۵-چالش‌های اصلی

مشکل شروع سرد: که از مهم‌ترین مسائل پیش روی سیستم‌های توصیه گر می‌باشد، زمانی رخ می‌دهد که ایجاد توصیه‌های قابل‌اعتماد به علت عدم وجود رتبه‌های اولیه، ممکن نیست. این مشکل در سه وضعیت به وجود می‌آید: قلم جدید و کاربر جدید. آخرین نوع در سیستم‌های توصیه گر از اهمیت بیشتری برخوردار است.

مشکل قلم جدید به دلیل اینکه اقلام جدید واردشده در سیستم توصیه گر معمولاً فاقد رتبه‌های اولیه نیستند، بنابراین سیستم توصیه گر مایل به توصیه‌ی این اقلام نمی‌باشد. قلمی که توصیه نشود، توسط جامعه بزرگی از کاربران بی‌توجه باقی می‌ماند و به دلیل اینکه آن‌ها از این قلم آگاه نیستند، درنتیجه به این قلم رتبه‌ای اختصاص نمی‌دهند. این مشکل تأثیر اندکی بر روی فرآیند توصیه دارد، زیرا می‌توان اقلام را به نحوی دیگری به کاربران معرفی کرد (مثلاً فیلم در تبلیغات). راه‌حل رایج دیگر برای این مشکل، داشتن مجموعه‌ای از کاربران مشتاق که مسئولیت آن‌ها امتیازدهی به قلم جدید در سیستم باشد.

مشکل کاربر جدید- یکی از مشکلات بزرگی است که سیستم‌های توصیه گر با آن روبرو می‌باشند. به دلیل اینکه کاربران جدید در سیستم هنوز هیچ امتیاز یا رتبه‌ای را ارائه نداده‌اند، این کاربران نمی‌توانند پیشنهاد شخصی‌سازی‌شده‌ای را بر اساس متدهای همکاری محور مبتنی بر حافظه دریافت کنند؛ زمانی که کاربران اولین امتیازات خودشان را وارد می‌کنند، آن‌ها از سیستم توصیه گر انتظار دارند تا توصیه‌های مرتبط با علایق خودشان را ارائه دهد، اما معمولاً این تعداد امتیاز ارائه‌شده در سیستم توصیه گر هنوز هم برای ارائه‌ی توصیه‌های همکاری محور قابل‌اعتماد، کافی نمی‌باشند و بنابراین کاربران جدید احساس می‌کنند که سیستم سرویس‌هایی را که انتظار دارند به آن‌ها ارائه نمی‌دهد و ممکن است از استفاده از سیستم صرف‌نظر کنند.

خلوت بودن داده‌ها- در هر سیستم توصیه گر، تعدادی از رتبه‌هایی که به‌دست‌آمده‌اند در مقایسه با تعداد رتبه‌هایی که برای پیش‌بینی نیاز است، بسیار کم است. پیش‌بینی مؤثر رتبه دهی‌ها از موارد با تعداد کم، حائز اهمیت می‌باشد. درواقع وقتی اقلام در دسترس، اغلب به‌طور فزاینده‌ای زیاد می‌شوند (برای مثال، کتاب‌فروشی برخط، چندین میلیون کتاب را ارائه می‌کند)، در چنین شرایطی اشتراک بین دو کاربر بسیار کم است و یا حتی اشتراکی میان آن‌ها نیست. زیرابه نسبت تعداد بسیار زیاد انواع اقلام، کاربران اقلام تعداد کمی از رتبه‌ها را ارائه/دریافت می‌کنند. موفقیت سیستم توصیه گر همکاری محور وابسته به در دسترس بودن تعداد بسیار زیاد کاربران است. برای مثال، در سیستم توصیه‌ی فیلم،

فیلم‌های بسیاری توسط تنها تعداد کمی از افراد رتبه دهی شده‌اند و حتی اگر این کاربران اندک، رتبه‌های بالایی را به این فیلم‌ها اختصاص داده باشند هنوز هم این فیلم‌ها به‌ندرت پیشنهاد خواهند شد. همچنین برای کاربرانی که سلایق آن‌ها در مقایسه با کل جمعیت، غیرمعمول باشد، کاربران دیگری وجود ندارند تا شبیه این کاربر باشند که بازهم به پیشنهادهای ضعیف منجر خواهد شد. یک‌راه حل برای فائق آمدن بر مشکل خلوت بودن رتبه دهی‌ها، استفاده از اطلاعات پروفایل کاربر در زمان محاسبه‌ی مشابهت کاربر است.

قابلیت مقیاس‌پذیری- درحالی‌که داده‌ها اکثراً پراکنده هستند، سایت‌های اصلی شامل چند میلیون کاربر و قلم است؛ بنابراین توجه به مسائل هزینه‌ی محاسباتی و جستجوی الگوریتم‌هایی که نیازمندی‌های کمی دارند و به‌آسانی قابل موازی‌سازی هستند، امری است حیاتی. درواقع راه‌حل ارائه‌شده در سیستم‌های اخیر، پیاده‌سازی سرویس‌دهنده و پایگاه‌های داده به‌صورت توزیع‌شده و در بستر ابر میباشد.

ارزش زمان- بسیاری از الگوریتم‌های توصیه، مهرهای زمان ارزیابی‌ها را نادیده می‌گیرند، درحالی‌که کاربران واقعی دارای تمایلات متنوعی در زمان‌های مختلف هستند (برای مثال، تمایلات کوتاه‌مدت مربوط به رفتن به مسافرت و علایق طولانی‌مدت مربوط به مکان زندگی). این‌که آیا و چگونه مقدار رأی‌های قدیمی با گذر زمان باید فاسد شوند و الگوهای موقتی در ارزیابی کاربر چه هستند، از مسائل مهم تحقیقاتی هستند.

رابط کاربر- نشان داده‌شده است که برای سهولت قبول توصیه‌ها از جانب کاربران، توصیه‌ها باید شفاف باشند: کاربران مایل هستند دلیل پیشنهادها به آن‌ها، شفاف باشد. مسئله‌ی دیگر زمانی است که لیست اقلام موردعلاقه بسیار طولانی باشد، آن‌وقت نیاز است تا این لیست به‌صورت ساده ارائه شود.

۱-۲-بررسی کلیات و روال پروژه

سامانه توصیه گر فیلم، یک کاربرد گسترده به همراه سکوهای چندرسانه‌ای آنلاین است که هدف آن کمک به مشتریان برای دسترسی هوشمندانه به فیلم‌های موردنظر از دسته وسیع فیلم‌هاست. تحقیقات فراوانی در فضای صنعتی و دانشگاهی‌ای جهت گسترش الگوریتم‌های جدید توصیه گر فیلم صورت گرفته است. اکثریت سیستم‌های نظریه‌ای موجود، مبتنی بر مکانیزم پالایش گروهی است که در چند سال اخیر رشد موفقیت‌آمیزی را طی کرده‌اند. در ابتدا، امتیازهای فیلم‌های ارائه‌شده توسط هر فرد را جمع کرده و سپس فیلم‌های موردنظر را برای مشتری هدف بر اساس افراد هم‌فکر با علایق و ترجیحات مشابه درگذشته، پیش نهاد می‌دهد. چندین پایگاه چندرسانه‌ای آنلاین معروف در مشارکت با الگوریتم پالایش گروهی وجود دارند تا

محصولات رسانه‌ای را به مشتریانشان پیش نهاد دهند. بااین‌وجود، سیستم‌های توصیه گر سنتی همیشه محدودیت‌هایی دارند: مقیاس‌پذیری ضعیف، پراکندگی داده و مسائل شورع سرد.

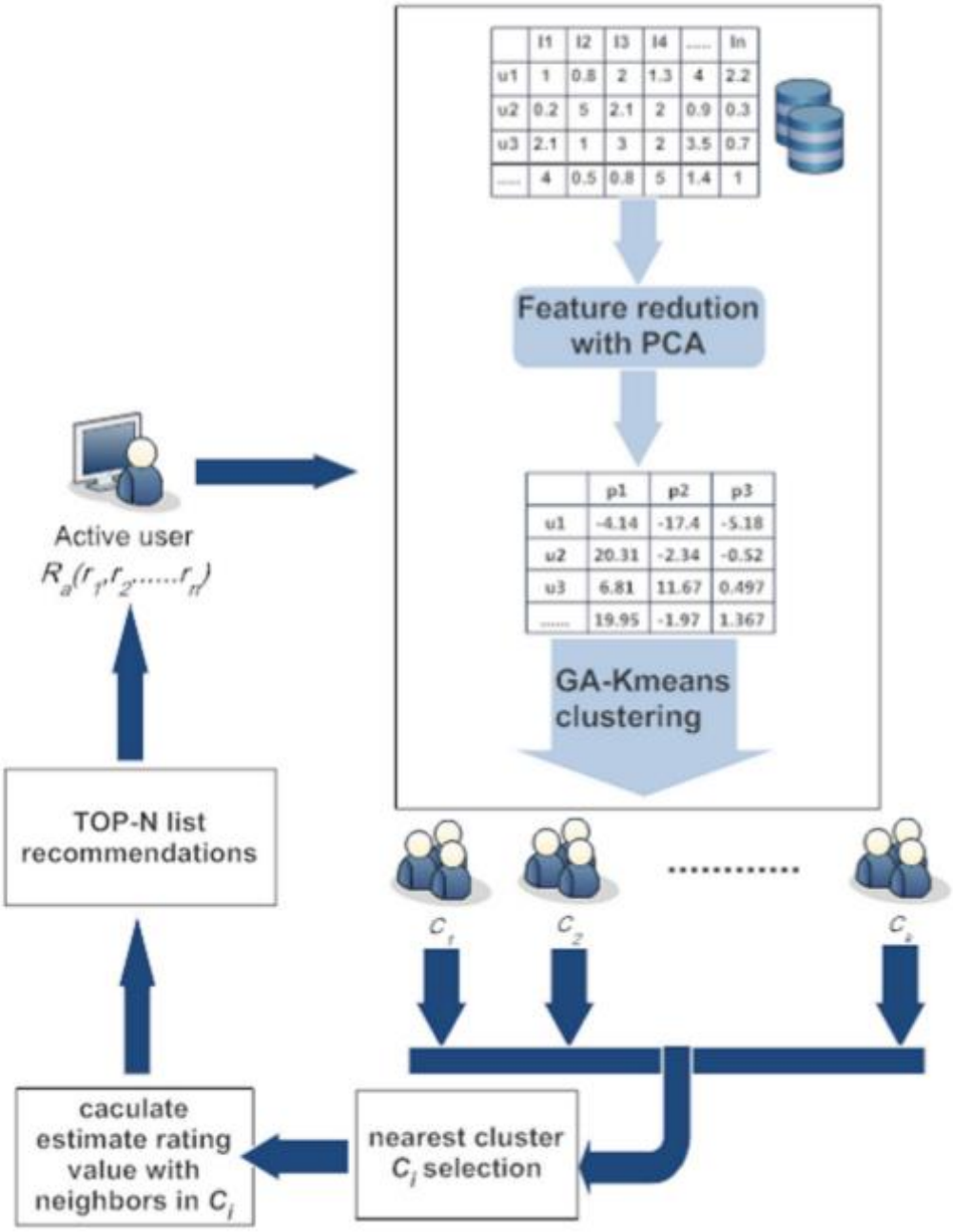
سامانه توصیه گر مبتنی بر مدل از نرخ‌بندی‌های کاربر – آیتم برای آموزش مدلی استفاده می‌کنند که برای تولید پیش‌بینی آنلاین به کار می‌رود. تکنیک‌های کاهش بعدیت و خوشه بندی اغلب در روش های مبتنی بر مدل برای ارزیابی مسئله پراکندگی داده ها به کار می روند. مسائل پراکندگی به دلیل ناکافی بودن امتیازات گذشته کاربران به وجود می آیند و در هنگام رشد کاربران و آیتم ها، پیچیده تر می شوند. به علاوه، مجموعه امتیازات با بعد بالا ممکن است استخراج کاربران را از طریق محاسبه مشابه مشکل کند که منجر به توصیه های ضعیف می شود. در تحقیقات انجام شده، سیستم های توصیه گر مبتنی بر مدل بسیاری وجود دارند که از طریق الگوریتم های خوشه بندی مزدوج مانند k-means و الگوریتم نقشه های خود سازمانده[۱۹] گسترش یافته اند. هدف خوشه بندی، تقسیم کاربران به گروه های مختلف جهت تشکیل همسایه های هم فکر (نزدیک موارد) به جای جستجوی کل فضای کاربر است که می تواند تا حد قابل ملاحظه ای مقیاس پذیری سیستم را بهبود دهد. ثابت شده است که سیستم های توصیه گر مبتنی بر خوشه بندی، نسبت به سیستم های مبتنی بر پالایش گروهی خالص، کیفیت پیش بینی و سودمندی بیشتری دارند.

در بسیاری از کارها، متدهای خوشه بندی با کل بعد های داده هدایت می شوند که می تواند منجر به عدم صحت شود و زمان محاسباتی بیشتری را صرف می کند. به‌طورکلی، انجام توصیه فیلم باکیفیت بالا هم چنان یک چالش هست و جستجوی یک متد خوشه‌بندی مؤثر و صحیح. مسئله مهمی در این موقعیت به شمار می‌رود.

برای مشخص نمودن چالش‌های ذکرشده، یک روش توصیه گر فیلم مبتنی بر مدل ترکیبی برای کم کردن مشکل بعدیت بالا و پراکندگی داده پیشنهاد می‌شود. در این پروژه، یک الگوریتم خوشه‌بندی بهینه برای تقسیم پروفایل‌های کاربری گسترش می‌دهیم که توسط بردارهای پروفایل متراکم‌تر پس از تبدیل تحصیل مؤلفه اساسی ارائه‌شده است. کل سیستم شامل دو فاز، یک‌فاز آنلاین و یک‌فاز آفلاین می‌باشد. در فاز آفلاین، مدل خوشه‌بندی در فضای بعد پایین آموزش داده می‌شود و برای هدف قرار دادن کاربران فعال در خوشه‌های مختلف آماده می‌شود. در فاز آنلاین، یک فهرست توصیه فیلم برای یک کاربر فعال از

[19] Self-Organize Map (SOM)

نرخ‌بندی‌های پیش‌بینی‌شده فیلم‌ها ارائه می‌شود. به‌علاوه، یک الگوریتم ژنتیک درروش جدید ما برای بهبود عملکرد خوشه‌بندی به کار گرفته می‌شود و الگوریتم خوشه‌بندی پیش نهادی با عنوان نامیده GA-KM می‌شود. سپس عملکرد روش پیش نهادی را در مجموعه داده movielensجستجو می‌کنیم. درزمینهٔ های صحت و دقت، نتایج آزمایش ثابت می‌کند که روش پیش نهادی قادر به ارائه توصیه‌های معتبر و با دقت می‌باشد.



شکل ۱-۳ فرآیند کل سیستم

۱-۳-سیستم‌های نظریه‌ای فیلم مبتنی بر پالایش گروهی

سیستم‌های توصیه گر فیلم مبتنی بر پالایش گروهی توسط پروژه Tapestry در سال ۱۹۹۲ معرفی شدند که یکی از موفق‌ترین سیستم‌های مدیریت اطلاعات محسوب می‌شود. کاربردهای عملی توصیه دهنده به

کاربر کمک می‌کند تا اطلاعات بدون استفاده گسترده را برای مواجهه باهم پوشانی اطلاعات فیلتر کند و پیشنهادات شخصی را ارائه کند. موفقیت عظیمی در بازرگانی الکترونیکی برای ایجاد دسترسی مشتری به محصولات ترجیح داده‌شده و بهبود سود تجارت وجود دارد. به‌علاوه، برای تقویت توانایی شخصی کردن، سیستم توصیه گر به‌طور گسترده‌ای در بیشتر وب‌سایت‌های چندرسانه‌ای برای هدف قرار دادن محصولات رسانه‌ای به مشتریان خاص به کار گرفته می‌شود. امروزه، پالایش گروهی، مؤثرترین تکنیک به کار گرفته‌شده توسط سیستم‌های توصیه گر فیلم است که مبتنی بر مکانیزم نزدیک‌ترین همسایه است. فرض می‌شود کسانی که دارای الگوی امتیازدهی مشابهی هستند و با احتمال ماکزیمم دارای سلایق یکسانی در آینده هستند. همه کاربران «هم‌فکر» با عنوان همسایه‌ها نامیده می‌شوند و از پایگاه داده امتیازات مشتق می‌شوند که به فیلم‌ها امتیاز داده‌اند. پیش‌بینی امتیازات داده نشده توسط کاربر هدف می‌تواند توسط شباهت اندازه‌گیری شده همسایه او استنباط شود.

تکنیک پالایش گروهی را به دودسته مهم سیستم‌های توصیه دهنده تقسیم می‌کند:

- پالایش گروهی مبتنی بر حافظه
- پالایش گروهی مبتنی بر مدل

پالایش گروهی مبتنی بر حافظه: بر روی فضای کل کاربر عمل می‌کند تا نزدیک‌ترین همسایه‌ها را برای یک کاربر فعال جستجو کرده و به‌طور اتوماتیک فهرستی از فیلم های پیش نهادی را برای ارائه فراهم نمایند. این متد پیچیدگی محاسباتی و پراکندگی داده دارد. برای حل مسائل کمبود حافظه و مسائل محاسباتی، یک پالایش گروهی مبتنی بر آیتم ای را پیش نهاد می دهند که در آن، ارتباطات بین آیتم ها برای تشکیل همسایگی برای یک آیتم هدف محاسبه می شوند. مطالعات تجربی ثابت می کند که روش مبتنی بر آیتم می تواند زمان محاسبه را کاهش دهند و صحت پیش بینی قابل مقایسه ای را ارائه دهند.

پالایش گروهی مبتنی بر مدل: مدلی پیش ساخت را برای ذخیره الگوهای امتیازات بر اساس پایگاه داده امتیازات کاربر توسعه می دهد که می تواند با مسائل پراکندگی و مقیاس پذیری مواجه شود. در زمینه های کیفیت توصیه، کاربردهای پالایش گروهی مبتنی بر مدل می تواند همانند موارد مبتنی بر حافظه عمل کنند. با این وجود روش های مبتنی بر مدل در ساخت و آموزش مدل آب لایه، زمان‌بر هستند که به‌سختی آپدیت می‌گردند.

۴-۱-نظریه شراکتی مبتنی بر خوشه‌بندی:

در سامانه توصیه گر فیلم، خوشه‌بندی یک روش پرکاربرد برای کم کردن مسئله مقیاس بندی است. بیشتر تحقیقات، مزایای ساختارهای پالایش گروهی مبتنی بر خوشه‌بندی را اثبات کرده‌اند. هدف الگوریتم‌های خوشه‌بندی، تقسیم اشیا بر خوشه‌هایی است که فاصله از بین اشیا موجود در یک خوشه مشابه را به حداقل برساند تا اشیاء مشابه شناسایی شوند. همانند یکی از متدهای پالایش گروهی مبتنی بر مدل، پالایش گروهی مبتنی بر خوشه‌بندی برای بهبود عملکرد نزدیک‌ترین همسایه[۲۰] از طریق پیش ساخت یک مدل آفلاین خوشه‌بندی مورداستفاده قرار می‌گیرد.

به‌طورمعمول، تعداد کاربران می‌تواند بر اساس شباهت امتیازات، به خوشه‌های مختلف گروه‌بندی شوند تا همسایگان هم‌فکر از طریق تکنیک های خوشه‌بندی پیدا شوند. سپس، پروسه خوشه‌بندی به‌صورت آفلاین برای ساخت مدل عمل می‌کند. وقتی کاربر هدف می‌رسد، ماژول آنلاین خوشه‌ای را با بالاترین وزن شباهت به آن را تعیین می‌کند و امتیازات پیش‌بینی آیتم تعیین‌شده، بر اساس اعضا خوشه مشابه، به‌جای جستجوی کل فضای کاربر، محاسبه می‌شود.

[20] K-NN

۲-فصل دوم الگوریتم‌های استفاده‌شده

۲-۱-تحلیل مؤلفه‌های اساسی

در این بخش از تکنیک استخراج ویژگی خطی برای تبدیل فضای با بعد بالای اصلی به فضای با بعد نسبتاً پایین استفاده می‌کنیم که حمل‌کننده اطلاعات متراکم‌تری می‌باشد. از آن جاییکه بعد بالای ماتریس امتیازدهی کاربر که بیشتر در ابتدا خالی است محاسبه شباهت را بسیار سخت می‌کند، روش ما با پروسه کاهش بعد مبتنی بر PCA شروع می‌شود. به‌عنوان یکی از موفق‌ترین تکنیک‌های استخراج ویژگی، PCA به طور گسترده‌ای در کاهش بعدی و پیش پر کردن داده‌های سیستم‌های فیلترینگ شراکتی مورداستفاده قرار می‌گیرد. ایده اصلی، تبدیل داده اصلی به فضای هماهنگ جدید است که از طریق جزء اصلی داده با بالاترین مقدار ویژه ارائه می‌شود. اولین بردار جزء اصلی، حمل‌کننده قابل‌ملاحظه‌ترین اطلاعات پس از مرتب کردن آن‌ها از بالا تا پایین می‌باشد. در حالت کلی، اجزا بااهمیت کمتر نادیده گرفته می‌شوند تا فضایی با بعدهای کمتر سنت به حالت اصلی تشکیل دهند. تصور کنید که ماتریس $m \times n$ امتیازات کاربران راداریم که در آن بردار هر سطر از این ماتریس ارائه دهند پروفایل کاربر است. پس از تجزیه مقدار ویژه، n جزء اصلی به دست می‌آید و ما فقط اولین d جزء آن را برای حفظ در فضای داده جدید انتخاب می‌کنیم که حاوی ۹۰ درصد تراکم داده اصلی می‌باشد. درنتیجه، بردارهای ویژگی کاهش‌یافته از PCA برای تغذیه با الگوریتم GA − KM جهت طبقه‌بندی آماده می‌شوند.

قبل از این‌که از آنالیز اجزا اصلی توصیفی به دست آوریم ابتدا به معرفی مفاهیمی ریاضی که در آنالیز اجزا اصلی استفاده می‌شود می‌پردازیم.

۲-۱-۱-انحراف معیار استاندارد:

برای فهمیدن انحراف معیار به یک مجموعه داده احتیاج داریم، آمارشناسان معمولاً علاقه‌مند به نمونه‌گیری از جامعه هستند. برای استفاده کردن از روش‌های نمونه‌گیری به‌عنوان‌مثال جامعه تمام مردم یک کشور است. درحالی‌که یک نمونه یک زیرمجموعه از جامعه است که آمارشناسان اندازه می‌گیرند.

مطلب مهم دیگر درباره آمار این است که اگر از سراسر جامعه استفاده می‌کنید فقط با اندازه‌گیری یک نمونه از جامعه شما می‌توانید با اندازه‌گیری احتمال (سنجش احتمال) کارکنید.

در این بخش آماری قصد داریم فرض کنیم که اطلاعات ما نمونه‌ای از جامعه است.

در اینجا یک مثال وجود دارد:

$$X = [1\ 2\ 4\ 6\ 12\ 15\ 25\ 45\ 68\ 67\ 65\ 98]$$

از علامت X برای اشاره به مجموعه اعداد استفاده می‌کنیم. اگر به یک عدد خاص در مجموعه داده‌ها بخواهیم اشاره‌کنیم از یک زیرنویس بر روی علامت استفاده کنیم که یک عدد خاص را نشان می‌دهد.

برای مثال ما میانگین نمونه‌ها را می‌توانیم حساب کنیم. چون با مفهوم میانگین نمونه‌ای آشنا هستیم فقط فرمول را ارائه می‌کنیم:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

برای مثال دو مجموعه داده مقابل دقیقاً میانگین یکسان ۱۰ دارند.

$$[0\ 8\ 12\ 20], [8\ 9\ 11\ 12]$$

اما تفاوت این دو مجموعه توزیع متفاوت اطلاعات است برای به دست آوردن چگونگی توزیع داده‌ها است از انحراف معیار استفاده می‌کنیم.

تعریف انحراف معیار: معدل فاصله از نقطه میانگین یک مجموعه داده.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

علامت S معمولاً برای نشان دادن انحراف معیار یک نمونه به کار می‌رود.

برای دو مجموعه بالا انحراف معیار در جداول زیر محاسبه‌شده‌اند:

جدول ۲-۱ انحراف معیار

X	$(X - \bar{X})$	$(X - \bar{X})^2$
0	-10	100
8	-2	4
12	2	4
20	10	100
Total		208
Divided by (n-1)		69.333
Square Root		8.3266

جدول ۲-۲ انحراف معیار ۲

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
8	-2	4
9	-1	1
11	1	1
12	2	4
Total		10
Divided by (n-1)		3.333
Square Root		1.8257

انتظار داریم مجموعه اول انحراف معیار بزرگ‌تری داشته باشد به این خاطر که داده‌ها از میانگین فاصله بیشتری دارند. مثال دیگر مجموعه داده‌های زیر میانگین و انحراف معیار ۱۰ دارند زیرا همه اعداد یکی هستند. هیچ‌کدام از آن‌ها از میانگین منحرف نمی‌شوند.

[10 10 10 10]

۲-۱-۲-واریانس[۲۱]:

واریانس معیار دیگری از پراکندگی مجموعه داده‌ها است. درواقع تقریباً با انحراف معیار برابر است. فرمول آن به‌صورت زیر است:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

با توجه هر دو علامت و فرمول متوجه می‌شوید که واریانس مربع انحراف معیار است.S^2 علامت معمولی برای واریانس یک نمونه است. هر دو این مقیاس از پراکندگی داده‌ها هستند. انحراف معیار مقیاس معمولی‌تری است؛ اما واریانس هم استفاده می‌شود.

۲-۱-۳-کوواریانس[۲۲]:

دو مقیاس آخر که ما به آن‌ها توجه داریم صرفاً کمی هستند. مجموعه داده‌ها مانند موارد زیر می‌تواند باشد: بلندی همه افراد در یک اتاق، نمره‌های آخرین امتحان و غیره؛ اما باوجوداین برای تعداد زیادی از مجموعه داده‌ها می‌تواند بیش از یک بعد وجود داشته باشد و هدف از تحلیل آماری این مجموعه داده‌ها معمولاً این است که ارتباطی که بین بعدها وجود دارد را بفهمیم. برای مثال ممکن است مجموعه داده‌هایمان هر دو بلندی همه دانش آموزان یک کلاس باشد.

---

[21] variance

[22] covariance

انحراف معیار و واریانس فقط بر روی یک بعد عمل می‌کنند. می‌توان انحراف معیار را به‌طور جداگانه برای هر بعد از مجموعه داده‌ها حساب کرد؛ اما نمی‌توان مقیاسی برای اندازه‌گیری اختلاف از میانگین نسبت به یکدیگر داشته باشیم. کوواریانس یک چنین مقداری است.

کوواریانس همیشه بین دو بعد اندازه‌گیری می‌شود. اگر کوواریانس را بین یک بعد و خودش حساب کنید درواقع همان واریانس به دست می‌آید. اگر شما یک سری داده سه‌بعدی (X,Y,Z) داشته باشید می‌توانید کوواریانس را بین دو بعد X,Y دو بعد X,Z و دو بعد Y,Z حساب کنید. اندازه‌گیری کوواریانس بین X,X یا Y,Y و یا Z,Z به شما واریانس بعدهای را به ترتیب می‌دهد.

فرمول محاسبه کوواریانس بسیار شبیه فرمول محاسبه واریانس است. فرمول محاسبه واریانس را نیز می‌توان مشابه این عبارت نوشت:

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

جمله درجه دوم نشان داده‌شده را به دو بخش بسط داده‌ایم زیرا این دانشی برای محاسبه کردن کوواریانس به ما می‌دهد.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

این دقیقاً همان فرمول واریانس است به‌جز آن‌که در دومین مجموعه از پرانتزها Y جایگزین X شده است.

تعریف کوواریانس: برای هر قلم داده تفاوت بین ارزش X و میانگین X را با تفاوت بین ارزش Y ضرب می‌کند و تقسیم‌بر n-1 می‌شود. فرض کنیم از یک گروه دانش‌آموز سؤال شده است که در درس خاصی چه نمره‌ای دریافت کرده‌اند و چه تعداد ساعت آن‌ها در کل صرف مطالعه کرده‌اند؛ بنابراین ما دو بعد داریم. اولین بعد، H تعداد ساعت مطالعه است و دومین بعد، M نمره کسب‌شده است. شکل زیر به ما اطلاعات فرضی را نشان می‌دهد؛ و cov(H,M) کوواریانس ساعت‌های مطالعه کردن و نمره گرفتن را محاسبه می‌کند.

جدول ۲-۳ واریانس

	Hours(H)	*Mark(M)*
Data	9	39
	15	56
	25	93
	14	61
	10	50
	18	75
	0	32
	16	85
	5	42
	19	70
	16	66
	20	80
Totals	167	749
Averages	13.92	62.42

بنابراین کوواریانس چه چیزی را نشان می‌دهد؟ اگر علامت کوواریانس مثبت باشد، نشان می‌دهد که هر دو بعد باهم افزایش می‌یابند، مثلاً افزایش ساعت مطالعه، نمره پایانی را افزایش می‌دهد. اگر علامت کوواریانس منفی باشد، هرگاه یک بعد افزایش یابد، بعد دیگر کاهش می‌یابد. پس آنچه به ما می‌گوید مخالف هم هستند که با افزایش ساعت مطالعه نمره پایانی کاهش می‌یابد. در بعضی موارد که کوواریانس صفر می‌شود نشان می‌دهد که دو بعد مستقل از هم هستند.

نتیجه‌ای که با افزایش نمره به ما می‌گوید مثلاً افزایش ساعت مطالعه می‌توان به‌آسانی با رسم یک نمودار از اطلاعات دید مانند شکل زیر:

جدول ۲-۴ کوواریانس

Covariance:

H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})$	$(H_i - \bar{H})(M_i - \bar{M})$
9	39	-4.92	-23.42	115.23
15	56	1.08	-6.42	-6.93
25	93	11.08	30.58	338.83
14	61	0.08	-1.42	-0.11
10	50	-3.92	-12.42	48.69
18	75	4.08	12.58	51.33
0	32	-13.92	-30.42	423.45
16	85	2.08	22.58	46.97
5	42	-8.92	-20.42	182.15
19	70	5.08	7.58	38.51
16	66	2.08	3.58	7.45
20	80	6.08	17.58	106.89
Total				1149.89
Average				104.54

ازآنجاکه علامت کوواریانس را بین هر دو بعد در مجموعه اطلاعات می‌توان حساب کرد این فن اغلب برای پیدا کردن ارتباط بین بعدها در ابعاد بزرگ مجموعه اطلاعات که تجسم آن مشکل است استفاده می‌شود.

۲-۱-۴- ماتریس کوواریانس:

می‌دانیم کوواریانس همیشه بین دو بعد اندازه‌گیری می‌شود. اگر مجموعه اطلاعاتی با بیش از دو بعد داشته باشیم، بیش از یک کوواریانس وجود دارد که می‌توان محاسبه کرد.

برای یک مجموعه اطلاعاتn بعدی می‌توان محاسبه کرد. یک پیکروش مفید برای به دست آوردن کوواریانس بین همه ابعاد این است که آن‌ها را محاسبه کرد و در یک ماتریس قرارداد. بنابراین ماتریس کوواریانس برای یک مجموعه از داده‌ها باn بعد به‌صورت زیر است:

$$C^{m\times n} = (c_{i,j},\ c_{i,j} = cov(Dim_i, Dim_j)),$$

که $C^{n\times n}$ یک ماتریس با n سطر و n ستون است. اگر ما یک مجموعه اطلاعات n بعدی داشته باشیم، می‌توان یک ماتریس مربعی داشته باشیم؛ که هر عنصر ماتریس نتیجه‌ای از محاسبه کوواریانس بین دو بعد جدا است.

یک مثال: ماتریس کوواریانس را برای یک مجموعه داده ۳ بعدی فرضی کنید. از ابعاد معمول X,Y,Z استفاده می‌کنیم. پس ماتریس کوواریانس ۳ ستون و ۳ سطر دارد و ارزش آن‌ها به‌صورت زیر است:

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

توجه به چند نکته لازم است، روی قطر شما می‌بینید که مقدار کوواریانس بین یکی از ابعاد و خودش است در حقیقت واریانس آن بعد است. نکته دیگر این‌که cov(a,b)=cov(b,a) ، ماتریسی متقارن مجاور با قطر اصلی است.

۲-۱-۵- بردارهای ویژه:

همان‌طور که می‌دانید دو ماتریس که اندازه‌های آن‌ها سازگار است را می‌توان در هم ضرب کرد. بردارهای ویژه یک مورد خاص از این مورد هستند. دو مورد ضرب از بین یک ماتریس و یک بردار را در مثال‌های زیر

ملاحظه می‌کنید. در مثال اول، نتیجه بردار یک مضرب صحیحی از بردار اصلی نیست درصورتی‌که در مثال دوم دقیقاً ۴ برابر بردار اصلی است ($\binom{3}{2}$بردار اصلی است).

$$2 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

بردار ویژه فقط می‌تواند برای ماتریس مربعی وجود داشته باشد؛ و هر ماتریس مربعی بردار مشخصه ندارد. ماتریس n×n معلوم n بردار ویژه برای آن وجود دارد. یک ماتریس ۳×۳ معلوم، ۳ بردار مشخصه دارد. خاصیت دیگر بردارهای مشخصه این است. اگر اندازه یک بردارمقداری کم باشد همه ما در حال انجام بلندتر ساختن آن هستیم اما مسیر آن را تغییر نمی‌دهیم. در ضمن تمام بردارهای مشخصه بر هم عمودند، مهم نیست شما چند بعد دارید. اطلاعات را می‌توان به ازای این بردارهای ویژه عمودی بیان کرد، به‌جای این‌که آن‌ها را به ازای محورهای X,Y بیان کرد که در قسمت آنالیز اجزای اصلی انجام می‌دهیم. مطلب مهم دیگری که باید دانست این است زمانی که بردار مشخصه را پیدا می‌کنیم علاقه‌مند به این هستیم که بردارهای مشخصه که طول واحد دارند پیدا کنیم طول یک بردار اثری بر روی این‌که آیا یک بردار مشخصه است یا نه نمی‌گذارد. به‌منظور این‌که بردارهای مشخصه استاندارد را پیدا کنیم هر زمان که بردار مشخصه را به دست می‌آوریم، معمولاً مقیاسی برای این‌که طول واحد داشته باشیم به دست می‌آوریم؛ بنابراین همه بردارهای مشخصه باید طول یکسان داشته باشند.

بردار$\binom{3}{2}$ یک بردار مشخصه و طولش به این صورت$\sqrt{3^2 + 2^2} = \sqrt{13}$ است بنابراین ما بردار اصلی برای ساختن برداری با طول یک تقسیم‌بر $\sqrt{13}$می‌کنیم.

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \div \sqrt{13} = \begin{pmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{pmatrix}$$

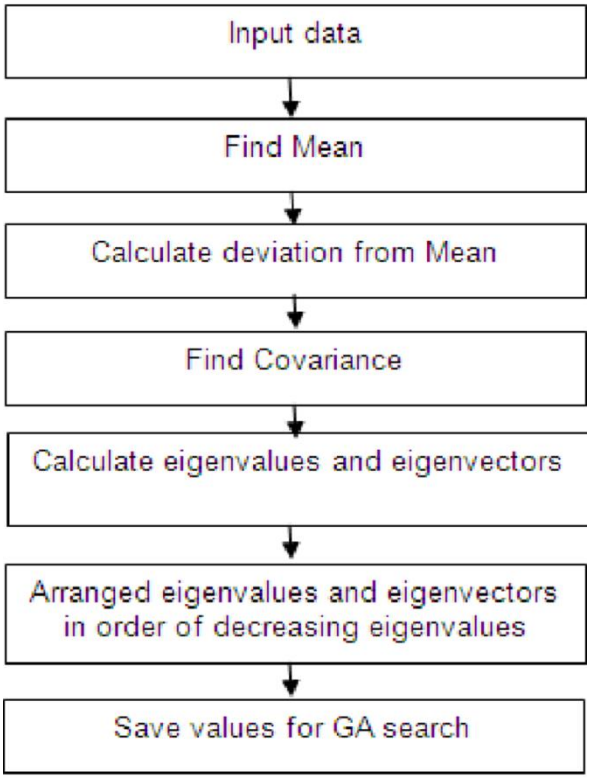
متأسفانه فقط برای ماتریس‌های نسبتاً کوچک آسان است به ماتریس‌های بزرگ‌تر از ۳×۳ علاقه نداریم. راه معمولی برای پیدا کردن بردارهای مشخصه با تعدادی روش‌های پیچیده تکراری است که در حوصله این مطلب نیست.

۲-۱-۶- مقادیر ویژه:

مقادیر ویژه بسیار وابسته به بردارهای مشخصه هستند. در حقیقت یک مقدار ویژه را در شکل‌های قبل دیدیم. در هر دو مثال‌ها، بعد ضرب‌در ماتریس مربعی نتیجه به‌دست‌آمده چند برابر بردار ویژه، این عدد مقدار ویژه مربوط به بردار مشخصه است. در مثال اول، ارزش ۴ بود؛ و ۴ مقدار ویژه وابسته به این بردار ویژه است. می‌بینیم که مقادیر ویژه و بردارهای ویژه باهم جفت هستند. زمانی که شما روشی برای محاسبه بردارهای ویژه به دست می‌آورید معمولاً به‌طور تمام و کمال به مقادیر ویژه هم می‌رسید.

۲-۱-۷- آنالیز اجزای اصلی:

این روش یکی از الگوهای تشخیص و شناسایی (تشخیص هویت) در یک مجموعه اطلاعات است. در این روش اطلاعات را بر اساس شباهت‌ها و تفاوت‌هایشان بیان می‌کنند. ازآنجاکه در اطلاعات از ابعاد بالا، نقشه و طرح خاصی را به‌سختی می‌توان در داده‌ها پیدا کرد در حقیقت آنالیز اجزای اصلی ارتباط بین داده‌ها را کشف می‌کند؛ و درجایی که نعمت نمایش گرافیکی در دسترس نیست، آنالیز اجزای اصلی یک ابزار نیرومند برای آنالیز اطلاعات است. دیگر مزیت اصلی آنالیز اجزای اصلی این است که شما یک‌بار این الگو را درداده تا پیدا می‌کنید و این اطلاعات را فشرده می‌کنید. با کاهش تعداد ابعاد بدون آن‌که مقدار زیادی از اطلاعات را از دست نمی‌دهید. هدف آنالیز اجزای اصلی خلاصه کردن داده‌ها است و به‌عنوان یک وسیله دسته کننده اطلاعات موردتوجه نیست. از این فن در فشرده‌سازی تصاویر استفاده می‌شود. در این قسمت مراحل موردنیاز برای اجرا کردن آنالیز اجزای اصلی در یک مجموعه داده را نشان خواهیم داد. تلاش می‌کنیم با یک مثال شرحی که برای هر نقطه استفاده می‌شود را فراهم کنیم.



شکل ۲-۱ فلوچارت الگوریتم تحلیل مؤلفه اساسی

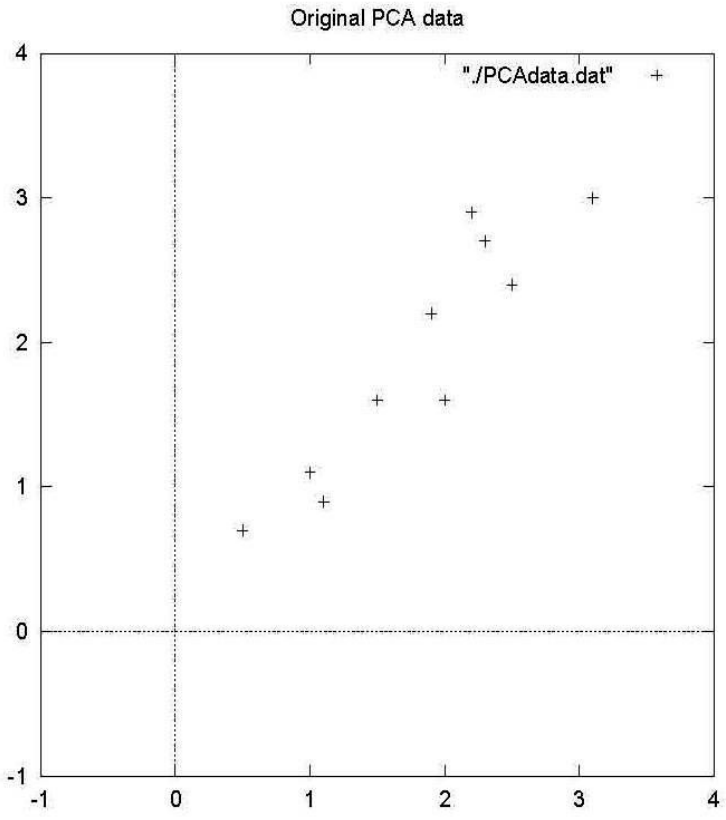
مرحله ۱: به دست آوردن اطلاعات

ما در اینجا از یک مثال ساده از یک مجموعه اطلاعات فرضی استفاده می‌کنیم. این اطلاعات فقط ۲ بعدی است. دلیل انتخاب دو بعد نمایش و رسم داده‌ها برای نشان دادن این است که آنالیز اجزای اصلی در هر مرحله چه‌کاری انجام می‌دهد.

جدول ۲-۵ داده‌های خام

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



شکل ۲-۲ داده‌ی اصلی

مرحله ۲: میانگین را کم کنید

برای این‌که آنالیز اجزای اصلی به‌طور صحیح کار کند شما باید میانگین را در هر بعد از داده‌ها کم کنید؛ که همان نرمال کردن داده است. کاستن میانگین یعنی میانگین سراسر هر بعد از مجموعه داده‌ها را کم کنیم؛ بنابراین در اینجا از همه X ها میانگین X کم شده است و از همه Y ها میانگین Y کم شده است؛ و میانگین هر بعد صفر می‌شود.

جدول ۲-۶ داده‌های تنظیم

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

مرحله ۳: محاسبه کردن ماتریس کوواریانس

ازآنجاکه در مثال ما اطلاعات ۲ بعدی است ماتریس کوواریانس باید ۲×۲ باشد. در زیر فقط نتایج را نشان داده‌ایم.

$$cov = \begin{pmatrix} .6165555556 & .615444444 \\ .615444444 & .7165555556 \end{pmatrix}$$

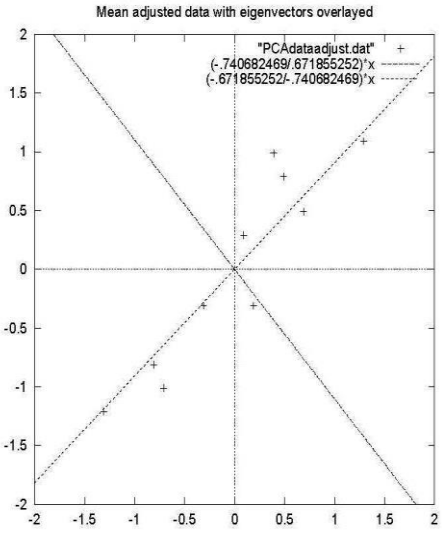
ازآنجاکه عناصر غیر قطری در این ماتریس کوواریانس مثبت است ما باید انتظار داشته باشیم که هر دو متغیر x,y باهم افزایش یابند.

مرحله ۴: بردارهای مشخصه و مقادیر ویژه را از ماتریس کوواریانس محاسبه کنید.

ازآنجاکه ماتریس کوواریانس مربعی است ما می‌توانیم بردارهای مشخصه و مقادیر ویژه را برای این ماتریس حساب کنیم. بردارهای مشخصه و مقادیر ویژه اطلاعات مفیدی درباره داده‌هایمان به ما می‌گویند. دلیل آن را به‌خوبی نشان خواهیم داد. در ضمن در این مثال بردارهای ویژه و مقادیر ویژه به‌صورت زیر است:

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

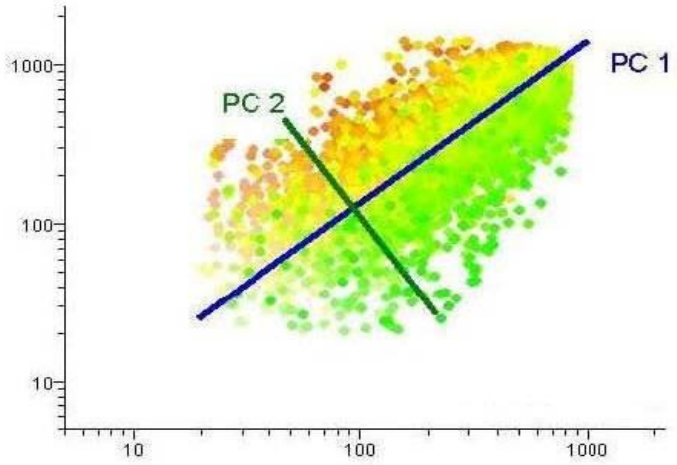
$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$



شکل ۲-۳ محورهای جدید

شما می‌توانید ببینید یک الگو قوی چگونه اطلاعاتی دارد. همان‌طور که از ماتریس کوواریانس انتظار می‌رفت آن دو متغیر باهم افزایش می‌یابند. در شکل هر دو بردار مشخصه رسم شده‌اند. آن‌ها مثل خطوط نقطه‌دار مورب روی نقشه ظاهرشده‌اند. بردارهای مشخصه که بر هم عمود بودند در اینجا نیز به‌آسانی دیده می‌شوند؛ اما این بردارهای مشخصه برای ما اطلاعاتی درباره داده‌هایمان فراهم می‌آورند. همان‌طور که می‌بینیم یکی از بردارهای مشخصه از مرکز داده‌ها می‌گذرد و بیشتر اطلاعات در حول‌وحوش این بردار مشخصه است. این بردار مشخصه به ما نشان می‌دهد که چگونه این دو مجموعه اطلاعات در امتداد خطوط به هم وابسته هستند. بردار مشخصه دوم به ما خط دیگری را می‌دهد که اهمیت کمتری دارد.

در الگو داده‌ها همه نقاط از خطی (یک بردار مشخصه) پیروی می‌کنند اما با مقدار کمی فاصله از گوشه خط حرکت می‌کنند. بنا براین به‌وسیله این فرآیند بردارهای مشخصه ماتریس کوواریانس را می‌توان به دست آورد، ما قادریم این خطوط را از داده‌ها استخراج‌کنیم. بقیه مراحل شامل تغییر شکل دادن داده‌ها است. در زیر یک‌شکل دیگر برای فهم بهتر این موضوع وجود دارد.



شکل ۲-۴ متراکم سازی در محورهای جدید

در اینجا فکر متراکم سازی داده‌ها و کاهش ابعاد به میان می‌آید. اگر به بردارهای مشخصه و مقادیر ویژه در بخش قبلی نگاه کنید. به این نتیجه می‌توان دست‌یافت که بردارهای مشخصه ارزش‌های کاملاً متفاوتی دارند. درواقع، بردار مشخصه با بزرگ‌ترین مقدار ویژه جز اساسی مجموعه داده‌ها است.

در مثال، بردار مشخصه با مقدار ویژه بزرگ آن نقطه پایینی وسط داده‌ها است. این پراهمیت‌ترین ارتباط بین ابعاد داده‌ها است.

عموماً، یک بردار ویژه از ماتریس کوواریانس به دست می‌آید و در مرحله بعدی طبق دستور به‌وسیله مقادیر ویژه از بلندترین به کوچک‌ترین مرتب می‌کنید. این به شما اجزای پراهمیت را می‌دهد.

حالا می‌توان از اجزای کم‌اهمیت‌تر چشم‌پوشی کرد زیرا مقدار کمی از اطلاعات را از دست می‌دهیم اگر مقدار ویژه حذف‌شده کوچک باشد اطلاعات زیادی را از دست نمی‌رود.

سرانجام مجموعه داده‌ها ابعاد کمتری نسبت به داده‌های اصلی به دست می‌آید. اگر ابعاد اصلی داده‌ها n بعدی است بنابراین n بردار مشخصه و مقدار ویژه را حساب کنید و اگر p بردار مشخصه اول را انتخاب کنیم، سرانجام مجموعه داده‌های فقط p بعد دارد.

مرحله ۶: مجموعه داده‌های جدید مشتق شده

این مرحله نهایی در آنالیز اجزای اصلی است و هم‌چنین آسان‌ترین قسمت است وقتی‌که ما جزای اصلی را انتخاب می‌کنیم (بردارهای ویژه). ما می‌خواهیم داده‌هایمان را حفظ و به شکل یک بردار ویژگی نشان دهیم و ما به‌سادگی ترانهاده بردار را به دست آورده و این را در سمت چپ ترانهاده مجموعه داده‌های اصلی ضرب می‌کنیم.

$$FinalData = RowFeature * RowDataAdjust$$

بردار ویژگی سطری[۲۳] ترانهاده ماتریس ستونی از بردارهای مشخصه هست، پس یک ماتریس سطری از بردارهای مشخصه است؛ که در آن بردارهای مشخصه پراهمیت‌تر در بالا قرار دارند؛ و داده‌های تعدیل‌شده سطری[۲۴] ترانهاده اطلاعات تعدیل‌شده است. در هر ستون از مجموعه داده‌ها قرار دارد و هر سطر یک بعد جداگانه را نگهدار می‌کند.

---

[23] RowFeatureVectore

[24] RowDataAdjust

داده نهایی[25]: مجموعه داده‌های نهایی با مجموع داده‌ها در ستون‌ها و همراه ابعاد سطرها است. مجموعه داده‌های اصلی ما دو محور دارد (X,Y) ؛ بنابراین داده‌های ما به ازای آن‌ها هستند. ممکن است شما دوست داشته باشید که اطلاعات اصلی را به ازای هر دو محور بیان کنید. اگرچه این محورها بر هم عمود هستند ولی این بیان بسیار مؤثر است. به این دلیل مهم است که بردارهای مشخصه همیشه قائم بر یکدیگر هستند.

فرم اطلاعاتمان را تغییر داده‌ایم. حال آن‌ها به ازای ۲ بردار ویژه هستند.X,Y به ازای محورهای

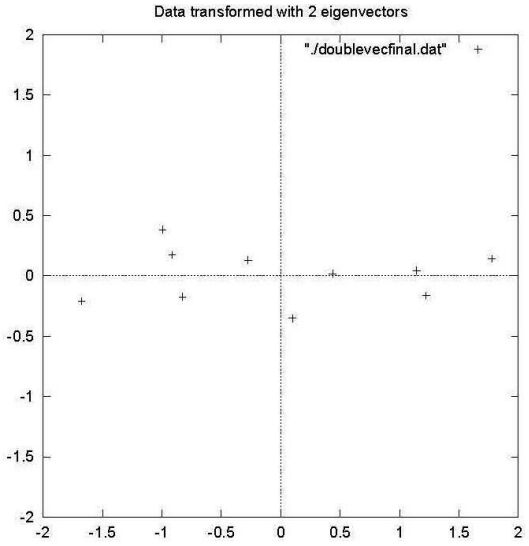
در شکل داده‌های نهایی و رسم آن با استفاده از هر دو بردار مشخصه را نشان می‌دهد:

جدول ۲-۷ داده‌های تبدیل‌شده

Transformed Data=

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

[25] FinalData



شکل ۲-۵ داده‌های تبدیل‌شده

تبدیل دیگری را می‌توانیم فقط با گرفتن بردار مشخصه با بزرگ‌ترین مقدار به دست آوریم. جدولی از اطلاعات در شکل بعدی تشکیل‌شده است. انتظار می‌رود فقط یک تک بعد باشد. اگر مجموعه داده‌ها را با شکل نتایج استفاده‌شده از هر دو بردار مشخصه مقایسه کنیم متوجه خواهیم شد که این مجموعه داده دقیقاً ستون اول مجموعه داده قبلی است؛ بنابراین اگر این مجموعه داده را رسم کنیم چون داده‌ها یک‌بعدی هستند دقیقاً موقعیت نقاط محور X در شکل قبلی است. ما تمام محور دیگر را کنار گذاشته‌ایم که مربوط به بردار مشخصه دیگر است.

شکل زیر داده‌های تعدیل‌شده با استفاده از بردار مشخصه پراهمیت‌تر را نشان می‌دهد:

جدول ۲-۸ نتیجه داده‌ها با ابعاد کمتر

Transformed Data (Single eigenvector)
x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

۲-۲- خوشه‌بندی k-means

خوشه‌بندی یکی از مهم‌ترین مسائل در حوزه‌ی یادگیری بدون ناظر است؛ مانند هر مسئله‌ی دیگر از این نوع، با یافتن یک ساختار در یک مجموعه داده‌ی بدون برچسب سروکار دارد. به‌طور غیررسمی، فرآیند سازمان‌دهی اشیا در چند دسته به‌طوری‌که اعضای هر دسته از جنبه‌هایی به هم شبیه باشند را خوشه‌بندی گویند. با این تعریف، یک خوشه مجموعه‌ای از اشیا است که به هم شبیه‌اند و با اشیای مربوط به دیگر خوشه‌ها متفاوت‌اند. هدف خوشه‌بندی، شناسایی دسته‌های طبیعی در یک مجموعه از اشیای برچسب نخورده است.

تقسیم‌بندی روش‌های خوشه‌بندی:

تاکنون الگوریتم‌های فراوانی برای خوشه‌بندی داده‌ها معرفی‌شده است باوجود گوناگونی روش‌های خوشه‌بندی، هنوز روشی یکتایی وجود ندارد که بتواند تمام انواع خوشه‌ها را به‌خوبی شناسایی کند؛ ازاین‌رو، این کاربر است که باید با توجه به نیازهایش روش مناسب را برگزیند. تقسیم‌بندی‌های گوناگونی برای روش‌های خوشه‌بندی وجود دارد:

- سلسله مراتبی و افراز بندی
- انحصاری و غیرانحصاری
- فازی و غیرفلزی
- جزی و کامل

از این میان، تقسیم‌بندی روش‌های خوشه‌بندی به دو نوع سلسله‌مراتبی و افراز بندی یا تودرتو و غیر تودرتو بیش از موارد دیگر موردتوجه است. در خوشه‌بندی افراز بندی، با مجموعه‌ای از خوشه‌ها سروکار داریم که روی‌هم افتادگی ندارند و هر شیء تنها به یک خوشه‌تعلق دارد. از سوی دیگر، در خوشه‌بندی سلسله‌مراتبی، خوشه‌ها به‌صورت تودرتو سازمان می‌یابند و تشکیل یک ساختار درختی می‌دهند.

افراز بندی:

همان‌طور که گفته شد، در خوشه‌بندی افراز بندی با مجموعه‌ای از خوشه‌ها سروکار داریم که روی‌هم افتادگی ندارند و هر شیء تنها به یک خوشه‌تعلق دارد. هدف از خوشه‌بندی افراز بندی، تقسیم داده‌ها به‌گونه‌ای است که داده‌های درون یک خوشه بیشترین شباهت را به هم داشته باشند و از سوی دیگر، بیشترین فاصله‌ها را با داده‌های موجود در خوشه‌های دیگر داشته باشند. الگوریتم‌های Forgy، Isodatو-K means چند نمونه از روش‌های خوشه‌بندی افراز بندی هستند.

۲-۲-۱- بررسی K-means

الگوریتم k means یکی از معمول‌ترین روش‌های خوشه‌بندی مورداستفاده به دلیل سادگی، انعطاف‌پذیری و محاسبه مؤثر به‌خصوص با در نظر گرفتن مقدار داده زیاد است. k means به‌طور تکرارشونده‌ای، مراکز خوشه k را برای تعیین اشیا در نزدیک‌ترین خوشه بر طبق محاسبه فاصله، محاسبه می‌کند. وقتی نقاط مرکزی تغییر درگیری انجام ندادند، الگوریتم خوشه‌بندی به همگرایی می‌رسد. بااین‌وجود k means فاقد توانایی انتخاب دانه اولیه صحیح است و ممکن است منجر به عدم صحت طبقه‌بندی گردد. انتخاب رندوم دانه اولیه می‌تواند منجر به راه‌حل بهینه مکانی شود که برای حالت بهینه کلی، فرعی محسوب می‌شود. به بیانی دیگر، دانه‌های اولیه مختلف در حال پیش روی بر روی دسته داده مشابه ممکن است نتایج قسمت‌بندی متفاوتی ایجاد نمایند. در زیر شبه کد این الگوریتم معرفی‌شده است.

1: Initialization: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids do not change

شکل ۲-۶- سودو کد خوشه‌بندی

با ارائه دسته‌ای از اشیا ($x_1, x_2, ..., x_x$) در جاییکه هر شیء یک بردار بعدی m هست، هدف الگوریتم k means قسمت‌بندی این اشیا به گروه‌های k، به‌طور خودکار می‌باشد. به‌طورمعمول، پروسه شامل مراحل زیر است:

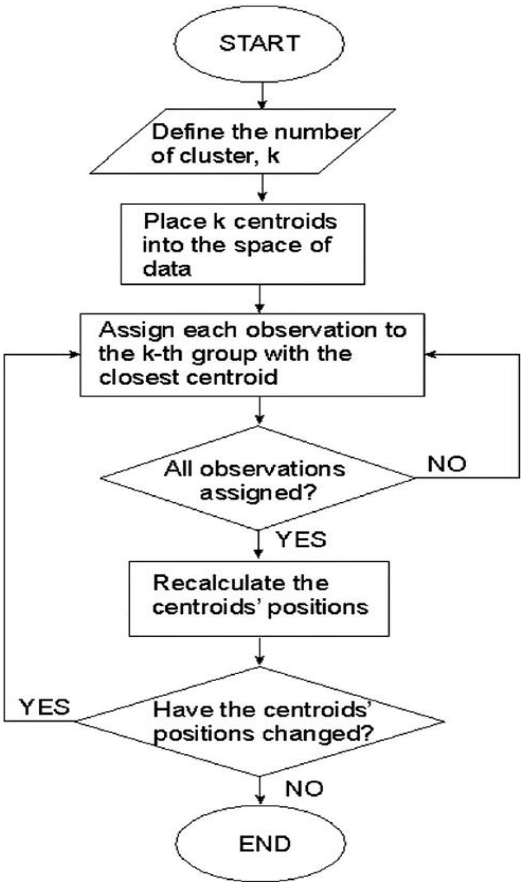
۱ مرکز خوشه اولیه k را انتخاب کنید $C_j, j = 1, 2, 3,...,k$

۲ هر x_i در نزدیک‌ترین مرکز خوشه بر طبق اندازه‌گیری فاصله تعیین می‌شود.

۳ مجموع فواصل مربع از همه اجزا در یک خوشه را محاسبه کنید:

۴ اگر تغییر دیگری وجود ندارد، الگوریتم همگرا شده است و کار خوشه‌بندی پایان گرفته است. در غیر این صورت، M_j مربوط به خوشه‌های k را به‌عنوان مراکز خوشه جدید دوباره محاسبه کنید و به مرحله ۲ بروید.

فلوچارت الگوریتم K-means:



شکل ۲-۷- فلوچارت الگوریتم خوشه‌بندی

در الگوریتم K-means می‌توان از معیارهای فاصله‌ی گوناگون بهره گرفت و خوبی یا بدی به‌کارگیری آن معیار بستگی به نوع داده‌هایی دارد که باید خوشه‌بندی شوند. تابع زیر به‌عنوان تابع هدف مطرح است که $\| \|$ معیار فاصله بین نقاط و c_j مرکز خوشهٔ j ام است.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

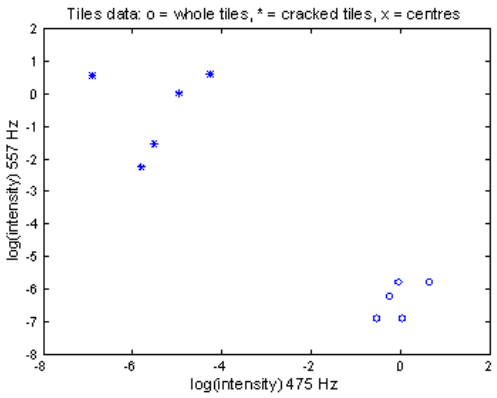
مشکلات روش خوشه‌بندي K-Means

علي‌رغم اينكه خاتمه پذیری الگوریتم بالا تضمین‌شده است ولي جواب نهايي آن واحد نبوده و همواره جوابي بهينه نيست. به‌طوركلی روش ساده بالا داراي مشكلات زير است.

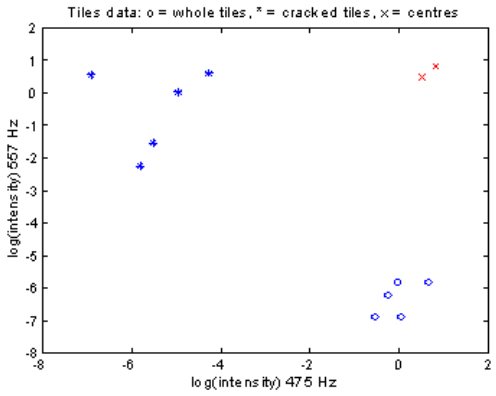
- جواب نهايي به انتخاب خوشه‌هاي اوليه وابستگي دارد.
- روالي مشخص براي محاسبه اوليه مراكز خوشه‌ها وجود ندارد.
- تعيين تعداد خوشه‌ها و صفر شدن خوشه‌ها می‌باشد.
- اگر در تكراري از الگوريتم تعداد داده‌هاي متعلق به خوشه‌اي صفر شد راهي براي تغيير و بهبود ادامه روش وجود ندارد.
- برای غلبه بر محدودیت‌های بالا، الگوریتم ژنتیکی را برای ترکیب با پروسه خوشه‌بندی $k-$ means جهت تقویت کیفیت طبقه‌بندی پیرامون یک k خاص معرفی می‌کنیم.

مثالي براي روش خوشه‌بندي K-Means:

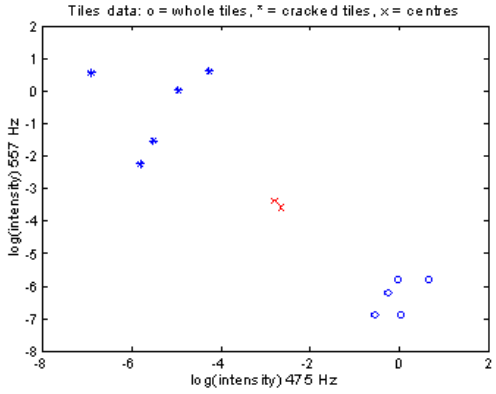
در شكل زير نحوه اعمال اين الگوريتم خوشه‌بندي روي يک مجموعه داده كه شامل دو گروه داده است نشان داده‌شده است. يک گروه از داده‌ها با ستاره و گروه ديگر با دايره مشخص‌شده‌اند (a). در مرحله اول نقطه‌اي به‌عنوان مركز خوشه‌ها انتخاب‌شده‌اند كه بارنگ قرمز نشان داده‌شده‌اند (b). سپس در مرحله دوم هر يک از نمونه داده‌ها به يكي از اين دو خوشه نسبت داده‌شده است و براي هر دسته جديد مركزي جديد محاسبه‌شده است كه در قسمت c نشان داده‌شده‌اند. اين روال تا رسيدن به نقاطي كه ديگر تغيير نمي‌كنند، ادامه پيداكرده است.



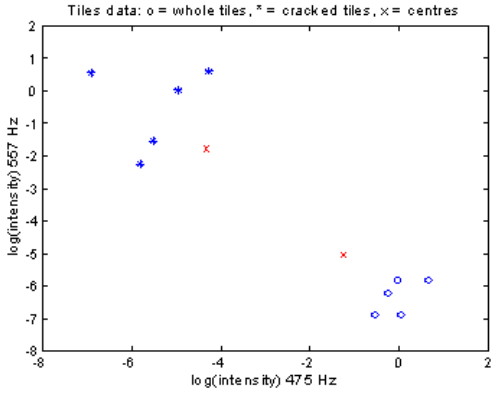
شكل ۲-۸- شكل ۱ مثال خوشه‌بندی



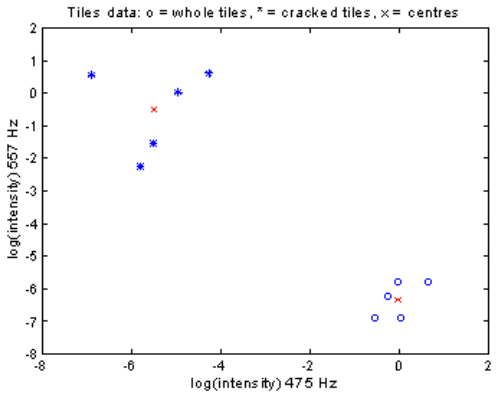
شکل ۲-۹- شکل ۲ مثال خوشه‌بندی



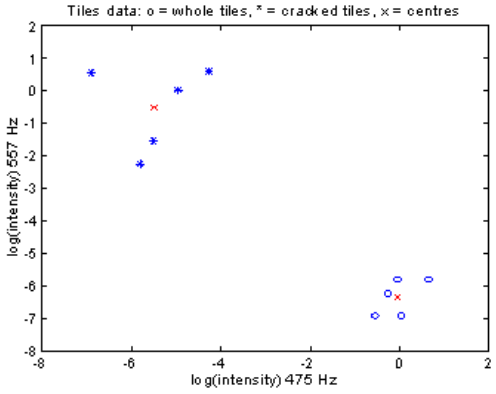
شکل ۲-۱۰- شکل ۳ مثال خوشه‌بندی



شکل ۲-۱۱- شکل ۴ مثال خوشه‌بندی



شکل ۲-۱۲- شکل ۵ مثال خوشه‌بندی



شکل ۲-۱۳- شکل ۶ مثال خوشه‌بندی

۲-۳- الگوریتم ژنتیک

محدوده کاري الگوريتم ژنتيک بسيار وسيع هست و هرروز با پيشرفت روزافزون علوم و فنّاوری استفاده از اين روش در بهينه‌سازی و حل مسائل بسيار گسترش‌يافته است. الگوريتم ژنتيک يکي از زيرمجموعه‌های محاسبات تکامل‌يافته می‌باشد که رابطه مستقيمي با مبحث هوش مصنوعي دارد. درواقع الگوريتم ژنتيک يکي از زيرمجموعه‌های هوش مصنوعي می‌باشد. الگوريتم ژنتيک را مي‌توان يک روش جستجوي کلي ناميد که از قوانين تکامل بيولوژيک طبيعي تقليد مي‌کند. الگوريتم ژنتيک بر روی يکسري از جواب‌هاي مسئله، به اميد به دست آوردن جواب‌های بهتر قانون بقاي بهترين را اعمال می‌کند. در هر نسل به کمک فرآيند انتخابي متناسب باارزش جواب‌ها و توليدمثل جواب‌های انتخاب‌شده به کمک عملگرهايي که از ژنتيک

طبيعي تقليدشده‌اند، تقريب‌هاي بهتري از جواب نهايي به دست مي‌آيد. اين فرايند باعث مي‌شود كه نسل‌های جدید با شرایط مسئله سازگارتر باشد.

۲-۳-۱- تاريخچه

حساب تكاملي، براي اولين بار در سال ۱۹۶۰ توسط آقاي ريچنبرگ ارائه شد كه تحقيق وي در مورد استراتژي تكامل بود. بعدها نظريه او توسط محققان زيادي موردبررسی قرار گرفت تا اينكه الگوريتم ژنتيک[۲۶] توسط جان هولند[۲۷] و در سال ۱۹۷۵ در دانشگاه ميشيگان، ارائه شد. در سال ۱۹۹۲ نيز جان كوزا[۲۸] از الگوريتم ژنتيک براي حل و بهينه‌سازی مسائل مهندسي پيشرفته استفاده كرد و توانست براي اولين بار روند الگوريتم ژنتيک را به زبان كامپيوتر درآورد و براي آن‌يک زبان برنامه‌نويسی ابداع كند كه به اين روش برنامه‌نويسی، برنامه‌نويسی ژنتيک گويند و نرم‌افزاری كه توسط وي ابداع گرديد به نرم‌افزار LISP مشهور است كه هم‌اكنون نيز اين نرم‌افزار كاربرد زيادي در حل و بهينه‌سازی مسائل مهندسي پيداكرده است.

۲-۳-۲- تاريخچه بيولوژيكي

بدن هر موجود زنده‌ای از سلول تشكيل يافته است و هر سلول هم از كروموزوم تشكيل يافته است. كروموزوم‌ها نيز از رشته‌های DNA تشكيل يافته‌اند. كروموزوم‌ها هم از ژن تشكيل يافته‌اند؛ و به هر بلوک DNA يک ژن می‌گويند و هر ژن نيز از يک پروتئين خاص و منحصربه‌فرد تشكيل يافته است؛ و به مجموعه از ژن‌ها يک ژنوم[۲۹] می‌گويند.

۲-۳-۳- ساختار الگوريتم‌هاي ژنتيكي

به‌طوركلی، الگوريتم‌هاي ژنتيكي از اجزاء زير تشكيل مي‌شوند:

---

26 Genetic Algorithm-GA

27 John Holland

28 John Koza

29 Genome

۲-۳-۴-کروموزوم[۳۰]

در الگوريتم‌هاي ژنتيكي، هر كروموزوم نشان‌دهنده يك نقطه در فضاي جستجو و يك راه‌حل ممكن براي مسئله موردنظر است. خود كروموزوم‌ها (راه‌حل‌ها) از تعداد ثابتي ژن[۳۱] (متغير) تشكيل مي‌شوند. براي نمايش كروموزوم‌ها، معمولاً از كدگذاري‌هاي دودويي (رشته‌هاي بيتي) استفاده مي‌شود.

۲-۳-۵-جمعيت

مجموعه‌اي از كروموزوم‌ها يك جمعيت را تشكيل مي‌دهند. با تأثير عملگرهاي ژنتيكي بر روي هر جمعيت، جمعيت جديدي با همان تعداد كروموزوم تشكيل مي‌شود.

۲-۳-۶- تابع برازندگي[۳۲]

به‌منظور حل هر مسئله با استفاده از الگوريتم‌هاي ژنتيكي، ابتدا بايد يك تابع برازندگي براي آن مسئله ابداع شود. براي هر كروموزوم، اين تابع عددي غير منفي را برمي‌گرداند كه نشان‌دهنده شايستگي يا توانايي فردي آن كروموزوم است.

---

[30] Chromosome

[31] Gene

[32] Fitness Function

۲-۳-۷- عملگرهاي الگوريتم ژنتيك

در الگوريتم‌هاي ژنتيكي، در طي مرحله توليدمثل[۳۳] از عملگرهای ژنتيكي استفاده مي‌شود. با تأثير اين عملگرها بر روي يك جمعيت، نسل[۳۴] بعدي آن جمعيت توليد مي‌شود. عملگرهاي انتخاب[۳۵], آميزش[۳۶] و جهش[۳۷] معمولاً بيشترين كاربرد را در الگوريتم‌هاي ژنتيكي دارند.

۲-۳-۸- عملگر انتخاب[۳۸]:

اين عملگر از بين كروموزوم‌هاي موجود در يك جمعيت، تعدادي كروموزوم را براي توليدمثل انتخاب مي‌كند. كروموزوم‌هاي برازنده‌تر شانس بيشتري دارند تا براي توليدمثل انتخاب شوند.

۲-۳-۹- انتخاب نخبگان[۳۹]:

مناسب‌ترین عضو هر اجتماع انتخاب می‌شود. با توجه به مقدار شایستگی که از تابع ارزیاب دریافت کرده است.

---

[33] Reproduction

[34] Generation

[35] Selection

[36] Crossover

[37] Mutation

[38] Selection

[39] Elitist Selection

۲-۳-۱۰- نمونه‌برداري به روش چرخ رولت

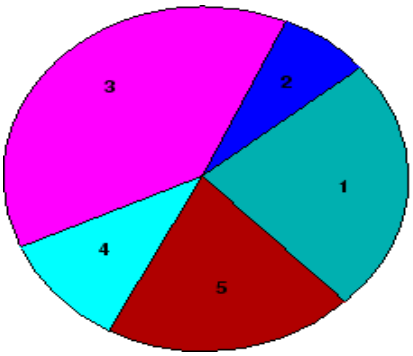
در اين روش، به هر فرد قطعه‌اي از يك چرخ رولت مدور اختصاص داده مي‌شود. اندازه اين قطعه متناسب با برازندگي آن فرد است. چرخ N بار چرخانده مي‌شود كه N تعداد افراد در جمعيت است. در هر چرخش، فرد زير نشانگر چرخ انتخاب مي‌شود و در مخزن والدين نسل بعد قرار مي‌گيرد. اين روش مي‌تواند به‌صورت زير پياده‌سازي شود:

نرخ انتظار كل افراد جمعيت را جمع كنيد و حاصل آن را T بناميد.

مراحل زير را N بار تكرار كنيد:

يك عدد تصادفي r بين صفر و T انتخاب كنيد.

در ميان افراد جمعيت بگرديد و نرخ‌هاي انتظار (مقدار شايستگی) آن‌ها را باهم جمع كنيد تا اين‌كه مجموع بزرگ‌تر يا مساوي r شود. فردي كه نرخ انتظارش باعث بيشتر شدن جمع از اين حد مي‌شود، به‌عنوان فرد برگزيده انتخاب مي‌شود.



Population	Fitness
1	25.0
2	5.0
3	40.0
4	10.0
5	20.0

شكل ۲-۱۴- نحوه ارزيابی شايستگی در چرخ رولت

۲-۳-۱۱- انتخاب رقابتی[۴۰]:

یک زیرمجموعه از صفات یک جامعه انتخاب می‌شوند و اعضای آن مجموعه باهم رقابت می‌کنند و سرانجام فقط یک صفت از هر زیرگروه برای تولید انتخاب می‌شوند.

۲-۳-۱۲- عملگر آمیزش:

در جریان عمل تلفیق به‌صورت اتفاقی بخش‌هایی از کروموزوم‌ها با یکدیگر تعویض می‌شوند. این موضوع باعث می‌شود که فرزندان ترکیبی از خصوصیات والدین خود را به همراه داشته باشند و دقیقاً مشابه یکی از والدین نباشند.

هدف تولید فرزند جدید می‌باشد به این امید که خصوصیات خوب دو موجود در فرزندشان جمع شده و یک موجود بهتری را تولید کند.

روش کار به‌صورت زیر است:

به‌صورت تصادفی یک نقطه از کروموزوم را انتخاب می‌کنیم

ژن‌های مابعد آن نقطه از کروموزوم‌ها را جابجا می‌کنیم

تلفیق تک نقطه‌ای[۴۱]

اگر عملیات تلفیق را در یک نقطه انجام دهیم به آن تلفیق تک نقطه‌ای می‌گویند. تلفیق بدین‌صورت انجام می‌گیرد که حاصل ترکیب کروموزوم‌های پدر و مادر می‌باشد. روش تولیدمثل نیز بدین‌صورت است که ابتدا به‌صورت تصادفي، نقطه‌ای که قرار است تولیدمثل ازآنجا آغاز گردد، انتخاب می‌گردد. سپس اعداد بعدازآن به ترتیب از بیت‌های کروموزوم‌های پدر و مادر قرار می‌گیرد که در شکل زیر نیز نشان داده‌شده است.

---

[40] Tournament Selection

[41] Single Point Crossover

Chromosome 1	11011 | 00100110110
Chromosome 2	11011 | 11000011110
Offspring 1	11011 | 11000011110
Offspring 2	11011 | 00100110110

شکل ۲-۱۵- شکل یک نمونه تلفیق (آمیزش)

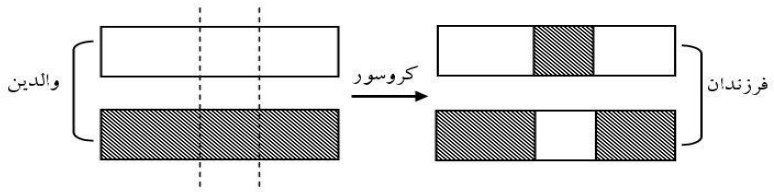
در شکل بالا کروموزوم‌های ۱ و ۲ در نقش والدین هستند؛ و حاصل تولیدمثل آن‌ها در رشته‌هایی بنام Offspring ذخیره‌شده است. دقت شود که علامت "|" مربوط به نقطه شروع تولیدمثل می‌باشد و در رشته‌های Offspring اعدادي که بعد از نقطه شروع تولیدمثل قرار می‌گیرند مربوط به کروموزوم‌های مربوط به خود می‌باشند. به‌طوری‌که اعداد بعد از نقطه شروع مربوط به Offspring1 مربوط به اعداد بعد از نقطه شروع مربوط به کروموزوم ۱ و اعداد بعد از نقطه شروع تولیدمثل مربوط به Offspring2 مربوط به اعداد بعد از نقطه شروع تولیدمثل مربوط به کروموزوم ۲ می‌باشند.

از روش فوق برای تلفیق دو کروموزوم در الگوریتم ژنتیک در این پروژه استفاده‌شده است.

روش ادغام دونقطه‌ای[۴۲]:

در این روش دو مکان را به‌صورت تصادفی انتخاب کرده و مقادیر بین این دونقطه را جابجا می‌کنیم.

[42] Two-point Crossover



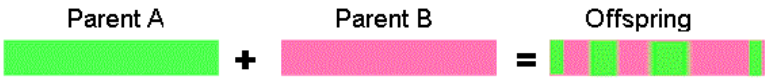
شکل ۲-۱۶- شکل تلفیق نقطه‌ای

می‌توانیم این عملیات را در چند نقطه انجام دهیم که به آن بازترکیبی چندنقطه‌ای می‌گویند

تلفیق جامع[۴۳]:

اگر تمام نقاط کروموزوم را به‌عنوان نقاط بازترکیبی انتخاب کنیم به آن بازترکیب جامع گوییم.

مثال



شکل ۲-۱۷- شکل تلفیق جامع

عملگر جهش[۴۴]:

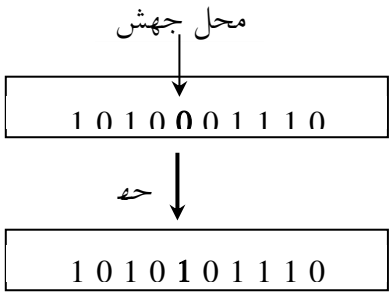
پس از اتمام عمل آميزش، عملگر جهش بر روي كروموزوم‌ها اثر داده مي‌شود. اين عملگر يك ژن از يك كروموزوم را به‌طور تصادفي انتخاب نموده و سپس محتواي آن ژن را تغيير مي‌دهد. اگر ژن از جنس اعداد دودويي باشد، آن را به وارونش تبديل مي‌كند و چنانچه متعلق به يك مجموعه باشد، مقدار يا عنصر ديگري از آن مجموعه را به‌جای آن ژن قرار مي‌دهد. در شكل ۲ چگونگي جهش يافتن پنجمين ژن يك كروموزوم نشان داده‌شده است.

---

[43] Uniform Crossover

[44] Mutation

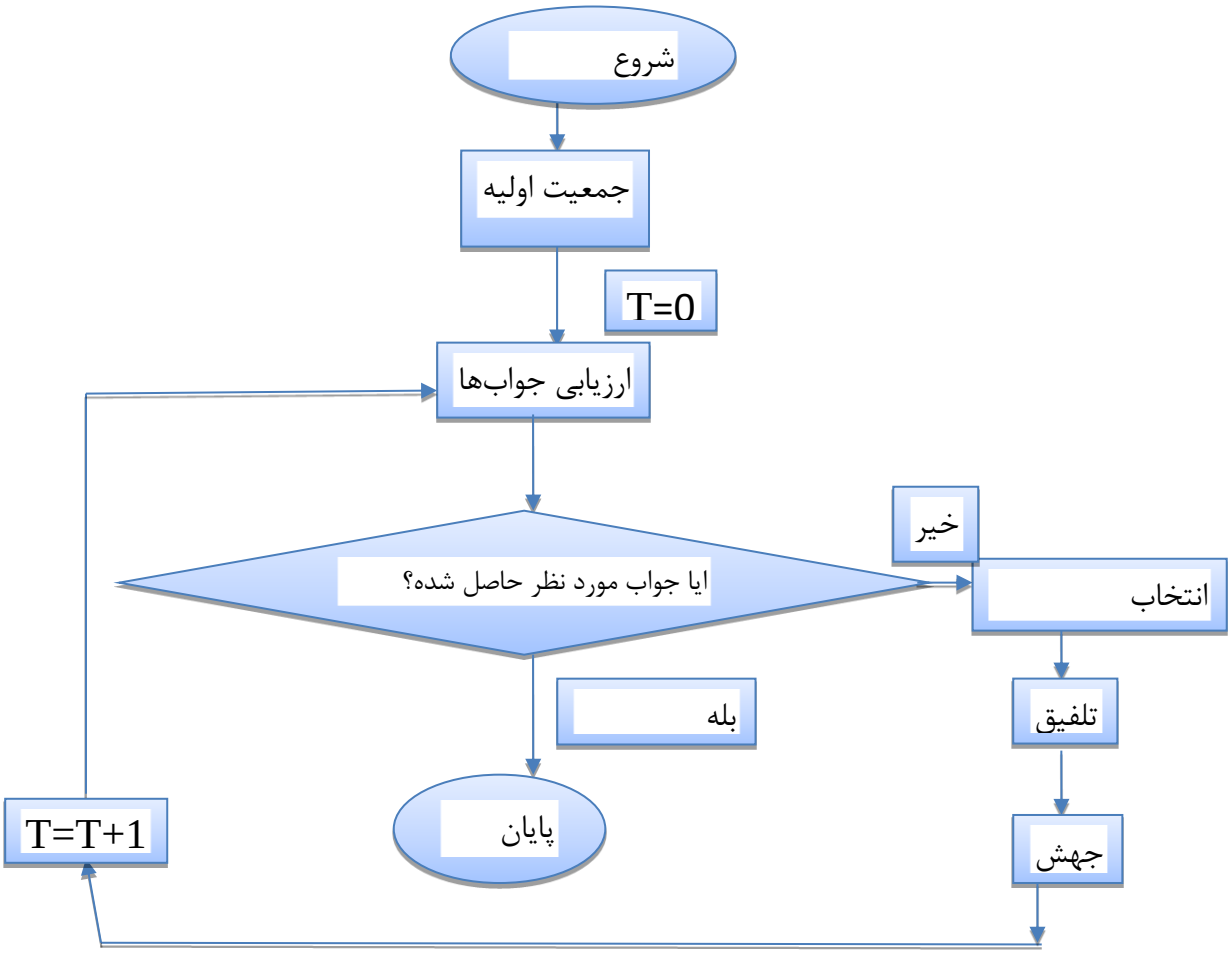
پس از اتمام عمل جهش، كروموزوم‌هاي توليدشده به‌عنوان نسل جديد شناخته‌شده و براي دور بعد اجراي الگوريتم ارسال مي‌شوند.



شكل ۲-۱۸- يك كروموزوم قبل و بعد از اعمال عملگر جهش

۲-۳-۱۳- روند كلي الگوريتم‌هاي ژنتيكي

قبل از اين‌كه يك الگوريتم ژنتيكي بتواند اجرا شود، ابتدا بايد كدگذاري (يا نمايش) مناسبي براي مسئله موردنظر پيدا شود. معمولی‌ترين شيوه نمايش كروموزوم‌ها در الگوريتم ژنتيک به شكل رشته‌های دودويي است. هر متغير تصميم‌گيری به‌صورت دودويي درآمده و سپس با كنار هم قرار گرفتن اين متغيرها كروموزوم ايجاد مي‌شود. گرچه اين روش گسترده‌ترين شيوه كدگذاري است اما شيوه‌های ديگري مثل نمايش با اعداد حقيقي در حال گسترش هستند. همچنين يك تابع برازندگي نيز بايد ابداع شود تا به هر راه‌حل كدگذاري شده ارزشي را نسبت دهد. در طي اجرا، والدين براي توليدمثل انتخاب مي‌شوند و با استفاده از عملگرهاي آميزش و جهش باهم تركيب مي‌شوند تا فرزندان جديدي توليد كنند. اين فرآيند چندين بار تكرار مي‌شود تا نسل بعدي جمعيت توليد شود. سپس اين جمعيت بررسي مي‌شود و درصورتی‌كه ضوابط همگرايي برآورده شوند، فرآيند فوق خاتمه مي‌يابد.

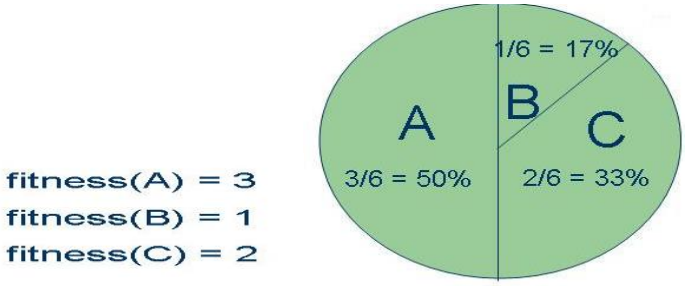


شکل ۲-۱۹- فلوچارت الگوریتم ژنتیک

۲-۳-۱۴- روند كلي بهينه‌سازی و حل مسائل در الگوريتم ژنتيک

1-شروع: توليد تصادفي يک جمعيت که شامل تعداد زيادي کروموزوم (روش‌های حل مسئله است) می‌باشد.

2-صحت و درستي: ارزيابي صحت براي تابع f(x) به ازای هر کروموزوم x در جمعيت.



شکل ۲-۲۰- نحوه ارزیابی تابع شایستگی در چرخ رولت

3-ایجاد یک جمعیت جدید: تولید یک جمعیت جدید با انجام تمامي زیرگروه‌های زیر تا آن‌که یک جمعیت جدید ایجاد گردد.

۳-۱ انتخاب: انتخاب کروموزوم‌های پدر و مادر از جمعیت قبلي با توجه به صحت و درستي آن به‌طوری‌که هر چه برازندگی بهتر باشد (دقت جواب در هم‌گرائی بیشتر باشد) شانس بیشتري براي انتخاب دارد.

۳-۲: تولیدمثل: انجام زادوولد و ایجاد یک نسل جدید.

۳-۳: جهش: مشخص شدن مکان فرزند تولیدشده در کروموزوم

۳-۴: پذیرش: جا دادن فرزند جدید در داخل جمعیت.

4-جایگزیني: جایگزیني جمعیت جدید به‌جای جمعیت قبلي و مورداستفاده قرار دادن جمعیت جدید در مراحل بعدي الگوریتم

5-امتحان: اگر شرایط مطلوب در حل مسئله ارضا شد اعلام می‌کنیم که به بهترین جواب رسیده‌ایم و از الگوریتم خارج می‌شویم در غیر این صورت به مرحله ۲ می‌رویم و دوباره همین روند را تکرار می‌کنیم.

۲-۳-۱۵- شرط پایان الگوریتم

چون‌که الگوریتم‌های ژنتیک بر پایه تولید و تست می‌باشند، جواب مسئله مشخص نیست و نمی‌دانیم که کدام‌یک از جواب‌های تولیدشده جواب بهینه است تا شرط خاتمه را پیدا شدن جواب در جمعیت تعریف کنیم. به همین دلیل، معیارهای دیگری را برای شرط خاتمه در نظر می‌گیریم:

تعداد مشخصی نسل: می‌توانیم شرط خاتمه را مثلاً ۱۰۰ دور چرخش حلقه اصلی برنامه قرار دهیم.

عدم بهبود در بهترین شایستگی جمعیت در طی چند نسل متوالی

بهترین شایستگی جمعیت تا یک‌زمان خاصی تغییری نکند.

شرایط دیگری نیز می‌توانیم تعریف کنیم و همچنین می‌توانیم ترکیبی از موارد فوق را به‌عنوان شرط خاتمه به کار ببندیم.

۲-۳-۱۶-بهبود الگوریتم خوشه‌بندی k-means به کمک الگوریتم ژنتیک

عیب معمول الگوریتم $k-means$ که در بالا توضیح داده‌شده این است که انتخاب حساسیت دانه‌های اولیه می‌تواند بر خروجی نهایی تأثیر بگذارد و به‌راحتی در بهینه مکانی قرار گیرد. جهت ممانعت از همگرایی پیش از بلوغ خوشه‌بندی $k-means$ ما الگوریتم ژنتیکی را به‌عنوان ابزار بهینه‌سازی برای بیرون دادن دانه‌های اولیه در مرحله اول پروسه $k-means$ جهت شناسایی قسمت‌های بهینه در نظر می‌گیریم. در این پروژه، یک کروموزوم با ژن‌های k برای مراکز خوشه k به‌صورت ($x_1, x_2, ..., x_{12}$) تعیین می‌شود.x_1 برداری با n بعد هست. در طی پروسه تحول، از تابع برازندگی برای ارزیابی کیفیت و از روش تلفیق تک نقطه‌ای برای تلفیق دو کروموزوم استفاده‌شده است

$$f(\text{chromosome}) = \sum_{x_j \in X} \min_{1 \le i \le k} (\text{dist}(C_i, x_j))$$

مقدار تناسب، مجموع فاصله‌ها برای همه نقاط داخلی به مراکز خوشه‌شان هست و سعی می‌کنند مقادیری که متناظر با قسمت‌های بهینه‌شده است را به حداقل برساند. در هر تکرار متوالی سه اپراتور ژنتیکی جلوتر می‌روند تا جمعیت‌های جدید را به‌صورت نوزادان بر طبق متناسب‌ترین اصول حیات را تولید کنند. جمعیت‌ها متمایل به نزدیک شدن به کروموزوم (راه‌حل) بهینه به هنگام تأمین ملاک برازندگی می‌باشند.

وقتی مراکز خوشه بهینه بیرون می‌آیند، ما آن‌ها را به‌عنوان دانه‌های اولیه جهت انجام الگوریتم $k-means$ در مرحله‌های آخر خوشه‌بندی استفاده می‌نماییم.

عملگر جهش استفاده‌شده در این پروژه به این صورت عمل می‌کند که ابتدا بهترین کروموزوم‌ها انتخاب می‌شود سپس با احتمال Ciمقادیر کروموزوم انتخاب‌شده را با یه کروموزوم تصادفی عوض می‌کنیم، برازندگی کروموزوم جدید را حساب می‌کنیم سپس اگر برازندگی کروموزوم جدید بهتر باشد کروموزوم جدید انتخاب می‌شود در غیر این صورت کروموزوم قدیمی نگه‌داشته می‌شود.

۲-۴- الگوریتم پالایش گروهی

به کاربر اقلامی توصیه خواهد شد که دیگران درگذشته با تمایلات و ترجیحات مشابه او این اقلام را پسندیدند؛ یعنی بر اساس رابطه بین کاربران و کالاها، اقلام جدید به کاربر توصیه می‌شود. در این روش‌ها، خود کالا اهمیتی ندارد و بر اساس انتخاب کاربران دیگر و انتخاب‌های گذشته خود کاربر، به او پیشنهادهای جدیدی ارائه می‌کنیم. این روش بر سلایق مشترک بنانهاده شده است.

۲-۴-۱-انواع مختلف دسته‌بندی پالایش گروهی

پالایش گروهی را می‌توان به‌طورکلی از دو دیدگاه دسته‌بندی کرد. دیدگاه اول مربوط به استخراج اطلاعات و دیدگاه دوم مربوط به جمع‌آوری داده‌ها می‌باشد.

در دیدگاه استخراج داده‌ها نحوه ارائه پیشنهادها به کاربران بررسی می‌شود که از دو نوع الگوریتم متفاوت بسته به شرایط استفاده می‌شود. نوع اول الگوریتم‌های مبتنی بر حافظه[۴۵] هستند که این الگوریتم‌ها برای ارائه پیشنهادها از ماتریس کامل رتبه‌بندی[۴۶] استفاده می‌کنند. نوع دوم الگوریتم‌های مبتنی بر مدل[۴۷] هستند که این الگوریتم‌ها برای ارائه پیشنهادها از ماتریس رتبه‌بندی برای ایجاد یک مدل استفاده می‌کنند و

---

[45] Memory-Based

[46] Rating matrix

[47] Model-Based

سپس با استفاده از این مدل پیشنهادها را ارائه می‌دهند. الگوریتم‌های مبتنی بر حافظه نسبت به الگوریتم‌های مبتنی بر مدل نتیجه بهتر و دقیق‌تری می‌دهند و هنگامی‌که ماتریس ارزیابی مرتباً تغییر می‌کند مناسب‌تر می‌باشند. از جهت دیگر این الگوریتم‌ها زمان محاسباتی زیادی نیاز دارند که این امر باعث می‌شود تا در پایگاه داده‌های بزرگ از الگوریتم‌های تقریبی مبتنی بر مدل استفاده شود.

در دیدگاه جمع‌آوری داده‌ها نحوه جمع‌آوری اطلاعات از کاربران بررسی می‌شود که به‌طورکلی داده‌ها به دودسته تقسیم می‌شوند. دسته اول داده‌های استخراج‌شده از رفتار کاربر[۴۸] هستند که در برخی از پیاده‌سازی‌های پالایش گروهی به دلیل اینکه دریافت رتبه‌بندی از کاربران به‌سادگی امکان‌پذیر نیست از داده‌هایی که کاربر در هنگام مشاهده صفحات از خود به‌جا می‌گذارد برای ارائه پیشنهادها استفاده می‌شود؛ مانند روند بازدید صفحات و مدت‌زمان مشاهده اقلام مختلف ارائه‌شده در وب‌سایت. دسته دوم داده‌های دریافت شده از خود کاربر[۴۹] هستند که کاربران ممکن است با مشخص کردن علاقه‌مندی‌های خود در هنگام خرید کالاهای قبلی و دادن رتبه به هرکدام به سیستم پالایش گروهی اجازه دهند تا پیشنهادها دقیق‌تری را به آن‌ها ارائه دهد. به‌عبارت‌دیگر کاربران در هنگام خرید کالاهای خود نظر خود را در ارتباط با آن کالا به‌صورت بازخورد (که معمولاً از طریق رتبه دهی انجام می‌شود) در سیستم ثبت می‌نمایند. پالایش گروهی در پیشنهادهای آینده خود از این اطلاعات استفاده کرده و کالاهای جدید را مطابق با علایق کاربر پیشنهاد می‌دهد. داده‌های استخراج‌شده از خود کاربر بسیار دقیق‌تر از داده‌های استخراج‌شده از رفتار کاربر می‌باشند زیرا کاربر نظر خود را دقیق اعلام می‌نماید.

۲-۴-۲-الگوریتم‌های متداول پالایش گروهی

در قسمت قبل به نحوه استخراج پیشنهادها اشاره شد که برای استخراج پیشنهادها از دو روش مبتنی بر حافظه و مبتنی بر مدل استفاده می‌شود. در این قسمت به سه نمونه از الگوریتم‌های متداولی که برای استخراج داده‌ها استفاده می‌شود اشاره می‌کنیم.

---

[48] Implicit data

[49] Explicit Data

الگوریتم تصادفی[۵۰]: در این روش به ازای هر کاربر U و هر کالای I یک عدد تصادفی ایجاد می‌شود که در هر اجرا همواره ثابت است. به‌عبارت‌دیگر در هر بار تلاش برای گرفتن پیشنهادها الگوریتم به کاربر و کالا عدد تصادفی یکسان با دفعات قبل اختصاص می‌دهد. این روند این قابلیت را می‌دهد که پیشنهادها به‌طور تصادفی ایجاد شود ولی در اجرایی متفاوت نتایج یکسانی به دست آید. این الگوریتم در دسته الگوریتم‌های مبتنی بر مدل جای می‌گیرد.

الگوریتم میانگین[۵۱]: در این روش برای هر کالا میانگین رتبه‌ای که دیگر کاربران به آن داده‌اند محاسبه می‌شود و با توجه به درخواست کاربر برای مشاهده نتایج تعداد K عدد از بیشترین میانگین‌ها پیشنهاد می‌گردد. این الگوریتم در دسته الگوریتم‌های مبتنی بر مدل جای می‌گیرد.

الگوریتم‌های بر پایه‌ی همسایگی[۵۲]: یکی از معروف‌ترین و پراستفاده‌ترین الگوریتم‌هایی که در پالایش گروهی استفاده می‌شود الگوریتم‌های بر پایه همسایگی هست. در این الگوریتم‌ها سعی می‌شود تا کاربرانی که علایق مشترکی با کاربر فعلی دارند ابتدا جستجو شوند و سپس کالاهایی که آن کاربران قبلاً تهیه‌کرده‌اند به کاربر فعلی پیشنهاد می‌شود. این الگوریتم با بررسی کالاهایی که کاربر فعلی و دیگر کاربران به‌طور مشترک خریداری کرده‌اند به این نتایج دست پیدا می‌کند. در پیاده‌سازی این الگوریتم‌ها از دو راهکار متداول استفاده می‌شود:

کاربربه‌کاربر[۵۳]: این الگوریتم‌ها در هنگام ارائه پیشنهاد به کاربر فعلی ابتدا در ماتریس ارزیابی (Rating Matrix) کاربران دیگری که علاقه‌مندی‌های مشابهی را در خریدهای گذشته خود نسبت به این کاربر فعلی داشته‌اند را جستجو کرده و سپس کالاهایی که این کاربران درگذشته انتخاب کرده‌اند را پیشنهاد می‌دهند.

---

[50] Random Algorithm

[51] Mean Algorithm

[52] Neighborhood-Based Algorithms

[53] User-to-User

کالا به کالا[۵۴]: این الگوریتم‌ها در هنگام ارائه پیشنهادها ابتدا کالاهایی که کاربر فعلی قبلاً انتخاب کرده است را بررسی کرده و سپس با توجه به کاربرانی که قبلاً نیز این کالاها را انتخاب کرده‌اند کالاهای دیگری که آن کاربران نیز انتخاب کرده‌اند را پیشنهاد می‌دهند. به‌عبارت‌دیگر ابتدا ماتریسی ارزیابی محدود به کاربرانی می‌شود که کالاهای مشترک را انتخاب کرده‌اند سپس تناظر بین کاربران و کالاهای خریداری‌شده انجام می‌پذیرد. در حقیقت در مورد اول ماتریسی رتبه‌بندی به‌صورت سطری و در مورد دوم به‌صورت ستونی تحلیل می‌شود.

یکی از متداول‌ترین معیارهایی که در به دست آوردن تشابهات استفاده می‌شود ضریب وابستگی پیرسون[۵۵] است. این ضریب، رابطه خطی بین دو متغیر مشخص می‌کند – حدی که دو متغیر باهم رابطه دارند؛ و مقدار آن از -۱ تا ۱ متغیر است. مقدار ۱ نشان‌دهنده ارتباط کامل دو متغیر و مقدار -۱ نمایش‌دهنده عدم ارتباط دو متغیر است. به‌عبارت‌دیگر ۱ نمایش می‌دهد که دو کاربر کاملاً علایق مرتبط باهم دارند درصورتی‌که عدد -۱ نمایش‌دهنده تضاد علایق دو کاربر است. را بطه بین کاربر فعال a و کاربر دیگر u به شرح زیر است:

$$W_{a,u}=\frac{\sum_{i=1}^{n}(r_{a,i}-\bar{r}_a)(r_{u,i}-\bar{r}_u)}{\sqrt{\sum_{i=1}^{n}(r_{a,i}-\bar{r}_a)^2}\sqrt{\sum_{i=1}^{n}(r_{u,i}-\bar{r}_u)^2}}$$

جدول زیر را در نظر بگیرید:

جدول ۲-۹- جدول امتیاز

	Movie1	Movie2	Movie3	Movie4	Movie5
User	4	4	1	4	3

[54] Item-to-Item

[55] Pearson’s correlation coefficients

A					
User B	2	1	4	2	5
User C	3	1	3	2	1
User D	5	4	2		3

چه عددی بین ۱ تا ۵ مشخص‌کننده علاقه کاربر D به مشاهده فیلم ۴ می‌باشد؟ r_a میانگین رتبه دهی کاربر D برابر ۳٫۵ است.r_u میانگین امتیازدهی دیگر کاربران است که برای کاربران A، B و C به ترتیب برابر ۳، ۳ و ۲ است. دقت شود فقط کاربرانی را در نظر می‌گیریم که مانند کاربر فعال فیلم‌های ۱، ۲، ۳ و ۵ را نیز مشاهده کرده‌اند؛ بنابراین n یعنی تعداد فیلم‌های مشترک برابر ۴ هست.$r_{a,i}$ رتبه‌ای است که کاربر فعال به فیلم I داده است و $r_{u,i}$ رتبه‌ای است که دیگر کاربران به فیلم i داده‌اند. با محاسبه فرمول مطابق داده‌های جدول به دست می‌آوریم:

$W_{D,A}=0.9, W_{D,B}=-0.7, W_{D,C}=0$

که بیانگر تشابه سلیقه کاربر D و A و عدم تشابه سلیقه کاربر D و B می باشد.

بعد از محاسبه میزان تشابه کاربر فعال با کاربران دیگر برای تخمین امتیاز فیلم‌های پیشنهادی از فرمول زیر استفاده می‌کنیم

$$P_{Ua,item} = \overline{R_u} + \frac{\sum_{y \in C_x} sim(U_a, y) \times (R_{y,i} - \overline{R_y})}{\sum_{y \in C_x} (|sim(U_a, y)|)}$$

در این معادله R_n عدد میانگین نرخ‌بندی ارائه‌شده توسط U_a می‌باشد. C_x دسته همسایگان متعلق به یک خوشه مشترک با U_a است؛ و R_y نشان‌دهنده میانگین نرخ‌بندی توسط آن همسایه است.

۲-۴-۳-مشکلات CF

۱. **فیلم‌های جدید**[۵۶]: الگوریتم‌هایی که بر اساس همسایگی مطرح شد تنها هنگامی به‌خوبی عمل می‌کنند که داده‌های زیادی وجود داشته باشند. به‌عبارت‌دیگر کاربران مختلف فیلم‌های مختلف را ارزیابی کرده باشند. در مواردی مانند خرید خانه که تنها یک کالا (خانه موردنظر) وجود دارد خرید کالای مشابه توسط کاربران مختلف معنایی ندارد. درنتیجه ارزیابی کاربران قبلی از این کالا بدون معنی هست و سیستم هیچ‌گاه این‌گونه کالاها را پیشنهاد نمی‌دهد. همچنین این مشکل برای فیلم‌های که به‌تازگی به لیست اضافه‌شده‌اند و هنوز هیچ کاربری آن‌ها را ارزیابی نکرده است به وجود می‌آید.

۲. **کمبود نظرها**[۵۷]: از جهت دیگر کاربران اغلب مایل به ارائه نظر خود در مورد فیلم‌ها نمی‌باشند. درنتیجه در ماتریسی ارزیابی بسیاری از خانه‌ها خالی خواهد ماند.

۳. **مقیاس‌پذیری**[۵۸]: به‌موازات اینکه داده‌ها افزایش پیدا می‌کند حجم محاسبات بر روی ماتریس ارزیابی نیز افزایش پیدا می‌کند که این مورد در سامانه‌های online مشکل‌ساز ظاهر می‌شود.

۴. **حریم خصوصی**[۵۹]: امنیت اطلاعات افراد یکی از مسائل مشکل‌آفرین در سامانه‌های CF هست. اغلب کاربران مایل نیستند اطلاعات خود را در معرض عموم قرار دهند درنتیجه راهکارهایی برای امنیت اطلاعات بایستی ایجاد شود.

۵. **اعتبار داده‌ها**[۶۰]: به‌مرورزمان علایق شخصی کاربران تغییر خواهد کرد. مشکل دیگری که در CF بایستی با آن روبه‌رو شد اعمال این تغییر علاقه کاربران در پیشنهادها جاری است.

---

[56] Cold Start

[57] Sparcity

[58] Scalability

[59] Privacy

[60] Recency

۶. **اعتماد به سامانه‌های پیشنهاددهنده**: اغلب کاربران مایل‌اند بدانند در پیشنهادی که به آن‌ها ارائه می‌شود چه معیارهایی در نظر گرفته می‌شود. درنتیجه بایستی با دلایل مناسب سیستم‌های CF کاربران خود را قانع نمایند. به‌عبارت‌دیگر چگونگی انتخاب معیار برای ارائه پیشنهاد نقش بسیار حیاتی در سامانه‌های پیشنهاددهنده دارد.

۷. **اعتماد به داده‌های موجود در سیستم‌های پیشنهاددهنده**: واردکردن داده‌های نادرست توسط مشاهده‌کنندگان فیلم‌ها می‌تواند روند پیشنهادها را از مسیر صحیح خود خارج نماید.

راه‌حل مشکل مقیاس‌پذیری و پراکندگی داده

برای مشخص نمودن چالش‌های ذکرشده، یک توصیه گر فیلم مبتنی بر مدل ترکیبی برای حل مشکل مقیاس‌پذیری بالا و پراکندگی داده پیشنهاد می‌شود. در این پروژه، یک الگوریتم خوشه‌بندی بهینه برای تقسیم پروفایل‌های کاربری گسترش می‌دهیم که توسط بردارهای پروفایل متراکم‌تر پس از تبدیل تحصیل جزء اصلی ارائه‌شده است. کل سیستم شامل دو فاز، یک‌فاز آنلاین و یک‌فاز آفلاین هست. در فاز آفلاین، مدل خوشه‌بندی در فضای بعد پایین آموزش داده می‌شود و برای هدف قرار دادن کاربران فعال در خوشه‌های مختلف آماده می‌شود. در فاز آنلاین، یک فهرست نظریه‌ای فیلم TOP Nبرای یک کاربر فعال از نرخ‌بندی‌های پیش‌بینی‌شده فیلم‌ها ارائه می‌شود. به‌علاوه، یک الگوریتم ژنتیکم در این روش برای بهبود عملکرد خوشه‌بندی k-means به کار گرفته می‌شود و الگوریتم خوشه‌بندی پیش نهادی با عنوان PCAGA-KM نامیده می‌شود. سپس عملکرد روش پیش نهادی را در مجموعه داده Movielens جستجو می‌کنیم.

۳-فصل سوم رابط کاربری سامانه توصیه گر

در این فصل در مورد رابط کاربری سامانه توصیه گر با کاربر صحبت می‌شود برای این منظور وب سایتی طراحی‌شده است که کاربر با ورود به سایت و امتیازدهی به فیلم‌ها می‌تواند فیلم‌های پیشنهادی که خروجی سامانه توصیه گر است را مشاهده کند این وب‌سایت با استفاده از زبان برنامه‌نویسی nodejs و همچنین دیتابیس mongodb پیاده شده است که این از نوع دیتابیس‌های nosql است که برای مدیریت دیتاهای کلان استفاده می‌شود.

HTML[۶۱]-۱-۳

عبارت HTML به معنی زبان نشانه‌گذاری فوق متن است. Html زبان استاندارد طراحی صفحات وب است و کلیه کدهای صفحه اعم از طرف سرور و طرف مشتری درنهایت به کدهای HTML تبدیل‌شده و توسط مرورگر نمایش داده می‌شوند. به‌عبارت‌دیگر مرورگرها هیچ‌کدام از کدها و کنترل‌های سمت سرور همچون کدهای asp و php را نمی‌شناسند و کد قابل‌فهم برای آن‌ها اچ تی ام ال می‌باشد. کامپایلرهای زبان‌های برنامه‌نویسی سروری درنهایت کدهای خود را برای نمایش به کد اچ تی ام ال تبدیل می‌کنند و برای مرورگر می‌فرستند تا به کاربران نمایش داده شود.

HTML یک‌زبان نشانه‌گذاری است، به این معنی که بخش‌های مختلف توسط اجزایی به نام تگ از هم جداشده که هرکدام دارای کاربرد و خواص مربوط خود هستند. این تگ‌ها به مرورگر اعلام می‌کنند که هر بخش از صفحه چه نوع عنصری است و باید به چه صورت نمایش داده شود.

در یک صفحه HTML می‌توان انواع عناصر از قبیل متن، تیتر، عکس، جدول و ... را قرارداد که برای هر عنصر باید از تگ مربوط به آن استفاده کرد. صفحات HTML فقط از کدها که به‌صورت متن هستند تشکیل‌شده‌اند. بدین معنا که برای تصویر کد مربوط به تمایش تصویر و جدول و ...کدهای اچ تی ام ال مربوط به هر یک را باید نوشت و مرورگر با رسیدن به این کدها و تگ‌ها، المنت‌های مرتبط با آن را نمایش می‌دهد.

---

[61] Hyper Text Markup Language

هر یک از کدهای HTML، معنا و مفهوم خاصی دارند و تأثیر مشخصی بر محتوا می‌گذارند. مثلاً برچسب‌هایی برای تغییر شکل ظاهری متن، نظیر درشت و ضخیم کردن یک کلمه یا برقراری پیوند به صفحات دیگر در HTML تعریف شده‌اند.

یک سند HTML، یک پروندهٔ مبتنی بر متن است که معمولاً با پسوندhtm. یاhtml. نام‌گذاری شده و محتویات آن از برچسب‌های HTML تشکیل می‌شود. مرورگرهای وب که قادر به درک و تفسیر برچسب‌های HTML هستند، تک‌تک آن‌ها را از داخل سند HTML خوانده و سپس محتوای آن صفحه را نمایان‌سازی رندر می‌کنند.

HTML زبان برنامه‌نویسی نیست، بلکه زبانی برای نشانه‌گذاری ابرمتن است و اساساً برای ساخت‌مند کردن اطلاعات و جدایش اجزای منطقی یک نوشتار به کار می‌رود. از سوی دیگر، HTML را نباید به‌عنوان زبانی برای صفحه‌آرایی یا نقاشی صفحات وب به کاربرد؛ این وظیفه اکنون بر دوش فناوری‌های دیگری همچون CSS است.

CSS[۶۲]-۲-۳

CSS زبان برنامه‌نویسی می‌باشد که کنسرسیوم بین‌المللی شبکه جهانی وب یا W3C برای غلبه بر مشکلاتی که در طی زمان با استفاده از HTML به وجود آمده است پیشنهاد داده است. عملاً این زبان برنامه‌نویسی، مکملی بر زبان باستانی HTML است و سعی در پر کردن نقاط ضعف و خلأهای آن دارد.

CSS زبانی است که توسط آن قادر خواهید تا استیل طراحی صفحات وب‌سایتتان را یک‌بار تعریف و به صفحات موردنیازتان اعمال نمایید. برای این منظور مثالی را می‌زنیم. تصور کنید که سایت شما شامل ۱۰۰ صفحه استاتیک می‌باشد و شما آن‌ها را تماماً به زبان HTML نوشته‌اید. بعد از یک هفته تصمیم گرفته‌اید تا فونت تمام کلمات راکمی بزرگ‌تر کنید. گفتن اینکه فونت تمام کلمات بزرگ‌تر شود بسیار کار راحتی است و تنها یک جمله است. ولی آیا در عمل تغییر ۱۰۰ صفحه نیز به همان راحتی خواهد بود؟

---

[62] Cascade Style Sheets

قطعاً نه. CSS دقیقاً همان زبانی است که جمله یک‌خطی شمارا تبدیل به همان یک جمله خواهد کرد. شما تنها کافی است تا استیل موردنیازتان را در طراحی تغییر دهید و آن‌هم تنها با تغییر یک یا چند مورد کوچک.

جمله آخر اینکه، استفاده از CSS باعث تمیزتر شدن کدهای برنامه‌نویسی‌تان می‌شود، تغییرات آتی را آسان می‌کند و همچنین دید شمارا بیشتر به طراحی معطوف می‌کند تا سروکله زدن با کدهای برنامه‌نویسی.

۳-۳-جاوا اسکریپت[۶۳]

جاوا اسکریپت برای اولین بار توسط شرکت Netscape و با نام LiveScript به‌عنوان نرم‌افزاری مفید جهت استفاده در دنیای وب به بازار عرضه شد ولی بعدها با حمایت شرکت Sun Microsystemsپدیدآورنده Java بانام جاوا اسکریپت شناخته شد.

جاوا اسکریپت یک‌زبان اسکریپت نویسی است که بیشتر با کدهای HTML در ارتباط است و دقیقاً همانند کدهای HTML روی پلتفرم‌های مختلف اجرا می‌شود یا به عبارتی به‌وسیله مرورگرهای وب interpret می شود.

اگرچه نت اسکیپ سازنده جاوا اسکریپت است اما درواقع جاوا اسکریپت به سیستم‌عامل یا PlatForm خاصی وابسته نیست.

۳-۳-۱-امکانات و قابلیت‌های جاوا اسکریپت

طراحان صفحات وب می‌توانند با استفاده از Function ها و Object های آماده و از پیش تعریف‌شده جاوا اسکریپت قابلیت‌های زیادی را برای صفحات وب ایجاد کنند. برای مثال:

---

[63] javascript

- قالب Html را طوری طراحی کنند که کاربران بتوانند خود اجزای صفحه وب مثل سایز لینک یا متن را داشته باشند.
- می‌توان با استفاده از کدهای گرافیکی پویانمایی ایجاد کرد و همچنین صفحاتی را طراحی کرد که کاربر به‌دلخواه قادر به جابجایی یا تغییر تصاویر گرافیکی باشد.
- Event ها را کنترل کند و با جاوا و Plug-in ها ارتباط داشته باشد.
- فرمهای Clint-Side ایجاد کند و اطلاعات واردشده توسط کاربر در برگه‌ها را ارزیابی کند و در صورت وجود هرگونه خطایی در نحوه پر شدن آن‌ها پیغام مناسب را نمایش دهد.

۳-۳-۲-تفاوت جاوا و جاوا اسکریپت

درعین‌حال که جاوا اسکریپت توانایی‌های بسیاری درزمینهٔ ایجاد و طراحی صفحات وب دارد به علت وجود بعضی از محدودیت‌ها در آن، تنها برای نوشتن برنامه‌های کوچک و ساده در صفحات وب بکار می‌رود.

برخلاف جاوا که برنامه‌های آن قبل از اجرا باید کامپایل شود و به بایت کد تبدیل شود برنامه‌های جاوا اسکریپت نیازی به کامپایل برای اجرا ندارند و در همان لحظه اجرا به‌وسیله مرورگر خوانده‌شده و interpret می‌شوند.

گرچه می‌توان به‌وسیله جاوا اسکریپت یک پرسشنامه یا فرم را به server فرستاد اما جاوا اسکریپت قدرت ایجاد ارتباط متقابل بین server و client را به‌اندازه جاوا ندارد.

Bootstrap-۴-۳

یکی از فریم‌ورک‌های متن‌باز CSS که با استفاده از آن می‌توان قالب و ظاهر سایت را به‌سرعت طراحی نمود. بوت استرپ درواقع از چند فایل CSS و Java Script تشکیل‌شده است که باعث می‌شود تا دیگر نیازی به نوشتن کدهای سی اس اس یا جاوا اسکریپت نباشد و از کلاس‌های آماده بوت استرپ استفاده نمود. به‌عبارت‌دیگر Bootstrap مجموعه‌ای از ابزارهای رایگان برای ایجاد صفحات وب و نرم‌افزارهای تحت وب جهت تولید و نمایش فرم‌ها، دکمه‌ها، تب‌ها، ستون‌ها و سایر المان‌های موردنیاز است.

یکی از موارد دشوار و وقت‌گیر در طراحی سایت اختصاص ویژگی‌های مختلف به عناصر صفحه مانند لینک‌ها، هدرها، دکمه‌ها و ... است که بوت استرپ به‌طور پیش‌فرض این کار را انجام داده است. درواقع به کمک فریم‌ورک Bootstrap طراحان صفحات وب قادر خواهند بود تا صفحاتی با پویایی بالا ایجاد کنند. بوت‌استرپ شامل قالب‌های آماده طراحی با محوریت HTML و CSS برای تایپوگرافی، فرم‌ها، دکمه‌ها،

نمودارها، منوهای راهبری و دیگر اجزاء رابط کاربری است. بوت‌استرپ یکی از محبوب‌ترین پروژه‌ها در سایت GitHub است که توسط سایت‌های مطرحی چون istockphoto و who.is و godaddy مورداستفاده قرارگرفته است.

از دیگر موارد مهم در طراحی وب‌سایت، نمایش صحیح سایت در دستگاه‌های مختلف نظیر تبلت و موبایل است. طراحی واکنش گرا قابلیتی است که Bootstrap برای این مسئله تدارک دیده است. برای استفاده از قابلیت طراحی Responsive باید از سیستم شبکه‌بندی استفاده نمود؛ سیستم GRID بوت استرپ به‌صورت پیش‌فرض با ۱۲ ستون و عرض ۹۴۰ پیکسل طراحی‌شده که قابل‌تغییر است.

AngularJS-۵-۳

AngularJS یک فریمورک اوپن سورس برای توسعه برنامه‌های تحت وب است. AngularJS در سال ۲۰۰۹ توسط Misko Hevery که کارمند شرکت گوگل بود توسعه یافت و در سال ۲۰۱۲ نسخه رسمی ۱٫۰ آن منتشر شد. AngularJS در حال حاضر توسط گوگل مدیریت و توسعه داده می‌شود. سایت AngularJS یک تعریف از این فریمورک قدرتمند ارائه کرده که ترجمه‌ی آن را در زیر ملاحظه نمایید:

AngularJS یک فریمورک ساختار پذیر برای توسعه برنامه‌های دینامیک و تحت وب است. AngularJS به توسعه‌دهنده این امکان را می‌دهد که از زبان نشانه‌گذاری HTML به‌عنوان طراحی قالب‌ها و تمپلیت‌ها استفاده کند و همچنین با توسعه‌ی سینتکس HTML کامپوننت‌های برنامه را به‌راحتی ایجاد نماید. مفاهیم data binding و dependency injection در Angular باعث می‌شود که حجم وسیعی از کدهایی که قرار است در برنامه‌ی عادی نوشته شود حذف گردند؛ و تمام موارد در مرورگر (سمت کاربر) اجرا می‌شود و این برای تمام تکنولوژی‌های سمت سرور عالی است

۳-۵-۱-ویژگی‌ها

AngularJS یک فریمورک قوی است که صدر صد بر پایه‌ی JavaScript طراحی‌شده و برای ایجاد برنامه‌های تحت وب قدرتمند، یک فریمورک ایده آل است. AngularJS امکاناتی خاص به توسعه‌دهندگان ارائه می‌کند که بتوانند برنامه‌هایی تولید کنند که سمت کاربر کار می‌کند و مسیر توسعه این کدها نیز یک مسیر مشخص و ساده مانند توسعه‌ی MVC هست.

برنامه‌هایی که توسط AngularJS نوشته می‌شود با تمام مرورگرها سازگار می‌باشد. AngularJS به‌صورت خودکار کدهای جاوااسکریپت مناسب برای هر مرورگری را هندل کرده و به همین دلیل در تمام مرورگرها به‌درستی نمایش داده می‌شود.

AngularJS اوپن سورس است، به‌صورت کامل رایگان است، هزاران نفر به‌عنوان توسعه‌دهنده در اقصی نقاط جهان روی آن کار می‌کنند؛ و تحت لایسنس MIT می باشد

به‌طورکلی AngularJS یک فریمورک برای ایجاد برنامه‌های تحت وب در مقیاس‌های بزرگ و با پرفرمنس بالا می‌باشد و یکی از بهترین ویژگی‌هایش این است که به‌راحتی می‌توان آن را مدیریت نمود.

Node.js-۶-۳

Node.js یک محیط[۶۴] برنامه‌نویسی تحت سرور است که بر پایه‌ی موتور جاواسکریپت V8 گوگل کروم توسعه پیداکرده است. Node.js می‌تواند برای ایجاد وب‌سرورهای ساده تا پیشرفته مورداستفاده قرار بگیرد. برای مثال ممکن است برای راه‌اندازی یک وب‌سایت همه‌منظوره که محتواهای متنی و چندرسانه‌ای را در اختیار مشتریان قرار می‌دهد از این ابزار استفاده شود.

ویژگی قابل‌توجهی که Node.js را از سایر محیط‌های برنامه‌نویسی متمایز می‌کند رویداد گرا بودن آن است. برنامه‌نویسی رویداد گرا به شیوه‌ای از برنامه‌نویسی گفته می‌شود که اجرا کدهای برنامه وابسته به رخداد رویدادهای خاص است. برای مثال در هنگام خواندن محتویات یک فایل، در برنامه‌نویسی سنتی، برنامه در هنگام خوانده شدن محتویات فایل از روی رسانه، متوقف می‌شود و پیشروی نمی‌کند. در معماری رویداد گرا، برنامه به پیشروی خود ادامه می‌دهد و هنگامی‌که محتویات از فایل خوانده شدند، تابع دلخواهی از برنامه توسط Node.js فراخوانی می‌شود.

به نظر می‌رسد زبان برنامه‌نویسی جاواسکریپت درآینده، زبان غالب باشد. استقبال چشم‌گیر از پلتفرم Node.js از دلایل اصلی این پیش‌گویی است.

---

[64] Platform

۳-۶-۱-سیستم چند سکویی

Node.js پشتیبانی به‌خوبی از سیستم‌عامل‌های گوناگون پشتیبانی می‌کند. یک برنامه‌ی نوشته‌شده با Node.js بدون توجه به سیستم‌عامل میزبان در تمام محیط‌ها، به شکلی مشابه عمل می‌کند.

۳-۶-۲-کارکردهای جانبی

اگرچه کارکرد اصلی Node.js برای ایجاد ابزارهای تحت سرور است، بااین‌حال کتابخانه‌های بسیاری برای آن توسعه پیداکرده‌اند که کاربردهای سمت مشتری زیادی را نیز برایش معرفی نموده‌اند. کتابخانه‌های Node.js تقریباً برای تمام نیازهای تولید یک نرم‌افزار کاربردی یا ابزار کمکی وجود دارند. کتابخانه‌هایی مانند ارتباط با بانک اطلاعاتی، ذخیره و بازیابی اطلاعات، ارتباطات شبکه، کدگذاری و کدگشایی اطلاعات، پردازشگرهای فایل‌های تصویری و بسیاری دیگر.

۳-۶-۳-نرم‌افزارهای بر پایه‌ی Node.js

علاوه بر کتابخانه‌های فراوان، برنامه‌های کاربری زیادی نیز با کمک Node.js ایجادشده‌اند که اغلب برای آسان کردن کار برنامه‌نویسان مورداستفاده قرار می‌گیرند. برای مثال ابزارهایی که می‌توانند جایگزین برنامه‌ی سنتی make باشند یا برنامه‌هایی که کار کامپایل زبان‌های برنامه‌نویسی جدید به زبان‌های متداول را انجام می‌دهند.

۳-۶-۴-سرعت

سرعت عملکرد Node.js در شروع عملیات اندکی اندکی کم است. ولی در عملکردهایی که زمان بیشتری لازم باشد، سرعت آن خوب و قابل قبول است.

۳-۷-MongoDB

mongoDBیک پایگاه داده سند-گرا[۶۵] است؛ و در گروه پایگاه‌های داده NOSQL قرار دارد. در این نوع پایگاه داده جدول و رکورد وجود ندارد و از مجموعه[۶۶] و سند استفاده می‌شود. می‌توان گفت مجموعه شبیه به جدول و سند شبیه به رکورد در پایگاه داده رابطه‌ای است.

در این پایگاه داده، داده‌ها ساختار ثابت ندارند و هر دو سند (شبیه رکورد در پایگاه رابطه‌ای) می‌تواند ساختار کاملاً متفاوت داشته باشد، به این نوع ساختار BSON می‌گویند.

BSON چیست؟ مونگودی‌بی داده‌ها را به شکل json ذخیره می‌کند. به این ساختار در مونگودی‌بی BSON می‌گویند. ساختار BSON به شکل زیر است.

از مزیت‌های پایگاه داده مونگودی‌بی نسبت سایر پایگاه‌های داده رابطه‌ای مانند MySQLامکان پردازش و جستجو در حجم بسیار بالاتری از داده‌ها در لحظه و همچنین امکان ذخیره حجم بالاتری از داده‌ها است.

۳-۸-مراحل استفاده از سایت

ابتدا کاربر باید در سامانه ثبت‌نام کند برای این منظور اطلاعات کاربری خودش را شامل نام نام خانوادگی و نام کاربری و رمز عبور را وارد می‌کند.

- بعد از ثبت‌نام کردن با ورد نام کاربری و رمز عبور می‌تواند وارد سیستم شود.
- بعد از ورود به سیستم لیست از فیلم‌های پیشنهادی روبه‌رو می‌شود ازآنجایی‌که ما هیچ اطلاعاتی راجب امتیازدهی کاربری به فیلم‌ها نداریم نمی‌توانیم از سامانه استفاده کنیم که هماهنطور در فصل قبل اشاره شد یکی از چالش‌های روش پالایش گروهی مشکل شروع سرد است که برای غلبه به این مشکل در ابتدا یک لیست از پرطرفدارترین فیلم‌ها برای کاربر نمایش داده می‌شود. سپس کاربر به فیلم‌های که تابه‌حال مشاهده کرده است امتیاز می‌دهد. فیلم‌های امتیاز داده‌شده از لیست پیشنهادی خارج می‌شود و در لیست فیلم‌های امتیاز داده‌شده قرار می‌گیرد.

---

[65] Document-oriented database

[66] collection

- سپس کاربر می‌تواند برای به‌روزرسانی لیست فیلم‌ها از دکمه برروز رسانی فیلم‌های پیشنهادی توصیه‌شده استفاده کند پس از فشردن این دکمه فایل پایتون حاوی الگوریتم پالایش گروهی و الگوریتم‌های پروژه که در فصل قبل صبحت کردیم اجرا می‌شود پس از اجرا این الگوریتم لیست موردعلاقه کاربر آپدیت می‌شود و کاربر فیلم‌های پیشنهادی رو مشاهده می‌کند
- همچنین کاربر می‌تواند امتیازهای داده‌شده به فیلم‌ها که در بخش فیلم‌های امتیاز داده‌شده است قرار دارد را تغییر دهد.

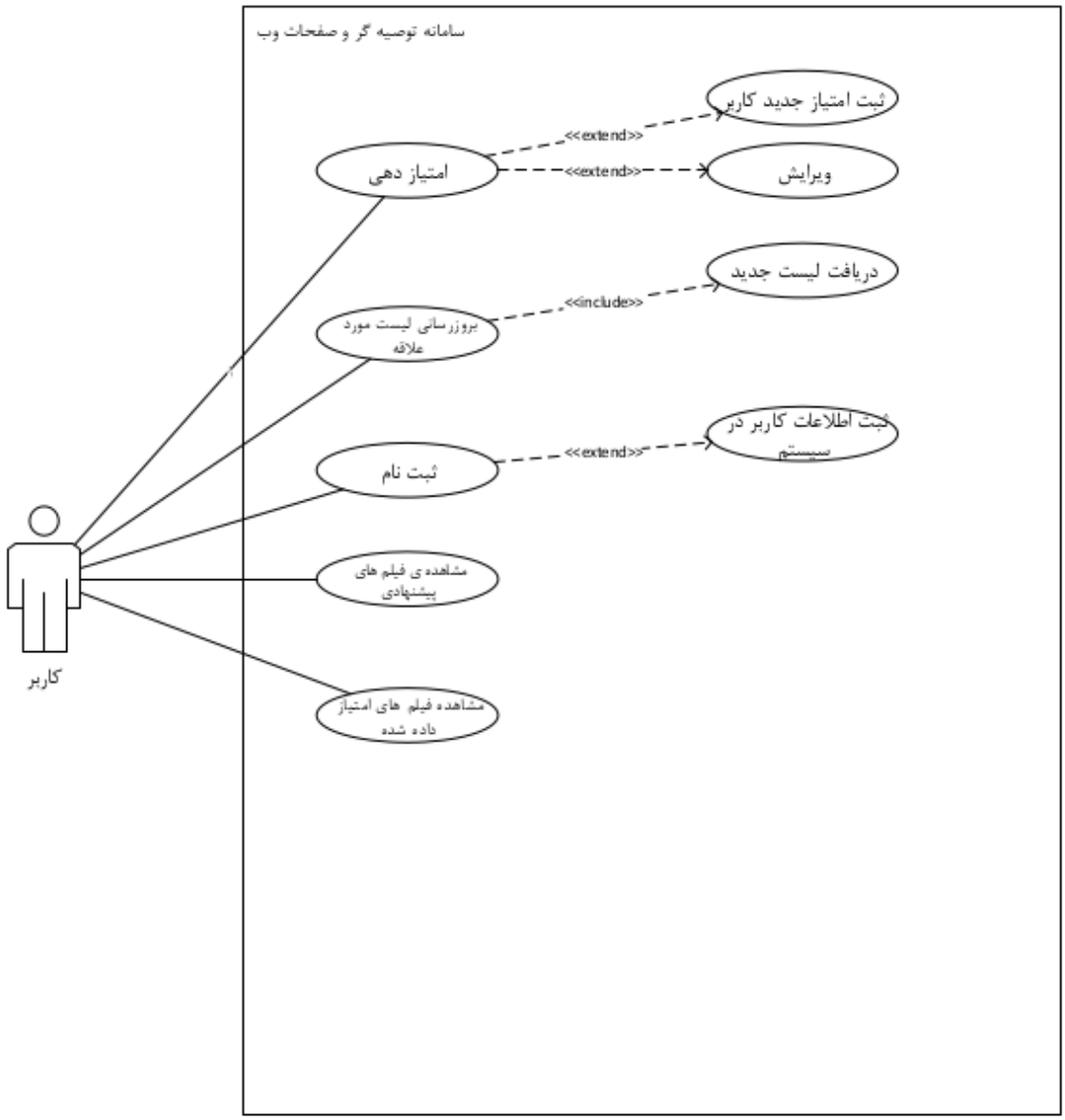
۳-۹-تحلیل سایت

در این بخش به تحلیل سیستم می‌پردازیم. در پروژه‌ی پیاده‌سازی شده سه سمت کاربر، استاد و مدیر کمیته پروژه مدنظر قرارگرفته‌اند که هرکدام دارای وظایف و اختیاراتی می‌باشند که به بررسی تک‌تک آن‌ها پرداخته می‌شود.

۳-۹-۱-نمودارها

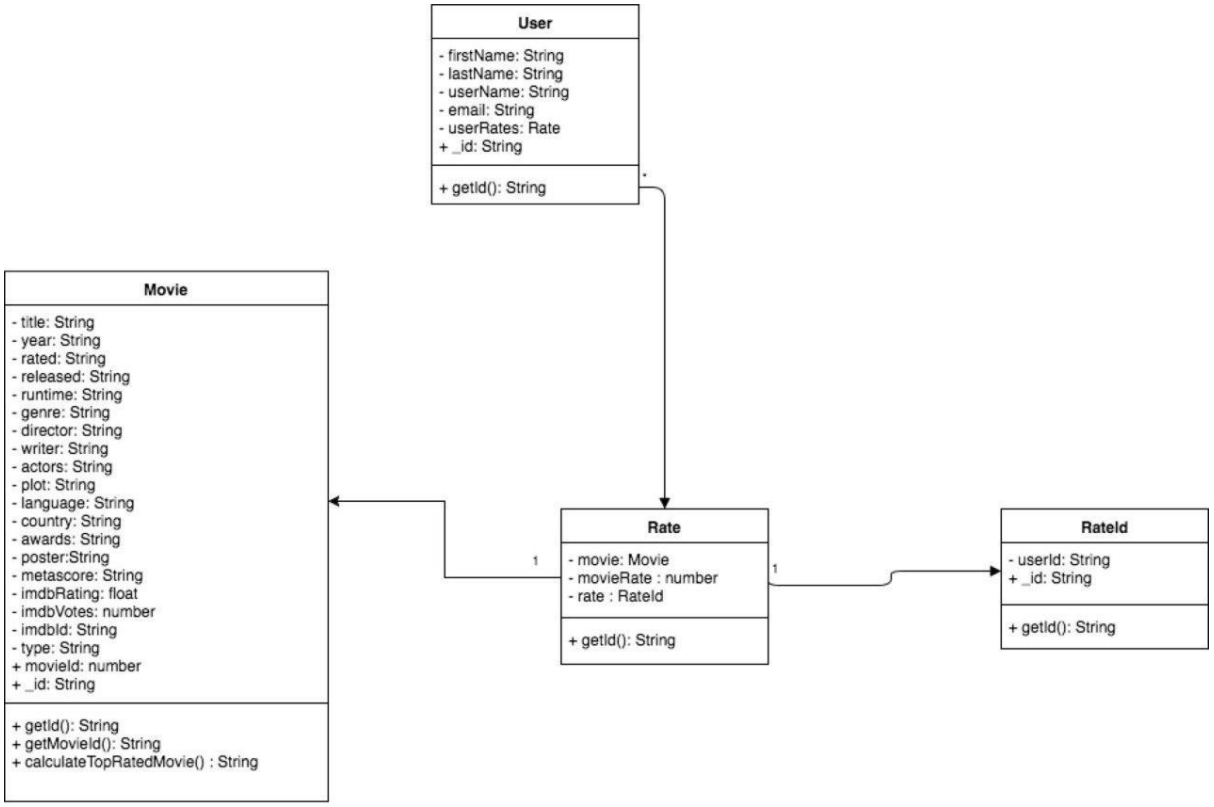
زبان مدل‌سازی یکنواخت یا UML زبانی برای مصورسازی، ساخت و مستندسازی سیستم‌های نرم‌افزاری و غیر نرم‌افزاری است که در این پروژه از برخی از نموداری آن استفاده‌شده است.

نمودار Use Case



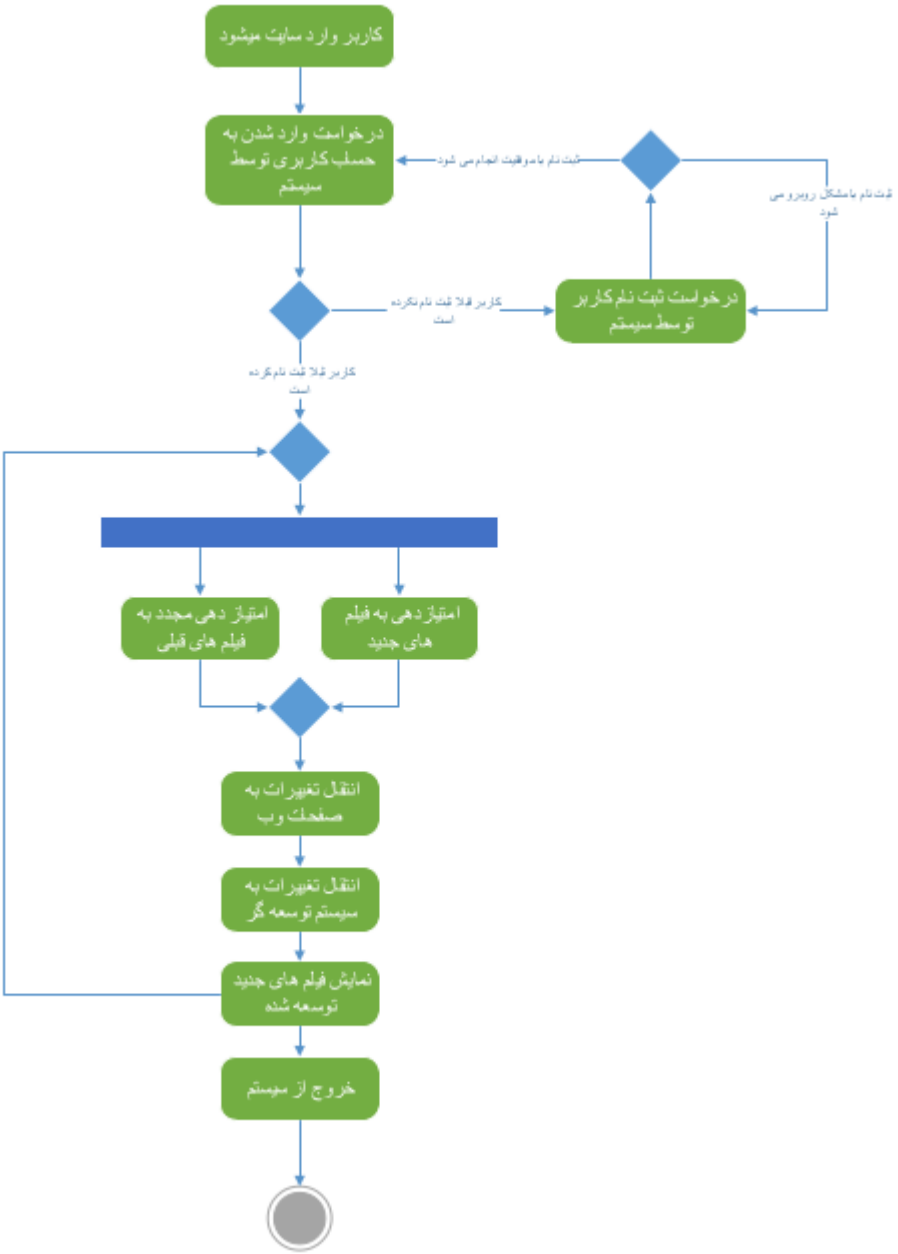
شکل ۳-۱- نمودارUse Case

نمودار Class



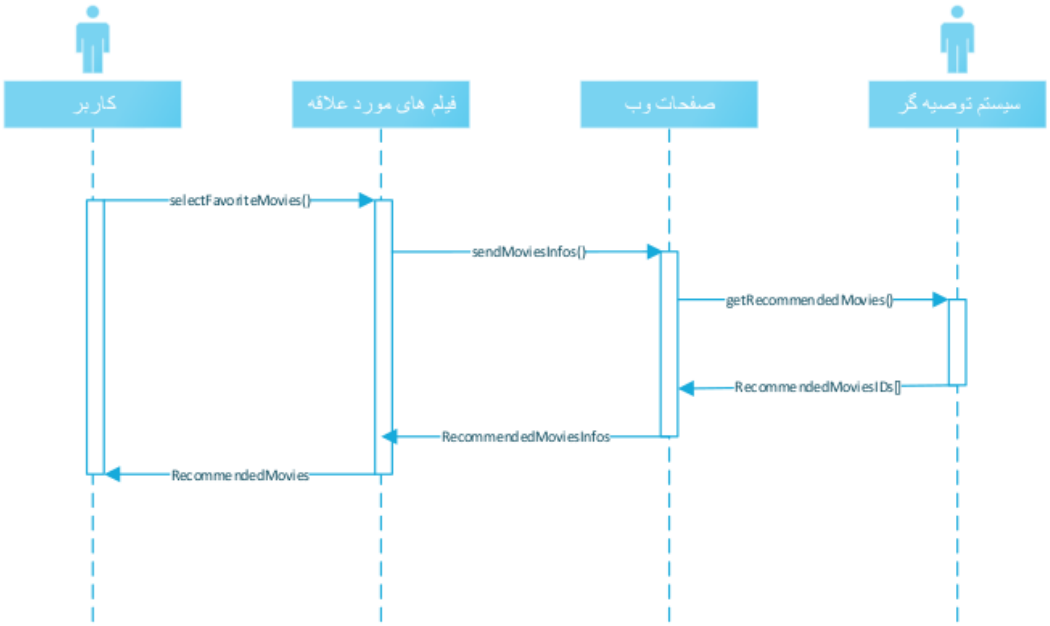
شکل ۳-۲- نمودار Class

نمودار Activity



شکل ۳-۳- نمودار Activity

نمودار Sequence



شکل ۳-۴- نمودار Sequence

۴-فصل چهارم ارزیابی سامانه ی توصیه گر

در این فصل یک مدل برای ارزیابی دقت پروژه ارائه می‌شود و موتور توصیه گر را از طریق تکنیک PCAGA-KM بررسی می‌نماییم. درنهایت نتایج تحلیل و توضیح داده می‌شوند. همه آزمایش‌های را بر روی کامپیوتر Intel i5 8GiG Ram و زبان برنامه‌نویسی python برای شبیه‌سازی مدل صورت گرفته‌اند.

۴-۱-مجموعه داده و معیار ارزیابی:

در این پروژه از مجموعه داده MovieLensاستفاده شد است که شامل یک‌میلیون امتیاز از ۶۰۰۰ کاربر برای ۴۰۰۰ فیلم می‌باشد که تا سال ۲۰۰۰ در این سایت ثبت‌نام کردن.

هر کاربر حداقل ۲۰ فیلم را امتیاز داده است. برای ارزیابی دقت پروژه از روش اعتبار سنجی چندلایه‌ای استفاده شد است به این صورت که دسته داده به‌صورت رندوم به ترتیب داده آموزشی و آزمایشی با نسبت ۹۰ درصد به ۱۰ درصد تقسیم می‌کنیم. از داده‌های آموزشی برای ساخت مدل آفلاین استفاده کردیم و داده‌های باقیمانده برای پیش‌بینی استفاده شدند. برای اثبات کیفیت موتور، از روش میانگین خطای مطلق[۶۷]، استفاده‌شده است و از precision، recall به‌عنوان محاسبات ارزیابی کمک گرفتیم که به‌طور گسترده‌ای برای مقایسه و محاسبه عملکرد سیستم‌های نظریه‌ای مورداستفاده قرار می‌گیرند.

۴-۱-۱-میانگین خطای مطلق

یک روش آماری اندازه‌گیری دقت است که میانگین تفاوت مطلق را بین امتیازهای پیش‌بینی‌شده و امتیازهای حقیقی بر روی کاربران آزمایش محاسبه می‌کند. بدیهی است که هر چه مقدار میانگین خطای مطلق پایین‌تر باشد پیش‌بینی‌ها صحیح‌تر است

[67] Mean Square Error

$$\mathrm{MAE} = \frac{\sum|\tilde{P}_{i,j} - r_{i,j}|}{M}$$

در این معادله M کل تعداد فیلم‌های پیش‌بینی‌شده است، ارائه‌دهنده مقدار پیش‌بینی‌شده برای کاربر i بر روی آیتم j است و $r_{i,j}$ نرخ‌بندی حقیقی است.

برای آگاهی از اینکه آیا کاربران به فیلم‌های پیشنهادی علاقه دارند، ما از روش اندازه‌گیری‌های دقت[۶۸] و صحت[۶۹] کمک گرفته‌ایم که به‌طور گسترده‌ای در سامانه‌های توصیه گر جهت ارزیابی سطح هوشمندی پیشنهادها استفاده می‌شود.

برای محاسبه این دو مقدار، ابتدا باید مقیاس امتیازات در موتور توصیه‌گر به حالت دودویی تبدیل شود. برای این کار می‌توان امتیازات را با استفاده از آستانه‌ای مشخص به دو قسمت مطلوب و نامطلوب تقسیم کرد که در سیستم‌های مختلف متفاوت است. ما این در این پروژه آستانه را برای امتیازات کاربران آزمایشی ۴ و برای امتیازهای پیش‌بینی‌شده توسط موتور توصیه گر ۳٫۶۵ در نظر گرفته‌ایم. این به این معناست که فیلم‌هایی که امتیاز ۴ یا ۵ از کاربر آزمایشی گرفته‌اند برای آن کاربر مطلوب و موردعلاقه بوده‌اند و فیلم‌هایی که امتیاز آن‌ها بیشتر از ۳٫۶۵ پیش‌بینی‌شده است فرض می‌شود برای کاربر مطلوب است. بعد از مقایسه نظیر به نظیر مقادیر واقعی امتیاز کاربر آزمایشی و مقادیر پیش‌بینی‌شده برای همان فیلم‌ها، برای هر فیلم امتیاز داده‌شده توسط کاربر آزمایشی یکی از حالت‌های زیر اتفاق می‌افتد.

TP: یعنی سیستم فیلم را مطلوب پیش‌بینی کرده و آن فیلم برای آن کاربر آزمایشی واقعاً مطلوب بوده است

TN: یعنی سیستم فیلم را نامطلوب پیش‌بینی کرده و آن فیلم برای آن کاربر آزمایشی واقعاً نامطلوب بوده است

---

[68] Presicion

Recall[69]

FP: یعنی سیستم فیلم را مطلوب پیش‌بینی کرده و آن فیلم برای آن کاربر آزمایشی واقعاً نامطلوب بوده است

FN: یعنی سیستم فیلم را نامطلوب پیش‌بینی کرده و آن فیلم برای آن کاربر آزمایشی واقعاً مطلوب بوده است

جدول زیر نشان دهدِ همین چهار حالت بالاست.

جدول ۴-۱- دقت و صحت

واقعاً \پیش‌بینی‌شده	مطلوب	نامطلوب
مطلوب	TP	FP
نامطلوب	FN	TN

حال می‌توان مفهوم و فرمول دقت و صحت رابین کرد:

۴-۱-۲-معیار Presicion

به حاصل تقسیم «تعداد مستندات بازیابی شدهٔ واقعاً باربط» بر «تعداد کل مستندات بازیابی شده» گفته می‌شود؛ که در این سیستم، می‌توان این‌گونه بیان کرد ((تعداد فیلم‌هایی که به‌درستی مطلوب بیش بینی شده‌اند در بین فیلم‌هایی که کاربر آزمایشی امتیاز داده)) تقسیم‌بر تعداد تمام فیلم‌هایی مطلوب پیش‌بینی‌شده‌اند در بین فیلم‌هایی که کاربر آزمایشی امتیاز داده است؛ که با توجه به جدول بالا می‌توان آن را با فرمول زیر نیز بیان کرد.

$$Percision = \frac{TP}{TP + F}$$

۴-۱-۳-معیار Recall

به حاصل تقسیم «تعداد مستندات بازیابی شدهٔ واقعاً باربط» بر «تعداد کل مستندات مرتبط موجود» گفته می‌شود؛ که در این سیستم، می‌توان این‌گونه بیان کرد ((تعداد فیلم‌هایی که به‌درستی مطلوب بیش بینی

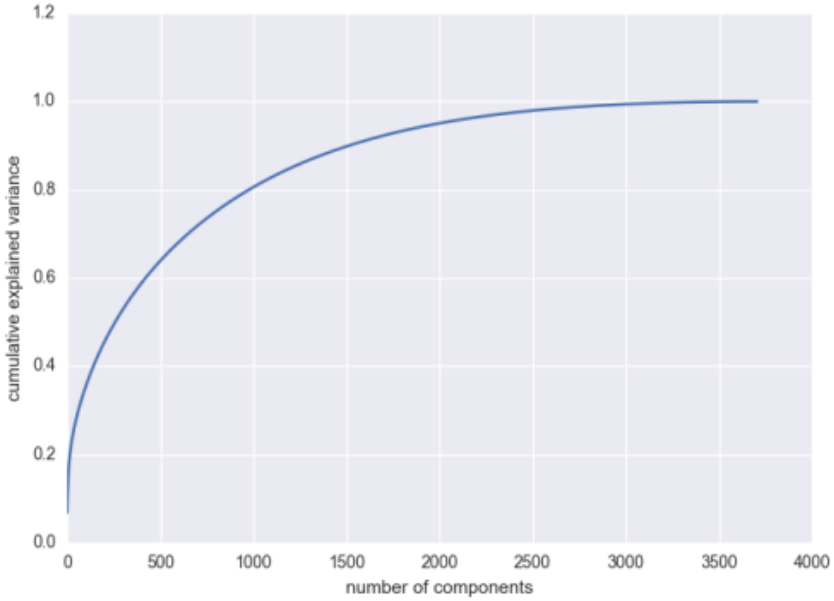
شده‌اند در بین فیلم‌هایی که کاربر آزمایشی امتیاز داده)) تقسیم‌بر ((تعداد تمام فیلم‌هایی که برای کاربر آزمایشی واقعاً مطلوب بوده))؛ که با توجه به جدول بالا می‌توان آن را با فرمول زیر نیز بیان کرد.

$$Recall = \frac{TP}{T}$$

۴-۲-بررسی الگوریتم‌های به‌کاررفته شده در پروژه

۴-۲-۱-تحلیل مؤلفه‌های اساسی

همان‌طور در فصل قبل در مورد الگوریتم تحلیل مؤلفه‌ی اساسی صحبت شد یکی از اهداف این الگوریتم کاهش ابعاد داده است به‌طوری‌که بیشترین مقدار پراکندگی داده‌ها حفظ شود. با توجه به نمودارهای زیر مشاهده می‌شود که با انتخاب ۲۵۰۰ کامپننت اول از ۳۷۰۶ کامپوننت موجود می‌توانیم از ۹۰ درصد پراکندگی داده برخوردار شویم.

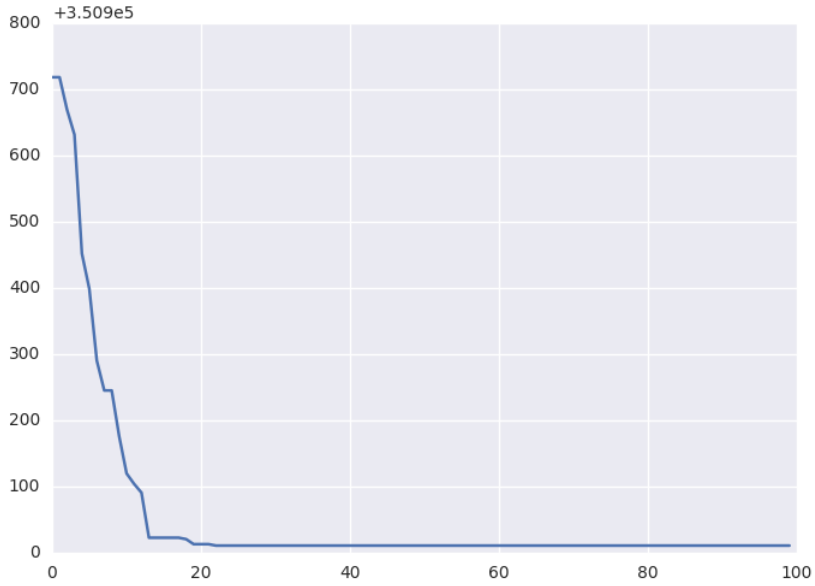


شکل ۴-۱- واریانس مؤلفه‌های اساسی

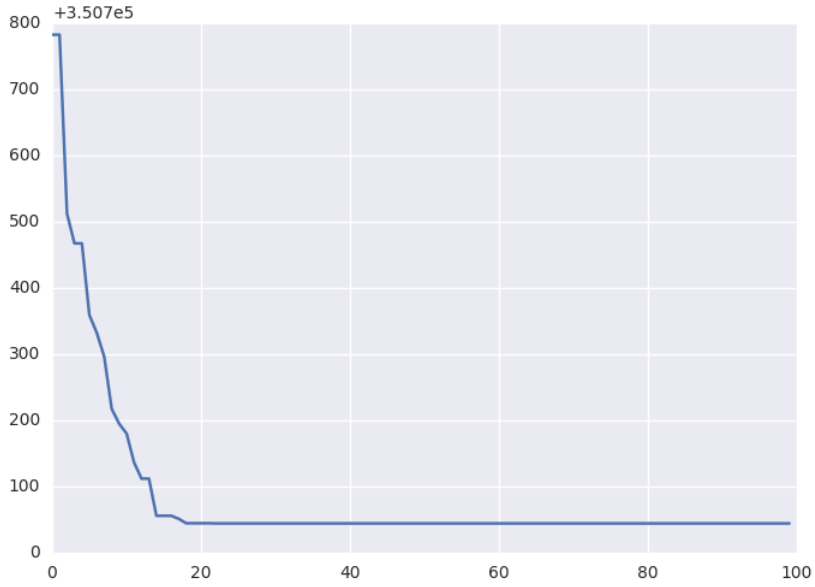
۴-۲-۲-الگوریتم ژنتیک

همان‌طور که در فصل قبل گفته شد الگوریتم ژنتیک استفاده‌شده در این پروژه به‌منظور انتخابیه‌ترین مقدارهای اولیه الگوریتم خوشه‌بندی با تعداد خوشه ۲۰ و ۲۵ به کار گرفته‌شده است. نتایج به دست

آمده تابع برازندگی برای هر کروموزوم بعد از ۱۰۰ گام که شرط پایان الگوریتم است به‌صورت زیر می‌باشد.



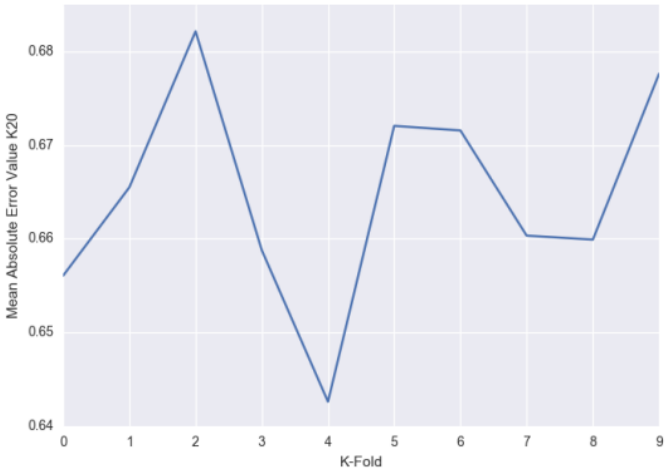
شکل ۴-۲- نمودار کاهش تابع برازندگی برای k=20



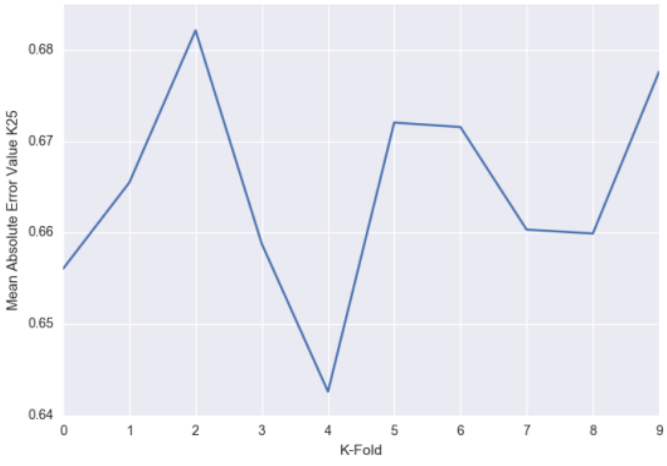
شکل ۴-۳- نمودار کاهش تابع برازندگی برای k=25

۴-۲-۳-الگوریتم پالایش گروهی

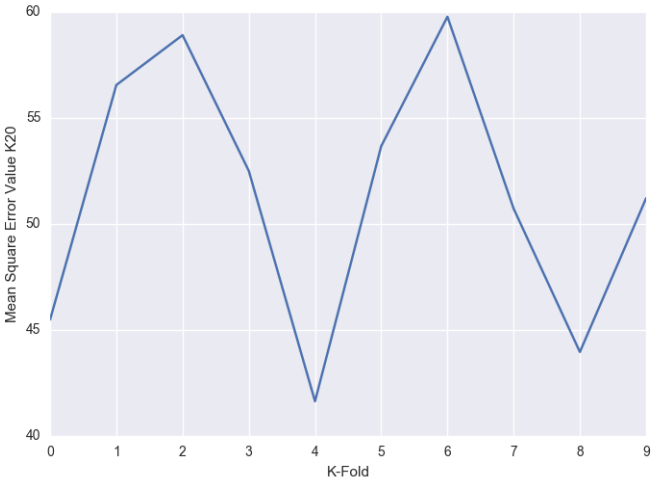
همان‌گونه توضیح داده شد معیارهای میانگین خطای مطلق، دقت و صحبت یکی از بهترین معیارهای ارزیابی این الگوریتم و سامانه‌های توصیه گر می‌باشد. نتایج ارزیابی این معیار ها برای نمایشش گروهی استفاده شده در این پروژه که با استفاده از الگوریتم‌های توضیح داده‌شده مانند تحلیلی مؤلفه اساسی و خوشه‌بندی بهینه‌شده است با نمودارهای زیر بیان می‌شود.



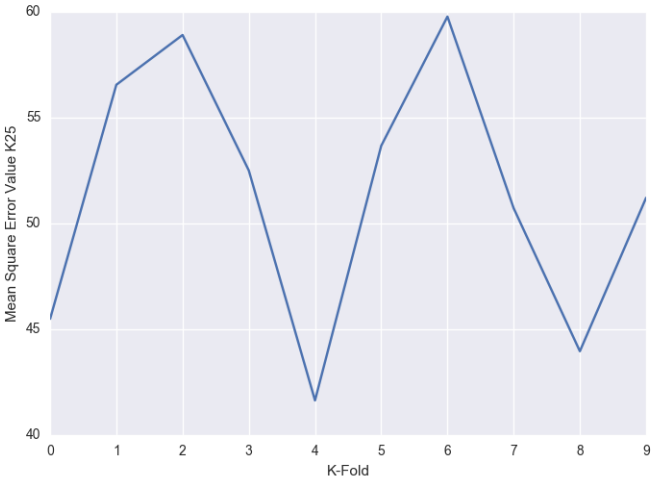
شکل ۴-۴- نمودار میانگین خطای مطلق برای k=20



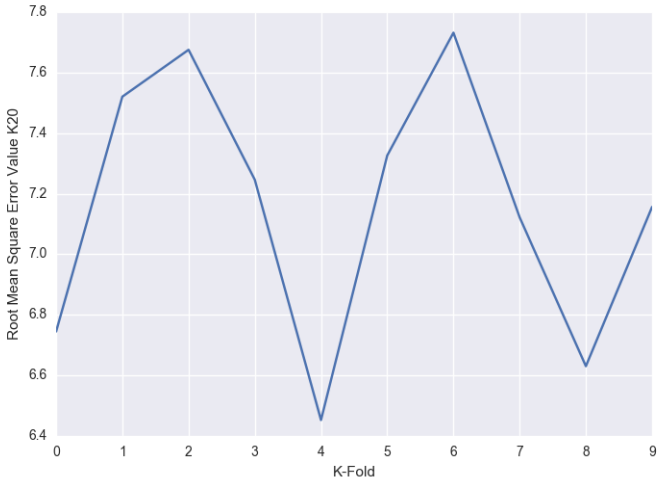
شکل ۴-۵- نمودار میانگین خطای مطلق برایk=25



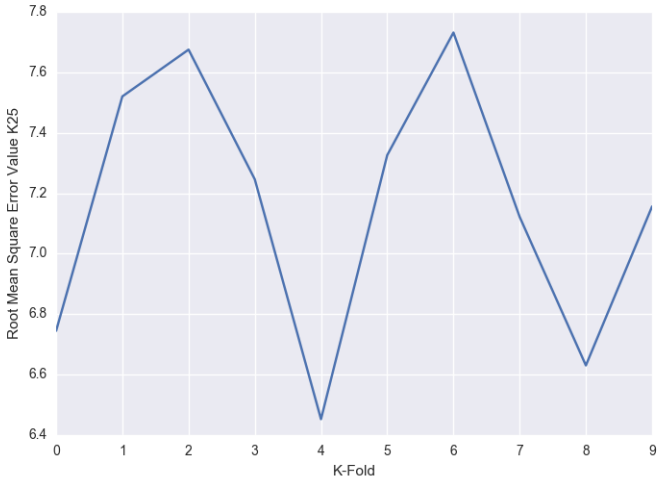
شکل ۴-۶- نمودار میانگین توان ۲ خطا برای$k=20$



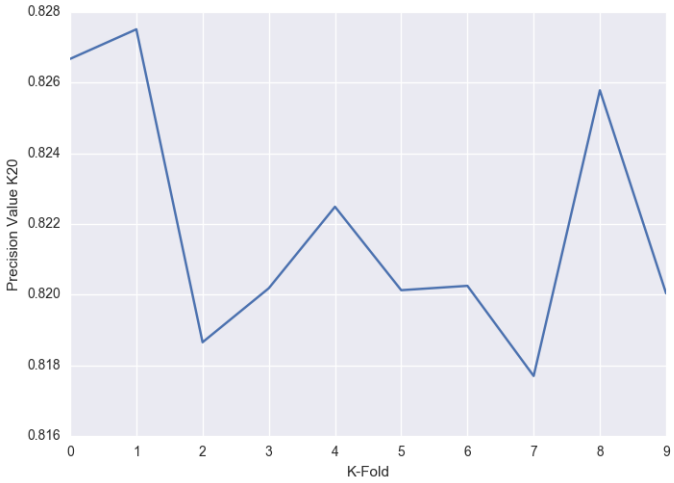
شکل ۴-۷- نمودار میانگین توان ۲ خطا برای$k=25$



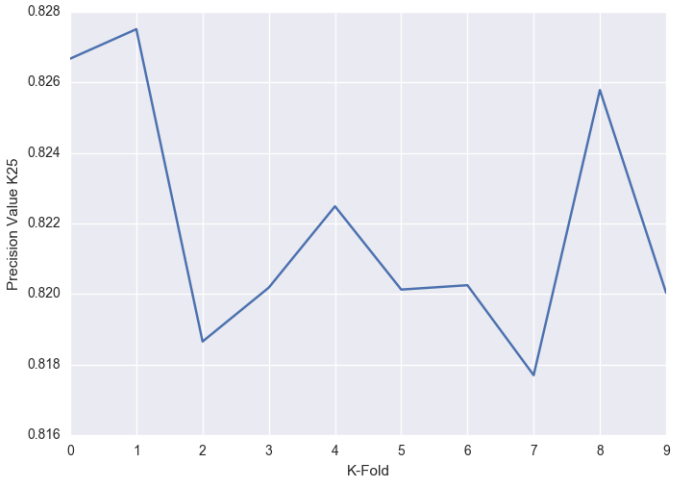
شکل ۴-۸- نمودار جذر میانگین توان ۲ خطا برای$k=20$



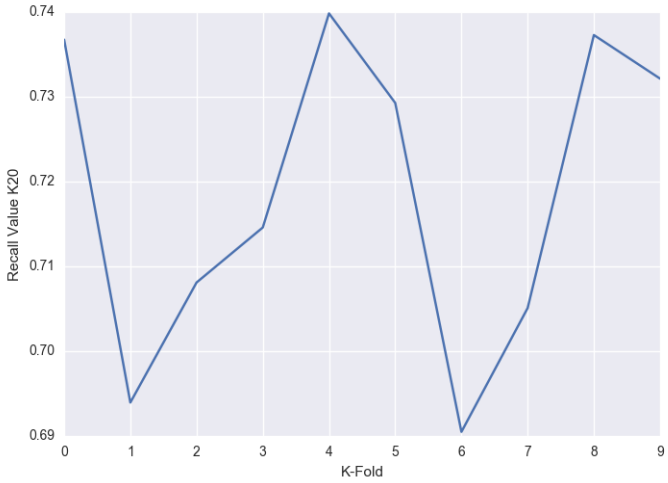
شکل ۴-۹- نمودار جذر میانگین توان ۲ خطا برای$k=25$



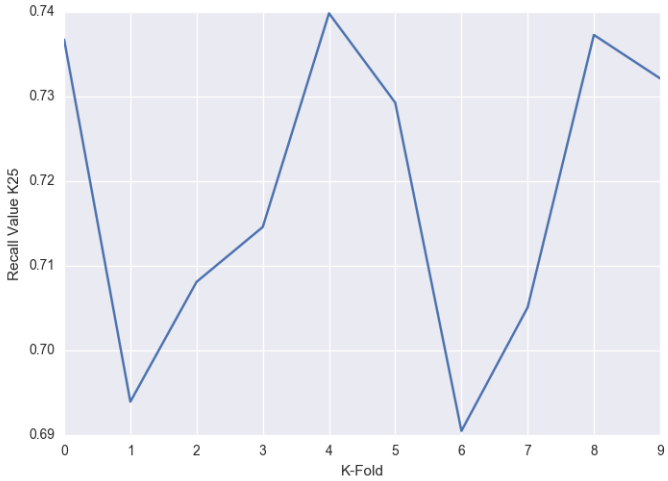
شکل ۴-۱۰- نمودار دقت برای k=20



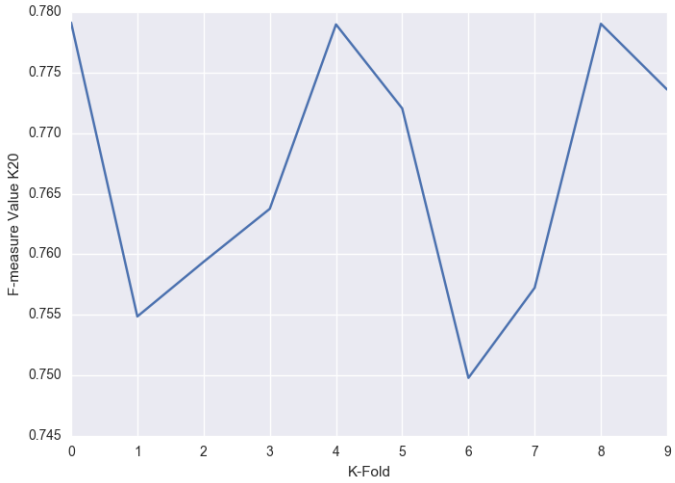
شکل ۴-۱۱- نمودار دقت برایk=25



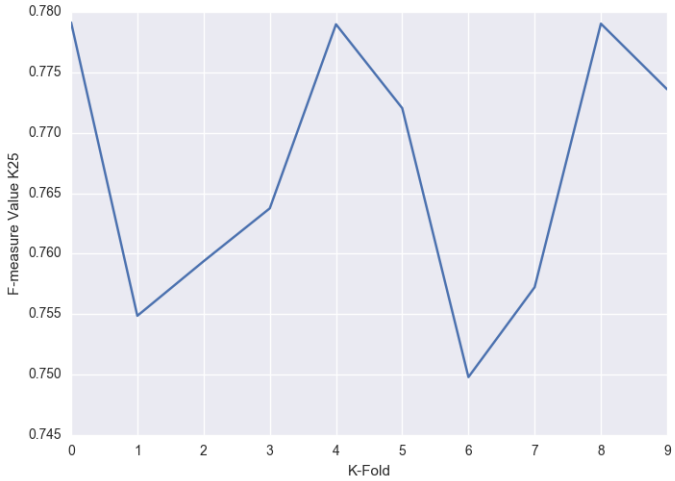
شکل ۴-۱۲- نمودار صحت برای$k=20$



شکل ۴-۱۳- نمودار صحت برای$k=25$



شکل ۴-۱۴- نمودار F-measure برای k=20



شکل ۴-۱۵- مقدار F-measure برای k=25

۴-۲-۴- بهترین نتیجه

جدول ۴-۲- بهترین نتایج ارزیابی سامانه روی مجموعه داده movielens

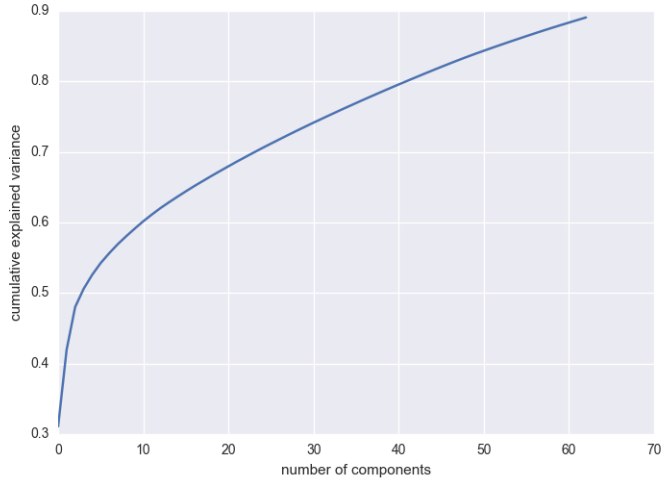
MAE	MSE	RMSE	Recall	precision	F1
۰٫۶۷۲۰	۵۳٫۶۷۲۰	۷٫۳۲۶۱	۰٫۷۲۹۲	۰٫۸۲۰۱	۰٫۷۷۲۰

۴-۳-بررسی روی مجموعه داده جک دانشگاه برکلی

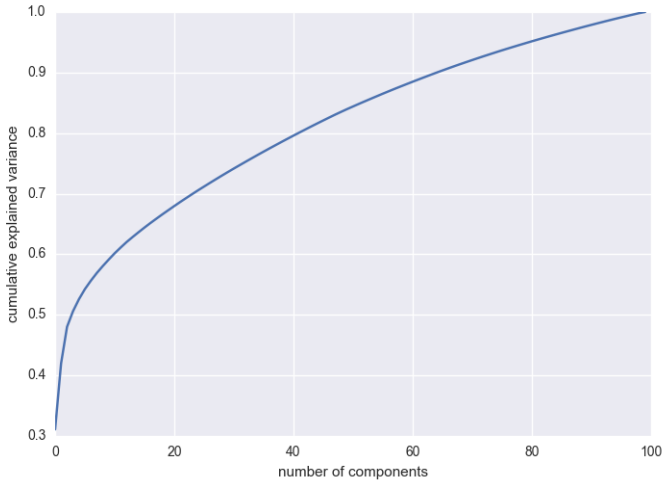
این مجموعه داده شامل ۲۵۰۰۰ کاربر و ۱۰۰ جک است که در آن هر کاربر حداقل به ۳۵ جک امتیاز داده است. امتیازات کاربران به‌صورت پیوسته بین بازه‌ی ۱۰ تا ۱۰- می‌باشد.

۴-۳-۱-تحلیل مؤلفه‌های اساسی

با توجه به نمودارهای زیر مشاهده می‌شود که با انتخاب ۶۳ کامپننت اول از ۱۰۰ کامپوننت موجود می‌توانیم از ۹۰ درصد پراکندگی داده برخوردار شویم.



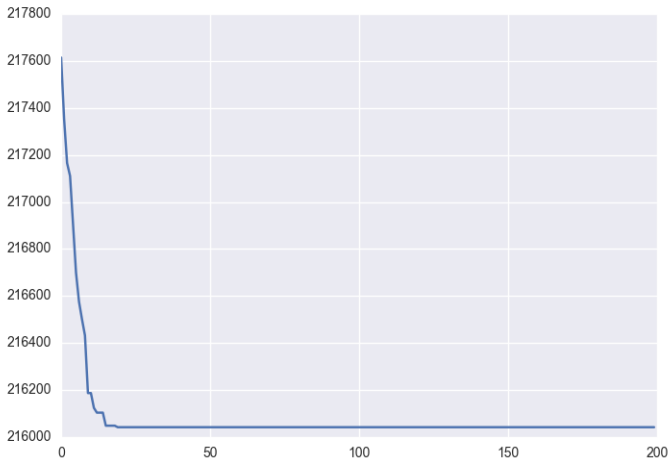
شکل ۴-۱۵- تحلیل مؤلفه اساسی بر روی مجموعه داده جک



شکل ۴-۱۷- تحلیل مؤلفه اساسی بر روی مجموعه داده جک یا ۶۳ مؤلفه

۴-۳-۲-الگوریتم ژنتیک

نتایج به دست آمده تابع برازندگی برای K ۲۰ برای هر کروموزوم بعد از ۲۰۰ گام که شرط پایان الگوریتم است به‌صورت زیر می‌باشد.



شکل ۴-۱۸- تابع برازندگی برای مجموعه داده جک

۴-۳-۳-الگوریتم پالایش گروهی

نتایج ارزیابی معیارها برای پالایش گروهی استفاده‌شده در این پروژه که با استفاده از الگوریتم‌های توضیح داده‌شده مانند تحلیلی مؤلفه اساسی و خوشه‌بندی بهینه‌شده است با جدول زیر بیان می‌شود.

جدول ۴-۳- بهترین نتایج ارزیابی سامانه روی مجموعه داده جک

MAE	MSE	RMSE	Recall	precision	F1
۲٫۹۴۵۱	۲۲۱٫۸۹۰۱	۱۴٫۸۹۵۹	۰٫۸۹۵۳	۰٫۵۶۳۷	۰٫۶۹۱۸

نتیجه‌گیری

در این پروژه ما روش پالایش گروهی ترکیبی مبتنی بر مدل را برای تولید فیلم‌های پیشنهادی پیاده‌سازی کردیم که روش‌های کاهش ابعاد داده را با الگوریتم‌های خوشه‌بندی ترکیب می‌کند. در محیطی که میزان خلوتی داده بالاست عمل انتخاب کاربران با شباهت بالاتر اساس امتیازات آن‌ها برای تولید پیشنهادهای باکیفیت بسیار حیاتی است. در این روش پیشنهادشده انتخاب ویژگی‌ها بر اساس تحلیل مؤلفه‌های اساسی در ابتدا بر روی تمام داده اجرا می‌شود. در مرحله‌ی بعد خوشه‌ها از روی بردارهایی با ابعادی کمتر که در مرحله قبل به‌دست‌آمده تولید می‌شود. به این صورت فضای اولیه کاربر به فضایی متراکم‌تر و مطمئن‌تر تبدیل می‌شود و برای انتخاب همسایگان مناسب استفاده می‌شود. علاوه بر این برای به دست آوردن بهترین همسایگان ما الگوریتم ژنتیک را برای بهینه کردن فرایند خوشه‌بندی کاربران مشابه استفاده کردیم.

به‌عنوان فعالیت‌های آینده ما برای افزایش دقت و سرعت پروژه از الگوریتم‌های خوشه‌بندی فازی و الگوریتم ژنتیک برای بهینه‌سازی الگوریتم پالایش گروهی استفاده خواهیم کرد و سیستم مبتنی بر دانش را برای حل مشکل شروع سرد و تغییر سلایق به آن اضافه می‌کنیم.

منابع

[1]http://stackoverflow.com/questions/41043234/implementing-euclidean-distance-based-formula-using-numpy

[2]http://stackoverflow.com/questions/40910492/conflict-in-recovering-features-names-after-apply-sklearn-pca

[3]http://stackoverflow.com/questions/40875469/add-new-vector-to-pca-new-space-data-python

[4]http://stackoverflow.com/questions/40842439/compute-eigenvalues-for-movielens-dataset

[5]http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.3488

[6]http://www.sciencedirect.com/science/article/pii/S0167865509002323

[7]https://en.wikipedia.org/wiki/K-means_clustering

[8]https://en.wikipedia.org/wiki/Principal_component_analysis

[9]https://en.wikipedia.org/wiki/Collaborative_filtering

[10]http://www.sciencedirect.com/science/article/pii/S1045926X14000901