مقاله پژوهشی

ارائه روشی جدید بر مبنای تجزیه ماتریس غیر منفی برای کاهش ابعاد

مهدی حسین‌زاده اقدم، مرتضی آنالویی و جعفر تنها

چکیده: یادگیری ماشین در طی دهه‌های گذشته به دلیل طیف گسترده کاربردهای آن مورد استفاده زیادی قرار گرفته است. در اکثر کاربردهای یادگیری ماشین مانند خوشه‌بندی و طبقه‌بندی، ابعاد داده‌ها زیاد می‌باشد و استفاده از روش‌های کاهش ابعاد داده ضروری است. تجزیه ماتریس غیر منفی با استفاده از استخراج ویژگی‌ها معنایی از داده‌های با ابعاد زیاد کاهش ابعاد را انجام می‌دهد و در تجزیه ماتریس غیر منفی فقط نحوه مدل‌سازی هر بردار ویژگی در ماتریس‌های تجزیه‌شده را در نظر می‌گیرد و روابط بین بردارهای ویژگی را نادیده می‌گیرد. ارتباطات میان بردارهای ویژگی، تجزیه بهتری را برای کاربردهای یادگیری ماشین فراهم می‌کنند. در این مقاله، یک روش بر مبنای تجزیه ماتریس غیر منفی برای کاهش ابعاد داده‌ها ارائه شده که محدودیت‌هایی را بر روی هر جفت‌بردارهای ویژگی با استفاده از معیارهای مبتنی بر فاصله ایجاد می‌کند. روش پیشنهادی از نرم فروبنیوس به عنوان تابع هزینه برای ایجاد قوانین به روز رسانی استفاده می‌کند. نتایج آزمایش‌ها روی مجموعه داده‌ها نشان می‌دهد که قوانین به روز رسانی ضربی ارائه‌شده، سریع همگرا می‌شوند و در مقایسه با الگوریتم‌های دیگر نتایج بهتری را ارائه می‌کنند.

کلیدواژه: **کاهش ابعاد، تجزیه ماتریسی غیر منفی، نرم فروبنیوس، قوانین به روز رسانی، خوشه‌بندی متن.**

۱- مقدمه

در بسیاری از مسایل مانند بازیابی اطلاعات، بینایی ماشین و شناسایی الگو، داده‌های ورودی ابعاد زیادی دارند که باعث می‌شود فرایند یادگیری با مشکل روبه‌رو شود [۱]. برای حل این چالش از روش‌های تجزیه ماتریسی به عنوان یک رویکرد جدید برای نمایش داده‌ها استفاده می‌شود و ایده اصلی در این روش‌ها پیداکردن دو یا چند ماتریس با ابعاد کم است، به طوری که بتوان با استفاده از ضرب آنها ماتریس اصلی را ساخت. روش‌های اولیه ارائه‌شده برای تجزیه ماتریس عبارت هستند از تجزیه- LU، تجزیه- QR و تجزیه مقادیر ویژه (SVD)[1].

SVD یکی از روش‌های پرکاربرد در تجزیه ماتریسی است. اگر ابعاد ماتریس اصلی (D) برابر $N\times M$ باشد، تجزیه مقادیر ویژه آن به صورت $D=W\sum H$ خواهد بود که در آن ماتریس W با ابعاد $N\times N$ و ماتریس H با ابعاد $M\times M$ ماتریس‌های متعامد هستند و $\sum$ با ابعاد $N\times M$ یک ماتریس قطری است که مقادیر قطر این ماتریس، مقادیر ویژه ماتریس D می‌باشد. با حذف بردارهای ویژه‌ای که مقادیر ویژه کمتری دارند می‌توان یک تقریب با رتبه کم از ماتریس اصلی به دست آورد. با در نظر گرفتن فضای اقلیدسی، این یک تقریب بهینه برای نمایش داده‌ها است و به همین دلیل SVD در خیلی از مسایل مانند شناسایی چهره و نمایش متون به کار برده شده است.

تحقیقات نشان داده که از نظر فیزیولوژیکی و روان‌شناسی، مغز انسان به صورت نمایش مبتنی بر بخش‌بندی است [۲]. از این رو محققین برای مدل‌سازی بهتر یادگیری مبتنی بر بخش‌بندی در شناسایی چهره و نمایش متون از روش تجزیه ماتریس غیر منفی (NMF)[2] به عنوان یک راهکار استفاده کردند [۳]. در روش NMF هدف، پیداکردن دو ماتریس غیر منفی است به طوری که حاصل‌ضرب آنها تقریبی از ماتریس اصلی باشد و از محدودیت غیر منفی بودن ماتریس‌های تجزیه‌شده برای نمایش مبتنی بر بخش‌بندی استفاده می‌شود، زیرا تنها اجازه عملیات جمع را می‌دهد. روش NMF نشان داده که در تشخیص چهره [۴] و خوشه‌بندی متن [۵] نسبت به SVD برتری دارد.

روش قوانین به روز رسانی ضربی[3] در NMF، همگرایی الگوریتم و غیر منفی بودن ماتریس‌های تجزیه‌شده را تضمین می‌کند [۶]. روش قوانین به روز رسانی ضربی به دلیل به دست آوردن نتایج خوب در مجموعه دادهای بزرگ، بسیار مورد استقبال قرار گرفته است ولی این روش تنها محدودیت غیر منفی بودن را به ماتریس‌های تجزیه‌شده اعمال می‌کند که در عمل ممکن است در بعضی از مسایل کافی نباشد. به همین دلیل محققین در این زمینه تحقیقاتی را برای پیشنهاد الگوریتم‌های جدید انجام داده‌اند [۷].

روش تجزیه ماتریس غیر منفی با محدودیت (CNMF)[4]، جزء مهم‌ترین روش‌ها در NMF است که محدودیت‌ها را به صورت عبارات قاعده‌مند اعمال می‌کند [۸]. در این روش بیشتر از نرم L۱ و نرم L۲ برای قاعده‌مندکردن استفاده می‌شود. از نرم L۱ برای کم‌کردن پراکندگی ماتریس‌های تجزیه‌شده و از نرم L۲ برای جلوگیری از بیش‌برازش استفاده می‌گردد. در روش CNMF محدودیت‌ها تنها بر روی سطرها یا ستون‌ها اعمال می‌شوند و روابط بین سطرها یا ستون‌ها در نظر گرفته نمی‌شود؛ در صورتی که این روابط در بسیاری از مسایل وجود دارند، مخصوصاً زمانی که ماتریس‌های تجزیه‌شده ویژگی‌ها را نشان می‌دهند. از این مسایل می‌توان به سیستم‌های توصیه‌گر و خوشه‌بندی تصویر و متن اشاره نمود.

خوشه‌بندی، فرایند تقسیم‌بندی مجموعه‌ای از داده‌ها در خوشه‌های متمایز بر اساس شباهت محتوای آنها است [۹] تا [۱۱] که از آن در سازمان‌دهی، استخراج، خلاصه‌سازی و بازیابی متون استفاده می‌شود

این مقاله در تاریخ ۱۱ فروردین ماه ۱۴۰۰ دریافت و در تاریخ ۴ آذر ماه ۱۴۰۰ بازنگری شد.

مهدی حسین‌زاده اقدم (نویسنده مسئول)، دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر، دانشگاه بناب، بناب، ایران، (email: mhaghdam@ubonab.ac.ir).

مرتضی آنالویی، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: analoui@iust.ac.ir).

جعفر تنها، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران، (email: jafar.tanha.pnu@gmail.com).

1. Singular Value Decomposition
2. Non-Negative Matrix Factorization
3. Multiplicative Updating Rules
4. Constrained Non-Negative Matrix Factorization

[۱۲]. خوشه‌بندی متن یک روش برای پیداکردن نزدیک‌ترین همسایگی برای هر متن در یک مجموعه از متون می‌باشد. در ابتدا از خوشه‌بندی متن برای بهبود دقت سامانه‌های بازیابی اطلاعات استفاده می‌شد ولی امروزه از خوشه‌بندی برای جستجوی یک مجموعه از متون و نرمال‌سازی نتایج داده‌شده توسط موتورهای جستجوی استفاده می‌گردد. اگرچه تحقیقات زیادی در زمینه خوشه‌بندی انجام شده است، ولی با این حال بسیاری از این روش‌ها نیاز به بهبود دارند. خوشه‌بندی طیفی یکی از روش‌هایی است که از بخش‌بندی ماتریس گراف استفاده می‌کند [۱۳]. این روش در مقایسه با روش‌های استاندارد خوشه‌بندی، توانایی شناسایی توزیع غیر محدب را دارد [۱۴] و [۱۵]. در این روش با توجه به مجموعه داده‌ها، یک گراف بدون جهت وزن‌دار برای پیداکردن خوشه‌بندی بهینه با در نظر گرفتن مقادیر ویژه و بردارهای ویژه گراف ساخته می‌شود. NMF با نگاشت ماتریس داده به فضای معنای می‌تواند روش مناسبی برای خوشه‌بندی باشد [۱۶].

در این مقاله یک روش جدید تجزیه ماتریس غیر منفی مبتنی بر نرم فروبنیوس (FNMF)[1] برای کاهش ابعاد و خوشه‌بندی ارائه شده است. در روش پیشنهادی، نرم فروبنیوس به عنوان تابع هزینه در نظر گرفته شده و قوانین به روز رسانی ضربی جدیدی برای تجزیه‌کردن ماتریس‌ها پیشنهاد گردیده است. در روش پیشنهادی، ویژگی‌های معنایی با استفاده از قوانین به هنگام‌سازی یاد گرفته می‌شوند و با استفاده از روابط بین بردارهای ویژگی، کاهش بهتری با حفظ محدودیت‌های اعمال‌شده انجام می‌گردد. برخلاف روش‌های دیگر، روش پیشنهادی بعد از تجزیه ماتریس اصلی، روابط موجود را حفظ می‌کند. اثبات همگرایی قوانین به روز رسانی پیشنهادشده و تحلیل پیچیدگی زمانی روش پیشنهادی در مقایسه با روش اولیه NMF نشان می‌دهد که اعمال محدودیت‌های جدید باعث افزایش پیچیدگی زمانی نمی‌شود و الگوریتم پیشنهادی سریع همگرا می‌گردد. نتایج پیاده‌سازی‌ها نشان‌دهنده برتری روش پیشنهادی در کاهش ابعاد و خوشه‌بندی نسبت به روش‌های دیگر است.

در بخش بعدی، پیش‌زمینه‌ای از روش NMF و کارهای مرتبط توضیح داده می‌شود. در بخش سوم، روش پیشنهادی همراه با اثبات همگرایی و تحلیل پیچیدگی زمانی ارائه می‌گردد. نتایج پیاده‌سازی‌ها در بخش چهارم آمده است و نتیجه‌گیری در بخش آخر بیان می‌شود.

۲- پیش‌زمینه

خوشه‌بندی، دسته‌بندی داده‌ها با ساختارهای معنایی یکسان در خوشه‌های مشابه است که هم می‌تواند به صورت سلسله‌مراتبی و هم به صورت بخش‌بندی باشد [۳]. روش NMF می‌تواند داده‌ها را بر اساس شباهت ویژگی‌های معنایی استخراج‌شده دسته‌بندی کند [۱۷]. ترکیب خطی بردارهای ویژگی استخراج‌شده در روش NMF غیر منفی است که نشان می‌دهد فقط می‌توان از عملگر جمع در ترکیب استفاده کرد؛ بنابراین هر داده می‌تواند به صورت ترکیب جمعی ویژگی‌های معنایی نشان داده شود و خوشه‌بندی می‌تواند در فضای ترکیب خطی بردارهای ویژگی انجام گردد. هر نمونه داده در ابتدا به صورت ویژگی‌ها نشان داده می‌شود. بعد از عملیات پیش‌پردازش در مجموعه داده‌ها، یک سری ویژگی‌ها می‌مانند که با استفاده از آنها بردار ویژگی‌ها ساخته می‌شود. همان طور که گفته شد، روش NMF، ماتریس اصلی D را به دو ماتریس W و ماتریس H تجزیه می‌کند که هر دوی آنها غیر منفی هستند $(D \approx WH)$.

NMF می‌تواند داده‌ها را در یک سیستم برداری جدید که بر اساس تحلیل ماتریس اولیه است نمایش دهد. در ماتریس اولیه، سطرها داده‌ها را نمایش می‌دهند و ستون‌ها معادل با ویژگی‌ها هستند. این ماتریس، تبدیل به دو ماتریس می‌شود که ماتریس‌های تجزیه‌شده، معمولاً کوچک‌تر از ماتریس اصلی هستند و از این طریق یک نمایش فشرده از ماتریس اصلی ایجاد می‌گردد. اگر سطرهای ماتریس اصلی، داده‌ها باشند آن گاه می‌توان ستون‌های ماتریس تجزیه‌شده دوم را به عنوان بردارهای پایه برای تشکیل داده‌ها در نظر گرفت که در این حالت هر سطر ماتریس تجزیه‌شده اول، میزان مشارکت هر پایه را در تشکیل یک نمونه داده نشان می‌دهد. در این روش بردارها، برچسب‌های خوشه‌ها را نشان می‌دهند و از این رو می‌توان عضویت هر نمونه داده در هر خوشه را به وسیله تعیین بزرگ‌ترین مؤلفه بردار معادل در ماتریس تجزیه‌شده اول تعیین کرد. مختصات هر نمونه داده در هر بردار همیشه غیر منفی است. نمایش داده‌ها به صورت ترکیب تجمعی از مفاهیم معنایی، درک بهتری را برای خوشه‌بندی فراهم می‌کند و به همین دلیل نگاشت NMF برای خوشه‌بندی بسیار مناسب است [۱۸]. قوانین به روز رسانی ضربی در روش اولیه NMF، غیر منفی بودن ماتریس‌های تجزیه‌شده و همگراشدن آنها به ماتریس داده اصلی را تضمین می‌کنند [۶]. همگراشدن این قوانین به کمینه محلی اثبات شده و این قوانین نتایج خوبی را در مجموعه داده‌های بزرگ تولید می‌کنند [۱۹]. با وجود این، روش اولیه NMF تنها محدودیت غیر منفی بودن را در تجزیه اعمال می‌کند که در بعضی کاربردها کافی نیست. به همین خاطر الگوریتم‌های جدیدی بر مبنای قوانین به روز رسانی ضربی برای افزایش کارایی ارائه شده است [۲۰] تا [۲۲]. تجزیه ماتریس غیر منفی منظم (RNMF)[2] از روش‌هایی است که محدودیت‌هایی را به صورت متغیرهای منظم در فرایند تجزیه اعمال می‌کند [۲۳]. در این روش، این محدودیت‌ها تنها بر روی هر کدام از بردارهای ویژگی اعمال می‌شود و روابط بین آنها را در نظر نمی‌گیرد. این روابط در بسیاری از کاربردها مانند خوشه‌بندی متن وجود دارند.

نرم فروبنیوس تابع هزینه‌ای است که اکثراً برای ارزیابی تقریب ماتریس‌های تجزیه‌شده به کار می‌رود. نرم فروبنیوس مربع فاصله اقلیدسی بین دو ماتریس است

$$Frobenius-norm = \|D-WH\|^2 = \sum_w \sum_h (D_{wh}-(WH)_{wh})^2 \quad (۱)$$

این تابع نسبت به تک‌تک ماتریس‌های W و H محدب[3] است ولی نسبت به هر دو محدب نیست [۲۴]، پس پیداکردن کمینه سراسری برای این تابع عملی نیست. با این حال روش‌های زیادی برای پیداکردن کمینه محلی ارائه شده است. لی و سیونگ الگوریتمی را برای کمینه‌کردن این تابع ارائه کردند [۶] که قوانین به روز رسانی آنها برای کمینه‌کردن به صورت زیر است

$$W_{wf} \leftarrow W_{wf} \times \frac{(DH^T)_{wf}}{(WHH^T)_{wf}} \quad (۲)$$

$$H_{fh} \leftarrow H_{fh} \times \frac{(W^T D)_{fh}}{(W^T WH)_{fh}} \quad (۳)$$

1. Frobenius-Norm Non-Negative Matrix Factorization

2. Regularized Non-Negative Matrix Factorization

3. Convex

لی و سیونگ اثبات کردند که این الگوریتم به روز رسانی تکراری می‌تواند کمینه محلی را برای تابع هزینه در (۱) پیدا کند [۶]. در روش NMF ستون hام ماتریس D به وسیله ترکیب خطی بردارهای سطری ماتریس W که با ستون hام ماتریس H وزن‌دهی شده‌اند ساخته می‌شود ($D_h \approx WH_h$). ماتریس W را می‌توان به عنوان پایه‌ای برای بهینه‌سازی ترکیب خطی ماتریس D در نظر گرفت. از آنجایی که تعداد محدودی از بردارهای پایه برای نمایش تعداد زیادی بردار به کار برده می‌شوند، از این رو پیش‌بینی قابل قبول وقتی به دست می‌آید که بردارهای پایه بتوانند ساختار معنایی داده‌ها را نشان دهند. معمولاً تعداد ویژگی‌های معنایی، کمتر از تعداد داده‌ها و ویژگی‌های اولیه در نظر گرفته می‌شوند. در واقع روش NMF، ماتریس اولیه با ابعاد زیاد را با بردارهایی در ابعاد کم نشان می‌دهد. برخلاف روش‌های دیگر تجزیه ماتریسی، محدودیت غیر منفی بودن، تنها اجازه عملیات جمع‌کردن در ماتریس‌های تجزیه‌شده را می‌دهد و عملیات منها در این روش مجاز نیست.

چندین روش بر مبنای NMF برای خوشه‌بندی ارائه شده است. در [۵] از روش اولیه NMF برای خوشه‌بندی متن استفاده گردیده و در این مقاله نشان داده شده که نتایج NMF از تجزیه مقدار ویژه (SVD) بهتر است. در [۱۸] یک الگوریتم برای یادگیری ویژگی‌های معنایی با استفاده از قوانین به روز رسانی آمده است. بیشتر روش‌های ارائه‌شده در مقالات، مشکل پیچیدگی زمان اجرا و مصرف حافظه زیاد را دارند.

۳- روش پیشنهادی

در این مقاله یک روش جدید به نام FNMF برای یادگیری ساختار معنایی ارائه شده است. این روش با استفاده از شباهت بردارهای ویژگی، ماتریس اصلی را تجزیه می‌کند و تضمین می‌کند که شباهت بین بردارها بعد از نگاشت حفظ شود. در این مقاله از شباهت کسینوسی برای اندازه‌گیری شباهت داده‌ها استفاده شده و همچنین از نرم فروبنیوس به عنوان محدودیت برای شباهت بین دو بردار ویژگی استفاده گردیده است. روش پیشنهادی با در نظر گرفتن محدودیت و همچنین نرم فروبنیوس به عنوان تابع هزینه، تابع هزینه زیر را برای تجزیه ماتریس به کار می‌برد

$$Cost\_Function(CF) = \sum_w \sum_h (D_{wh} - (WH)_{wh})^2 + \lambda \sum_w \sum_t \sum_f (W_{wf} - W_{tf})^2 S_{wt} \quad (۴)$$

در این فرمول ماتریس‌های W و H و پارامتر λ نامنفی هستند و λ میزان تأثیر محدودیت را در تابع هزینه مشخص می‌کند. اگر دو نمونه داده w و t شبیه به هم باشند، انتظار می‌رود که بردارهای معادل آنها (W_w و W_t) به وسیله کمینه‌کردن تابع هزینه به هم شبیه شوند. S ماتریس شباهت بین بردارهای ویژگی است که توسط شباهت کسینوسی به دست می‌آید و یک ماتریس نامنفی می‌باشد. این تابع هزینه نسبت به هر کدام از ماتریس‌های W و H محدب است ولی نسبت به هر دو محدب نیست، بنابراین پیداکردن کمینه سراسری برای این تابع هزینه غیر ممکن می‌باشد. به همین دلیل در این مقاله الگوریتمی تکرارشونده برای پیداکردن کمینه محلی ارائه شده است. در این بخش در ابتدا نحوه کمینه‌کردن تابع هزینه توضیح داده می‌شود.

۳-۱ قوانین به روز رسانی پیشنهادی برای کمینه‌کردن تابع هزینه

برای کمینه‌کردن تابع هزینه در (۴)، نیاز به قوانین به روز رسانی برای ماتریس‌های W و H می‌باشد. در این مقاله برای پیداکردن این قوانین از تابع لاگرانژ استفاده شده است

$$\ell = \sum_w \sum_h (D_{wh} - (WH)_{wh})^2 + \lambda \sum_w \sum_t \sum_f (W_{wf} - W_{tf})^2 S_{wt} + \sum_w \sum_f \alpha_{wf} W_{wf} + \sum_f \sum_h \beta_{fh} H_{fh} \quad (۵)$$

به طوری که α و β ضرایب لاگرانژ برای محدودیت‌های $W \geq 0$ و $H \geq 0$ هستند. مشتق جزئی لاگرانژ نسبت به W_{wf} و H_{fh} برابر است با

$$\frac{\partial \ell}{\partial W_{wf}} = -2(DH^T)_{wf} + 2(WHH^T)_{wf} + 2\lambda \sum_t (W_{wf} - W_{tf}) S_{wt} + \alpha_{wf} \quad (۶)$$

$$\frac{\partial \ell}{\partial H_{fh}} = -2(W^T D)_{fh} + 2(W^T WH)_{fh} + \beta_{fh} \quad (۷)$$

با استفاده از شرایط KKT، $\alpha_{wf} W_{wf} = 0$ و $\beta_{fh} H_{fh} = 0$ است و قوانین به روز رسانی زیر به دست می‌آید

$$W_{wf} \leftarrow W_{wf} \times \frac{(DH^T)_{wf} + \lambda (SW)_{wf}}{(WHH^T)_{wf} + \lambda W_{wf} \sum_t S_{wt}} \quad (۸)$$

$$H_{fh} \leftarrow H_{fh} \times \frac{(W^T D)_{fh}}{(W^T WH)_{fh}} \quad (۹)$$

۳-۲ اثبات همگرایی قوانین به روز رسانی پیشنهادی

اثبات همگرایی در این مقاله شبیه اثبات همگرایی در مقاله لی و سیونگ است [۶]. با توجه به قوانین به روز رسانی (۸) و (۹) می‌توان قضیه زیر را در نظر گرفت.

قضیه: تابع هزینه در (۴) نسبت به قوانین به روز رسانی (۸) و (۹) غیر افزایشی است.

اگر $\lambda = 0$ باشد، آن گاه قوانین به روز رسانی (۸) و (۹) شبیه قوانین اولیه NMF خواهند بود. در این مقاله برای اثبات همگرایی از یک تابع کمکی استفاده شده است.

تعریف: اگر $G(x,x^t) \geq F(x)$ و $G(x,x) = F(x)$ باشد، $G(x,x^t)$ یک تابع کمکی برای $F(x)$ است.

لم ۱: اگر $G(x,x^t)$ یک تابع کمکی برای $F(x)$ باشد، آن گاه $F(x)$ تحت قانون به روز رسانی (۱۰) غیر افزایشی است

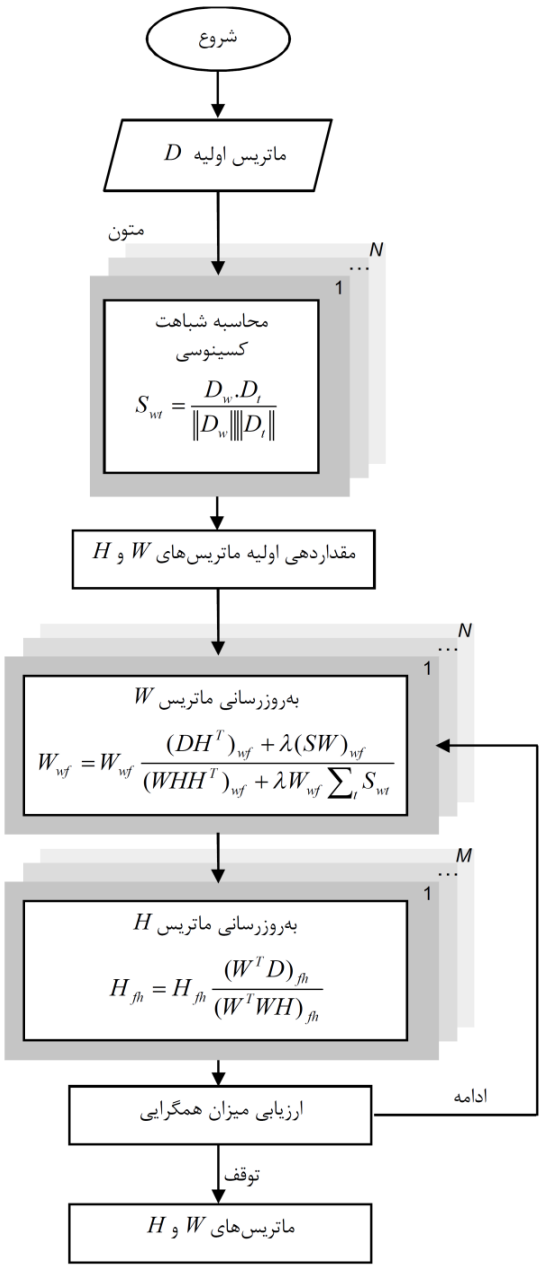
$$x^{t+1} = \arg\min_x G(x,x^t) \quad (۱۰)$$

اثبات: $F(x^{t+1}) \leq G(x^{t+1},x^t) \leq G(x^t,x^t) = F(x^t)$

اگر x^t یک کمینه محلی برای $G(x,x^t)$ باشد، آن گاه $F(x^{t+1}) = F(x^t)$ برقرار است. پس یک سری تقریب‌ها را می‌توان با تکرار قانون به روز رسانی (۱۰) برای همگراشدن تابع هزینه به دست آورد:

$$F(x_{\min}) \leq \ldots \leq F(x^{t+1}) \leq F(x^t) \leq \ldots \leq F(x^0)$$

با در نظر گرفتن یک تابع کمکی مناسب می‌توان نشان داد که قانون به روز رسانی در (۸) یک تابع به روز رسانی در (۱۰) است.



شکل ۱: روند روش پیشنهادی برای خوشه‌بندی.

لم ۲: تابع

$$G(W,W^t)=CF(W^t)+CF'(W^t)(W-W^t)_{wf}+\frac{(W^tHH^T)_{wf}+\lambda W^t_{wf}\sum_t S_{wt}}{W^t_{wf}}(W-W^t)^2_{wf} \quad (۱۱)$$

یک تابع کمکی برای $CF(W)$ در (۴) است که $CF(W)$ اشاره به قسمت W تابع CF دارد.

اثبات: $G(W,W)=CF(W)$ واضح است. برای پیداکردن $G(W,W^t)\geq CF(W)$، (۱۱) با سری تیلور بسط داده‌شده روی $CF(W)$ مقایسه شده است

$$CF(W)=CF(W^t)+CF'(W^t)(W-W^t)_{wf}+((HH^T)_{ff}+\lambda\sum_t S_{wt})(W-W^t)^2_{wf} \quad (۱۲)$$

برای اثبات $G(W,W^t)\geq CF(W)$ نیاز به اثبات نامعادله زیر است

$$\frac{(W^tHH^T)_{wf}+\lambda W^t_{wf}\sum_t S_{wt}}{W^t_{wf}}\geq (HH^T)_{ff}+\lambda\sum_t S_{wt} \quad (۱۳)$$

برای اثبات این نامعادله می‌توان از نامعادله زیر استفاده کرد

$$(W^tHH^T)_{wf}=\sum_s W^t_{ws}(HH^T)_{sf}\geq W^t_{wf}(HH^T)_{ff} \quad (۱۴)$$

پس نامعادله $G(W,W^t)\geq CF(W)$ برقرار است و به این ترتیب می‌توان یک تابع کمکی برای قانون به روز رسانی (۹) تعیین کرد.

لم ۳: تابع

$$G(H,H^t)=CF(H^t)+CF'(H^t)(H-H^t)_{fh}+\frac{(W^TWH)_{fh}}{H^t_{fh}}(H-H^t)^2_{fh} \quad (۱۵)$$

یک تابع کمکی برای $CF(H)$ در (۴) است که $CF(H)$ اشاره به قسمت H تابع CF دارد

$$CF(H)=CF(H^t)+CF'(H^t)(H-H^t)_{fh}+(W^TW)_{ff}(H-H^t)^2_{fh} \quad (۱۶)$$

لم ۳ اثباتی شبیه لم ۲ دارد و از تکرار آن صرف نظر شده است. با در نظر گرفتن این لم‌ها همگرایی قضیه اثبات می‌شود.

اثبات قضیه: جایگذاری $G(W,W^t)$ در (۱۰) با (۱۱) و (۱۵) منجر به ایجاد قوانین به روز رسانی زیر می‌شود

$$W^{t+1}_{wf}=\arg\min_w G(W,W^t)=W^t_{wf}-W^t_{wf_x}\frac{CF'(W^t)}{2(W^tHH^T)_{wf}+2\lambda W^t_{wf}\sum_t S_{wt}}=W^t_{wf_x}\frac{(DH^T)_{wf}+\lambda(SW)_{wf}}{(W^tHH^T)_{wf}+\lambda W^t_{wf}\sum_t S_{wt}} \quad (۱۷)$$

$$H^{t+1}_{fh}=\arg\min_h G(H,H^t)=H^t_{fh}-H^t_{fh_x}\frac{CF'(H^t)}{2(W^TWH^t)_{fh}}=H^t_{fh_x}\frac{(W^TD)_{fh}}{(W^TWH^t)_{fh}} \quad (۱۸)$$

از آنجایی که (۱۱) و (۱۵) توابع کمکی هستند، پس می‌توان نتیجه گرفت که $CF(W)$ و $CF(H)$ تحت این قوانین به روز رسانی، غیر افزایشی هستند و در کل تابع هزینه در (۴) غیر افزایشی است. شکل ۱ روش پیشنهادی برای خوشه‌بندی را نشان می‌دهد.

۳-۳ تحلیل پیچیدگی زمانی روش پیشنهادی

در روش پیشنهادی چون تابع هزینه توسط قوانین به روز رسانی به صورت تکرارشونده کمینه می‌شود، از این رو محاسبه پیچیدگی زمانی آن مهم است. برای محاسبه ماتریس شباهت اگر N تعداد داده‌ها و M تعداد ویژگی‌ها باشد، آن گاه پیچیدگی زمانی آن برابر با $O(N^2M)$ خواهد بود. از طرف دیگر با توجه به ابعاد ماتریس W که برابر با $N\times F$ و ابعاد ماتریس H که برابر با $F\times M$ می‌باشد (F تعداد ویژگی‌های معنایی یا همان تعداد خوشه‌ها است)، پیچیدگی روش پیشنهادی در هر تکرار برابر با $O(N\times M\times F)$ خواهد بود که برابر با روش اولیه NMF می‌باشد. لازم به ذکر است که تعداد تکرار قوانین به روز رسانی در الگوریتم، وابسته به سرعت همگرایی می‌باشد. از آنجایی که

روش پیشنهادی شباهت را به عنوان محدودیت در قوانین به روز رسانی وارد می‌کند، بر اساس نتایج به دست آمده، روش پیشنهادی سرعت همگرایی زیادی دارد و در تعداد تکرارهای کمتری نسبت به الگوریتم اولیه NMF همگرا می‌شود.

٤- نتایج آزمایش‌ها

در این بخش آزمایش‌های انجام‌شده جهت نشان‌دادن کارایی روش پیشنهادی در کاهش ابعاد و خوشه‌بندی ارائه گردیده است. در ابتدا روش‌های مقایسه‌شده توضیح داده می‌شوند و سپس معیارهای ارزیابی بیان می‌گردند و نهایتاً نتایج آمده است.

٤-١ روش‌های مقایسه‌شده

چندین روش خوشه‌بندی برای مقایسه با روش پیشنهادی بررسی شده که این روش‌ها به شرح زیر هستند:

K-means: در این روش، داده‌ها در خوشه‌هایی که واریانس برابر دارند گروه‌بندی می‌شوند و هدف، کمینه‌کردن جمع مربعات فواصل نقاط داده‌ها در داخل هر خوشه است. در این روش، تعداد خوشه‌ها باید از قبل مشخص باشد. این الگوریتم مقیاس‌پذیر است و در مجموعه داده‌های بزرگ می‌توان از آن استفاده کرد و هم‌اکنون در کاربردهای زیادی از این روش استفاده می‌شود [۲۵] و [۲۶].

خوشه‌بندی طیفی[1]**:** این روش در ابتدا با استفاده از مقادیر ویژه ماتریس شباهت، کاهش ابعاد انجام می‌دهد و سپس خوشه‌بندی می‌کند. روش کلی خوشه‌بندی طیفی به کارگیری یک خوشه‌بندی استاندارد بر روی بردارهای ویژه ماتریس لاپلاسین ماتریس شباهت است [۲۷] و [۲۸].

خوشه‌بندی Agglomerative: الگوریتم‌های خوشه‌بندی سلسله‌مراتبی، خوشه‌های تودرتو را به وسیله ادغام یا تقسیم‌کردن پی‌درپی ایجاد می‌کنند [۲۹] و سلسله‌مراتب خوشه‌ها به صورت درخت یا دندروگرام[2] نشان داده می‌شود. ریشه درخت یک خوشه تنها است که هم نمونه‌ها را در بر می‌گیرد و برگ‌های آن خوشه‌هایی است که تنها یک نمونه دارند [۳۰]. خوشه‌بندی agglomerative یک خوشه‌بندی سلسله‌مراتبی است که از روش بالا به پایین استفاده می‌کند. این الگوریتم در ابتدا هر نمونه را یک خوشه در نظر می‌گیرد و به صورت پی‌درپی آنها را با هم ادغام می‌کند و از معیار پیوند[3] برای ادغام خوشه‌ها استفاده می‌کند.

NMF: در پیاده‌سازی‌ها از روش اولیه NMF برای مقایسه استفاده شده است [۶] و [۳۱].

٤-٢ معیارهای ارزیابی

اکثراً از دو معیار برای ارزیابی کارایی الگوریتم‌های خوشه‌بندی استفاده می‌شود. اولین معیار، اطلاعات متقابل[4] (MI) بین دو خوشه‌بندی است. این معیار شباهت بین دو نوع خوشه‌بندی بر روی یک داده را نشان می‌دهد [۳۲]. برای دو خوشه‌بندی U و V اطلاعات متقابل به صورت زیر محاسبه می‌شود

$$MI(U,V)=\sum_{i}^{|U|}\sum_{j}^{|V|}\frac{|U_i\cap V_j|}{N}\log\frac{N|U_i\cap V_j|}{|U_i||V_j|} \qquad (۱۹)$$

در این فرمول $|U_i|$ تعداد نمونه‌ها را در خوشه U_i نشان می‌دهد و $|V_j|$ تعداد نمونه‌ها در خوشه V_j است. این معیار مستقل از برچسب نمونه‌ها می‌باشد و یک معیار برای ارزیابی دو خوشه‌بندی است. دو نسخه برای نرمال‌کردن MI وجود دارد: اطلاعات متقابل تنظیم‌شده[5] (AMI) و اطلاعات متقابل نرمال‌شده[6] (NMI) [۳۳]. اکثراً از NMI در مقالات استفاده می‌شود و معیار NMI، نرمال‌شده MI در مقیاس صفر (بدون اطلاعات متقابل) و یک (وابستگی کامل) است و به صورت رابطه زیر تعریف می‌شود

$$NMI(U,V)=\frac{MI(U,V)}{mean(E(U),E(V))} \qquad (۲۰)$$

که $E(U)=-\sum_i^{|U|}P(i)\log P(i)$ و $E(V)=-\sum_j^{|V|}P(j)\log P(j)$ آنتروپی یا میزان عدم قطعیت خوشه‌بندی U و V را نشان می‌دهد و $P(i)=|U_i|/N$ احتمال این را که یک متن، متعلق به خوشه U_i باشد بیان می‌کند. در این معیار ترتیب برچسب‌گذاری خوشه‌ها، امتیاز NMI را تغییر نمی‌دهد. امتیاز نزدیک به یک نشان می‌دهد که دو خوشه‌بندی بسیار شبیه به هم هستند و امتیاز نزدیک به صفر نشان می‌دهد که دو خوشه‌بندی از همدیگر مستقل می‌باشند. معیارهای که بر پایه MI هستند برای ارزیابی نیاز دارند که برچسب‌گذاری دستی خوشه‌ها موجود باشد که این در عمل همیشه امکان‌پذیر نیست. اگر برچسب‌های دستی در دسترس نباشند، می‌توان ارزیابی را توسط خود مدل انجام داد. ضریب سیلوته[7] (SC)، معیاری است که در آن امتیاز زیاد نشان می‌دهد که خوشه‌بندی به صورت مناسب توزیع شده است. این ضریب برای هر نمونه از داده به صورت زیر محاسبه می‌شود

$$SC(d)=\frac{b-a}{\max(a,b)} \qquad (۲۱)$$

در این فرمول a میانگین فاصله نمونه داده d با نمونه‌های دیگر در خوشه خودش است و b میانگین فاصله d با نمونه‌های نزدیک‌ترین خوشه را نشان می‌دهد. این معیار برای یک مجموعه از داده‌ها به صورت میانگین ضریب سیلوته همه داده‌ها بیان می‌شود. بازه این ضریب بین ۱- و ۱+ است؛ ۱+ یک خوشه‌بندی با چگالی زیاد را نشان می‌دهد و ۱- برای خوشه‌بندی نامناسب است. مقادیر نزدیک صفر نیز خوشه‌بندی با همپوشانی را نشان می‌دهند.

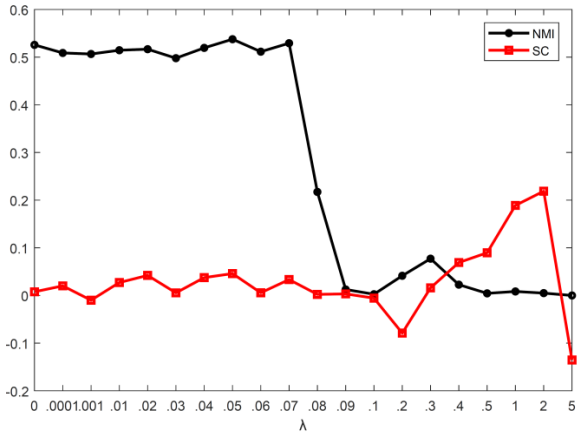
٤-٣ نتایج

در این پژوهش از مجموعه داده‌های رویترز- ۲۱۵۷۸ و WebKB برای ارزیابی روش پیشنهادی در مقایسه با روش‌های دیگر استفاده شده است. تمامی پیاده‌سازی‌ها با زبان پایتون بر روی کامپیوتری هشت‌هسته‌ای با حافظه ۸ گیگابایت و سیستم عامل ویندوز ۷ انجام شده است. قبل از انجام خوشه‌بندی متن نیاز است که بر روی مجموعه داده، پیش‌پردازش انجام شود. پیش‌پردازش متن یکی از مراحل مهم در خوشه‌بندی متن است. در این مقاله از روش‌هایی مانند توکن‌سازی[8]، حذف کلمات متناوب و ریشه‌یابی کلمات استفاده شده است. توکن‌سازی روشی است که در آن محتوای یک متن به عبارات و کلمات تقسیم می‌شود. خیلی از کلمات در

1. Spectral Clustering
2. Dendrogram
3. Linkage
4. Mutual Information
5. Adjusted Mutual Information
6. Normalized Mutual Information
7. Silhouette Coefficient
8. Tokenization

جدول ۱: توزیع متون در مجموعه داده رویترز– ۲۱۵۷۸.

نام خوشه	تعداد متون
Earn	۳۹۶۴
Acq	۲۳۶۹
money-fx	۷۱۷
Grain	۵۸۲
crude	۵۷۸
Trade	۴۸۵
interest	۴۷۸
Ship	۲۸۶
wheat	۲۸۳
Corn	۲۳۷



شکل ۲: مقدار *NMI* و سیلوته روش پیشنهادی در مقایسه با تغییرات λ بر روی مجموعه داده رویترز– ۲۱۵۷۸.

جدول ۲: کارایی روش‌های مقایسه‌شده در مجموعه داده رویترز– ۲۱۵۷۸.

	NMI					سیلوته				
تعداد ویژگی‌ها انتخابی	۵۰۰	۱۰۰۰	۲۰۰۰	همه ویژگی‌ها	میانگین	۵۰۰	۱۰۰۰	۲۰۰۰	همه ویژگی‌ها	میانگین
K-means	۰٫۴۳۵	۰٫۴۶۶	۰٫۴۶۱	۰٫۴۹۱	۰٫۴۶۳	۰٫۰۸۸	۰٫۰۶۰	۰٫۰۴۳	۰٫۰۲۵	۰٫۰۵۴
طیفی	۰٫۴۱۴	۰٫۴۱۶	۰٫۴۱۹	۰٫۴۱۴	۰٫۴۱۶	۰٫۰۸۸	۰٫۰۶۳	۰٫۰۴۳	۰٫۰۱۹	۰٫۰۵۳
Agglomerative	۰٫۴۱۰	۰٫۴۸۲	۰٫۴۷۷	۰٫۴۶۷	۰٫۴۵۹	۰٫۰۹۱	۰٫۰۴۳	۰٫۰۳۷	۰٫۰۱۸	۰٫۰۴۷
NMF	۰٫۴۳۴	۰٫۴۶۴	۰٫۴۴۱	۰٫۴۷۲	۰٫۴۵۳	۰٫۰۴۷	۰٫۰۳۷	۰٫۰۱۴	۰٫۰۱۹	۰٫۰۲۹
FNMF	۰٫۴۸۴	۰٫۴۸۷	۰٫۴۹۴	۰٫۵۱۲	۰٫۴۹۴	۰٫۰۹۸	۰٫۰۶۵	۰٫۰۵۵	۰٫۰۲۹	۰٫۰۶۲

بیشتر متن‌ها تکرار می‌شوند و اهمیت زیادی ندارند، زیرا اکثراً برای اتصال کلمات به یکدیگر به کار می‌روند. آنها معمولاً به عنوان کلمات متناوب در نظر گرفته می‌شوند و حذف می‌گردند چون به محتوای متن ربطی ندارند. ریشه‌یابی کلمات روشی است که در آن فرم‌های مختلف یک کلمه به صورت یک فرم کلی به نام ریشه کلمه نشان داده می‌شود [۱۷].

بعد از پیش‌پردازش متون با استفاده از روش $tf-idf$ [1]، میزان اهمیت یک عبارت نسبت به یک متن در مجموعه‌ای از متون نشان داده می‌شود. در واقع هدف این سیستمِ وزن‌دهی، نشان‌دادن اهمیت عبارت در متن است. مقدار $tf-idf$ متناسب با تعداد تکرار عبارت در متن افزایش می‌یابد و توسط تعداد متونی که شامل عبارت می‌باشند متعادل می‌شود. به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود، احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی متن ندارد

$$tf-idf(t,d)=tf(t,d)\times idf(t)$$
$$idf(t)=\log(\frac{N+1}{df(t)+1})+1 \qquad (۲۲)$$

که $df(t)$ تعداد متونی است که در آن عبارت t تکرار شده و $tf(t,d)$ تعداد تکرار عبارت t در متن d است. انتخاب ویژگی، یک روش کاهش تعداد ویژگی‌ها است که در آن، تعداد کلمات یا عبارت‌ها در ماتریس متن–عبارت کاهش پیدا می‌کند. این روش با کاهش تعداد ویژگی‌ها، هم باعث کاهش پیچیدگی زمانی روش‌های خوشه‌بندی و هم باعث افزایش دقت خوشه‌بندی می‌گردد. در این پژوهش از بهره اطلاعات (IG) [2] برای ارزیابی ویژگی‌ها استفاده شده که کاربرد وسیعی در یادگیری ماشین دارد [۳۴]. بهره اطلاعات به صورت میزان اطلاعات فراهم‌شده توسط کلمات برای خوشه‌بندی متن تعریف می‌شود. فرمول این روش برای اندازه‌گیری میزان اهمیت کلمات در خوشه‌بندی به صورت زیر است

$$IG(U,h)=E(U)-E(U|h) \qquad (۲۳)$$

در این فرمول $E(U|h)$ آنتروپی شرطی را نشان می‌دهد.

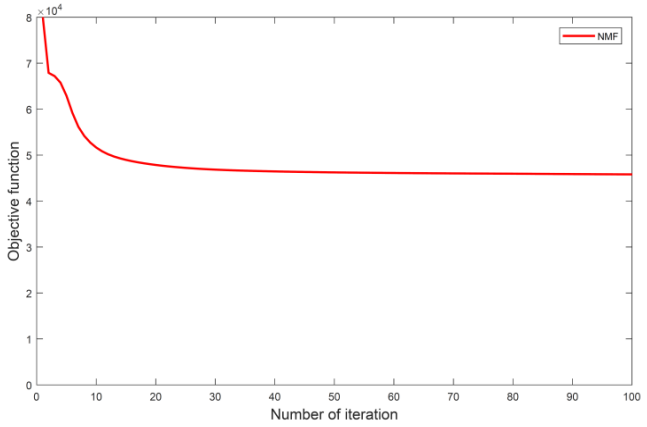
مجموعه داده رویترز– ۲۱۵۷۸ از سال ۱۹۸۷ توسط رویترز در دسترس است که شامل ۲۱۵۷۸ متن و ۱۳۵ خوشه می‌باشد که به صورت دستی تعیین شده‌اند. هر متن در این مجموعه بر اساس محتوایش به یک یا چند خوشه به صورت دستی منتسب شده و متونی که چند برچسب دارند در این پژوهش به هر کدام از خوشه‌ها منتسب شده‌اند. اندازه خوشه‌های یعنی تعداد متون در آنها در بازه ۱۰ تا ۴۰۰۰ هستند. در این مقاله فقط از ۱۰ خوشه بزرگ استفاده گردیده که کلاً ۹۹۷۹ متن دارند. توزیع اندازه خوشه‌ها بالانس نیست؛ بزرگ‌ترین خوشه ۳۹۶۴ متن دارد که ٪۳۹٫۷۲ مجموعه داده و کوچک‌ترین خوشه ۲۳۷ متن دارد که ٪۲٫۳۷ مجموعه داده می‌شود. جدول ۱ تعداد متون داخل خوشه‌ها را نشان می‌دهد.

در روش پیشنهادی باید پارامتر λ و تعداد ویژگی‌های معنایی در ابتدا تعیین شوند. شکل ۲ نشان می‌دهد که *NMI* و سیلوته روش پیشنهادی چگونه با استفاده از پارامتر λ بر روی مجموعه داده رویترز تغییر می‌کند. این شکل نشان می‌دهد که کارایی روش پیشنهادی با ۵۰۰ ویژگی و ۲۰۰ تکرار، وقتی λ از صفر تا پنج تغییر می‌کند، چه تغییراتی دارد.

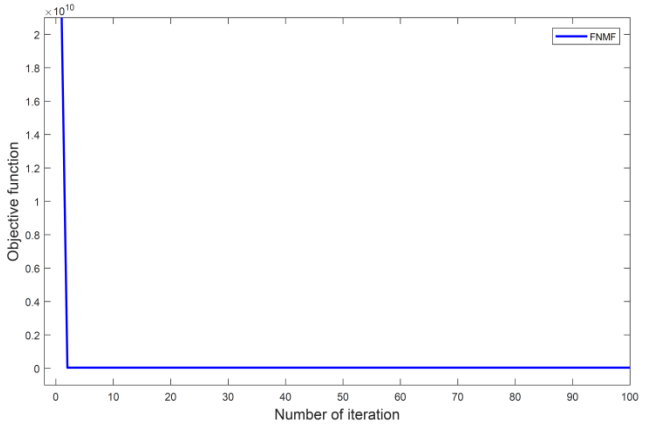
همان طور که در شکل ۲ دیده می‌شود، روش پیشنهادی وقتی λ برابر با ۰٫۰۵ باشد، بیشترین *NMI* را دارد و به همین ترتیب وقتی λ برابر با ۲ است، بیشترین سیلوته را دارد. در این آزمایش‌ها تعداد ویژگی‌های معنایی برابر با تعداد خوشه‌ها در نظر گرفته شده است. در روش پیشنهادی با در نظر گرفتن معیار *NMI* به عنوان معیار اصلی، مقدار λ برابر ۰٫۰۵ در نظر گرفته شده است.

جدول ۲ جزئیات نتایج خوشه‌بندی روش پیشنهادی و روش‌های

1. Term Frequency - Inverse Document Frequency
2. Information Gain



شکل ۳: میزان همگرایی *NMF* بر روی مجموعه داده رویترز- ۲۱۵۷۸.



شکل ۴: میزان همگرایی *FNMF* بر روی مجموعه داده رویترز- ۲۱۵۷۸.

جدول ۳: توزیع متون در مجموعه داده WEBKB.

نام خوشه	تعداد متون
Student	۱۶۴۱
Faculty	۱۱۲۴
Course	۹۳۰
Project	۵۰۴

مقایسه‌شده را بر روی مجموعه داده رویترز نشان می‌دهد. در این جدول با استفاده از روش *IG* ابتدا ۵۰۰ ویژگی برای ارزیابی در نظر گرفته شده و سپس از ۱۰۰۰، ۲۰۰۰ و همچنین تمام ویژگی‌ها یعنی بدون انتخاب ویژگی و با ۲۱۶۸۷ ویژگی برای ارزیابی روش پیشنهادی استفاده گردیده است. به ازای هر مقدار ثبت‌شده، ۳ اجرا انجام شده و مقدار نهایی بر اساس میانگین ۳ اجرا تعیین گردیده است. روش پیشنهادی کارایی خوبی در هر تعداد ویژگی انتخاب‌شده نسبت به روش‌های دیگر دارد و به صورت میانگین *NMI* آن برابر با ۰/۵۱۲ و سیلوته‌اش برابر با ۰/۰۶۲ است. نتایج نشان می‌دهند که محدودیت شباهت اضافه‌شده در روش پیشنهادی باعث افزایش کارایی خوشه‌بندی گردیده است. روش اولیه *NMF* بعد از ۱۰۰ تکرار، توانایی افزایش *NMI* را نداشت و در حالت کلی بدتر از روش K-means عمل کرد. *NMF* به خاطر در نظر نگرفتن محدودیت شباهت، نمی‌تواند شباهت بین بردارها را بعد از نگاشت داده‌ها به فضای جدید حفظ کند.

دلیل کم‌بودن مقدار سیلوته برای روش پیشنهادی و روش‌های دیگر، توزیع برچسب خوشه‌ها در مجموعه داده رویترز بود. زیرا وقتی تنها با روش $tf-idf$ مجموعه داده به بردارهای ویژگی تبدیل می‌شود و بدون انتخاب ویژگی و با برچسب‌های خود مجموعه داده رویترز سیلوته محاسبه می‌گردد، مقدار سیلوته برابر ۰/۰۲۷ می‌شود. این نشان می‌دهد که افزایش *NMI* در راستای افزایش سیلوته نمی‌باشد، یعنی برچسب‌گذاری دستی خوشه‌ها در مجموعه داده با معیار فاصله متناسب نیست و به همین دلیل با افزایش *NMI* مقدار سیلوته کاهش می‌یابد. با این حال همان طور که در شکل ۲ دیده می‌شود، می‌توان برای به دست آوردن مقدار سیلوته بالا در روش پیشنهادی، مقدار λ را تا عدد ۲ بالا برد و به مقدار سیلوته ۰/۲۲۱ رسید. در این پیاده‌سازی‌ها هدف، افزایش معیار *NMI* بود.

در این مقاله برای بررسی بهتر روش پیشنهادی، علاوه بر مجموعه داده رویترز- ۲۱۵۷۸ از مجموعه داده WebKB نیز استفاده شده است. صفحات وب در این مجموعه داده متعلق به گروه‌های علمی دانشگاه‌های مختلف می‌باشد که توسط دانشگاه کارنگی ملون جمع‌آوری شده است. صفحات وب این مجموعه داده به صورت دستی به هفت خوشه تقسیم شده است. در این مقاله، سه خوشه از هفت خوشه به خاطر وجود تعداد کم داده‌ها در آنها حذف گردیده است. جدول ۳ تعداد داده‌ها در هر خوشه را نشان می‌دهد.

جدول ۴ نتایج خوشه‌بندی بر روی مجموعه داده WebKB را نشان می‌دهد. همانند نتایج رویترز- ۲۱۵۷۸، در این جدول نیز با استفاده از روش *IG* ابتدا ۵۰۰ ویژگی برای ارزیابی در نظر گرفته شده و سپس از ۱۰۰۰، ۲۰۰۰ و تمام ویژگی‌ها با ۷۶۴۷ ویژگی برای ارزیابی روش پیشنهادی استفاده گردیده است. همان طور که در جدول دیده می‌شود روش پیشنهادی نتایج بهتری نسبت به روش‌های دیگر دارد.

همان طور که قبلاً ذکر شد، قوانین به روز رسانی برای پیداکردن بهینه محلی به کار برده می‌شوند و در بخش ۳، همگرایی این قوانین اثبات شد و پیچیدگی زمانی روش پیشنهادی مورد تحلیل قرار گرفت. در این بخش میزان همگرایی روش پیشنهادی در مقایسه با *NMF* در پیاده‌سازی‌ها نشان داده می‌شود. شکل‌های ۳ و ۴ سرعت همگرایی *NMF* و روش پیشنهادی را بر روی مجموعه داده رویترز نشان می‌دهند. همان طور که در شکل ۴ دیده می‌شود، روش پیشنهادی خیلی سریع همگرا می‌شود و در کمتر از ۱۰ تکرار کارایی خوبی را نشان می‌دهد، ولی روش اولیه *NMF* بعد از ۱۰۰ تکرار همگرا می‌شود (شکل ۳). این آزمایش مطالب گفته‌شده در بخش ۳-۳ را تأیید می‌کند که اگرچه روش پیشنهادی محاسبات زیادی نسبت به روش اولیه *NMF* دارد ولی در تعداد تکرار کمتری همگرا می‌شود و در کل فرایند به روز رسانی، روش پیشنهادی خیلی سریع‌تر عمل می‌کند.

۵- نتیجه‌گیری

در این پژوهش یک روش تجزیه ماتریس غیر منفی با به کارگیری شباهت بین بردارهای ویژگی متون برای کاهش ابعاد و خوشه‌بندی ارائه گردیده است. در این روش محدودیت‌های شباهت به صورت متغیرهایی برای افزایش سرعت همگرایی به تابع هزینه اضافه شده‌اند. اثبات همگرایی، تحلیل پیچیدگی زمانی و نتایج به دست آمده از آزمایش‌های انجام‌شده بر روی مجموعه داده رویترز نشان‌دهنده کارایی روش پیشنهادی در مقایسه با روش‌های دیگر است. در این مقاله نشان داده شد که با استفاده از خصوصیات روش تجزیه ماتریس غیر منفی و همچنین محدودیت شباهت بین بردارهای ویژگی، روش پیشنهادی خیلی بهتر می‌تواند داده‌ها را در ابعاد کمتر مدل کند.

با تنظیم صحیح مقدار λ، روش پیشنهادی می‌تواند نتایج بهتری نسبت به روش‌های دیگر به دست آورد. همچنین روش پیشنهادی را

جدول ۴: کارایی روش‌های مقایسه‌شده در مجموعه داده WEBKB.

	NMI								
تعداد ویژگی‌ها انتخابی	۵۰۰	۱۰۰۰	۲۰۰۰	همه ویژگی‌ها	میانگین	سیلوته ۵۰۰	۱۰۰۰	۲۰۰۰	همه ویژگی‌ها	میانگین
K-means	۰٫۳۴۱	۰٫۳۲۶	۰٫۳۲۵	۰٫۳۵۸	۰٫۳۳۸	۰٫۰۱۳	۰٫۰۱۳	۰٫۰۰۵	۰٫۰۱۲	۰٫۰۱۱
طیفی	۰٫۳۱۵	۰٫۳۱۰	۰٫۲۹۵	۰٫۲۹۸	۰٫۳۰۵	۰٫۰۱۷	۰٫۰۱۵	۰٫۰۴۵	۰٫۰۱۴	۰٫۰۲۳
Agglomerative	۰٫۲۵۶	۰٫۲۴۱	۰٫۲۶۲	۰٫۰۰۱	۰٫۱۹۰	۰٫۰۵۷	۰٫۰۶۵	۰٫۰۴۰	۰٫۰۱۵	۰٫۰۴۴
NMF	۰٫۳۰۴	۰٫۳۲۷	۰٫۳۳۰	۰٫۳۳۳	۰٫۳۲۴	−۰٫۰۰۲	۰٫۰۰۱	۰٫۰۰۷	۰٫۰۱۰	۰٫۰۰۴
FNMF	۰٫۳۵۶	۰٫۳۶۲	۰٫۳۷۷	۰٫۳۶۰	۰٫۳۶۴	۰٫۰۵۳	۰٫۰۵۰	۰٫۰۴۱	۰٫۰۲۴	۰٫۰۴۲

می‌توان در مسایل دیگری به کار برد و به عنوان کارهای آتی در نظر داریم که این روش را بر روی خلاصه‌سازی متن اعمال کنیم و نیز انتخاب مقدار بهینه برای λ یک چالش در روش پیشنهادی است که در نظر داریم در تحقیقات آتی، انتخاب خودکار مقدار λ را بررسی نماییم.

مراجع

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, Hoboken, NJ, 2nd Edition, 2000.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, Jan. 1999.

[3] D. Kuang, J. Choo, and H. Park, "Nonnegative matrix factorization for interactive topic modeling and document clustering," In M. Emre Celebi (Ed.), *Partitional Clustering Algorithms*, pp. 215-243, Springer Cham, 2015.

[4] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR'01*, pp. 207-212, Kauai, HI, USA, 8-14. Dec. 2001.

[5] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in *Proc. of the 26th Annual Int ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 267-273, Toronto, Canada, 28 Jul.-1 Aug. 2003.

[6] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of the 2001 Conf. Advances in Neural Information Processing Systems*, pp. 556-562, Vancouver, Canada, 3-8 Dec. 2001.

[7] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric for image analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1590-1602, Aug. 2011.

[8] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear Algebra and Its Applications*, vol. 416, no. 1, pp. 29-47, Jul. 2006.

[9] M. H. Aghdam and M. D. Zanjani, "A novel regularized asymmetric non-negative matrix factorization for text clustering," *Information Processing & Management*, vol. 58, no. 6, Article ID: 102694, 6 pp., Nov. 2021.

[10] A. J. Mohammed, Y. Yusof, and H. Husni, "Document clustering based on firefly algorithm," *J. of Computer Science*, vol. 11, no. 3, pp. 453-465, Mar. 2015.

[11] A. J. Mohammed, Y. Yusof, and H. Husni, "Determining number of clusters using firefly algorithm with cluster merging for text clustering," in *Proc. Int. Visual Informatics Conf.*, pp. 14-24, Bangi, Malaysia, 17- 19 Nov. 2015.

[12] A. Abraham, S. Das, and A. Konar, "Document clustering using differential evolution," in *Proc. IEEE Int. Conf. on Evolutionary Computation*, pp. 1784-1791, Vancouver, Canada, 16-21 Jul. 2006.

[13] A. Kumar and H. Daum´e, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. on Machine Learning, ICML'11*, pp. 393-400, Bellevue, WA, USA, 28 Jun.-2 Jul. 2011.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proc. of the 2002 Conf. Advances in Neural Information Processing Systems*, pp. 849-856, Vancouver, Canada, 9-14 Dec. 2002.

[15] D. Yogatama and K. Tanaka-Ishii, "Multilingual spectral clustering using document similarity propagation," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, vol. 2, pp. 871-879, Singapore, 6-7 Aug. 2009.

[16] D. Toli´c, N. Antulov-Fantulin, and I. Kopriva, "A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering," *Pattern Recognition*, vol. 82, pp. 40-55, Oct. 2018.

[17] C. C. Aggarwal and C. Zhai, *Mining Text Data*, Springer Science & Business Media, 2012.

[18] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373-386, Mar. 2006.

[19] E. F. Gonzalez and Y. Zhang, *Accelerating the Lee-Seung Algorithm for Nonnegative Matrix Factorization*, Tech. Rep., 2005.

[20] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698-714, Nov. 2013.

[21] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition," *IEEE Trans. on Signal Processing*, vol. 62, no. 1, pp. 211-224, Oct. 2013.

[22] K. Kimura, M. Kudo, and Y. Tanaka, "A column-wise update algorithm for nonnegative matrix factorization in bregman divergence with an orthogonal constraint," *Machine Learning*, vol. 103, no. 2, pp. 285-306, May 2016.

[23] F. Shang, L. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognition*, vol. 45, no. 6, pp. 2237-2250, Jun. 2012.

[24] X. Ma, P. Sun, and G. Qin, "Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability," *Pattern Recognition*, vol. 71, pp. 361-374, Nov. 2017.

[25] A. David and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. of the 18th annual ACM-SIAM Symp. on Discrete Algorithms, SODA'07*, pp. 1027-1035, New Orleans, LA, USA, 7-9 Jan. 2007.

[26] J. N. Myhre, K. Ø. Mikalsen, S. Løkse, and R. Jenssen, "Robust clustering using a knn mode seeking ensemble," *Pattern Recognition*, vol. 76, pp. 491-505, Apr. 2018.

[27] W. Jiang, W. Liu, and F. L. Chung, "Knowledge transfer for spectral clustering," *Pattern Recognition*, vol. 81, pp. 484-496, Sept. 2018.

[28] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, Aug. 2007.

[29] J. V. Munoz, M. A. Goncalves, Z. Dias, and R. da S. Torres, "Hierarchical clustering-based graphs for large scale approximate nearest neighbor search," *Pattern Recognition*, vol. 96, Article ID: 106970, Dec. 2019.

[30] L. Rokach and O. Maimon, *Clustering Methods*, in Data Mining and Knowledge Discovery Handbook. Springer, pp. 321-352, 2005.

[31] H. Xiong and D. Kong, "Elastic nonnegative matrix factorization," *Pattern Recognition*, vol. 90, pp. 464-475, Jun. 2019.

[32] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance," *J. of Machine Learning Research*, vol. 11, pp. 2837-2854, Dec. 2010.

[33] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Scientific Reports*, vol. 6, no. 1, pp. 1-18, Aug. 2016.

[34] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30-36, Nov. 1999.

مهدی حسین‌زاده اقدم در سال ۱۳۸۵ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه آزاد اسلامی واحد قزوین و در سال ۱۳۸۷ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه اصفهان دریافت نمود. دکتر حسین‌زاده اقدم دوره دکترای مهندسی کامپیوتر را در دانشگاه علم و صنعت ایران در سال ۱۳۹۵ به پایان رساند و از سال ۱۳۹۶ در دانشکده فنی و مهندسی دانشگاه بناب مشغول به فعالیت گردید و اینک

نیز عضو هیأت علمی این دانشکده می‌باشد. زمینه‌های علمی مورد علاقه نامبرده متنوع بوده و شامل موضوعاتی مانند یادگیری ماشین، هوش محاسباتی، سیستم‌های توصیه‌گر و پردازش متن می‌باشد.

مرتضی آنالویی تحصیلات خود را در مقطع کارشناسی مهندسی برق در سال ۱۳۶۳ از دانشگاه علم و صنعت ایران و در مقطع دکتری مهندسی برق و الکترونیک در سال ۱۳۶۹ از دانشگاه اکایاما ژاپن به پایان رسانده است و هم‌اکنون عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران می‌باشد. نام‌برده قبل از پیوستنش به دانشگاه علم و صنعت ایران در سالهای ۱۳۷۰ الی ۱۳۷۳ عضو هیأت علمی دانشکده مهندسی دانشگاه اکایاما ژاپن و در سال ۱۳۷۴ عضو هیأت علمی دانشکده مهندسی دانشگاه تربیت مدرس بوده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: هوش مصنوعی، بهینه‌سازی و شبکه‌های کامپیوتری.

جعفر تنها تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد علوم کامپیوتر به‌ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۱ از دانشگاه صنعتی امیرکبیر به پایان رسانده است و پس از آن به دوره دکترای مهندسی کامپیوتر در دانشگاه آمستردام در هلند وارد گردید و در سال ۱۳۹۲ موفق به اخذ درجه دکترا در علوم کامپیوتر از دانشگاه مذکور گردید. دکتر تنها اینک عضو هیأت علمی دانشکده مهندسی برق و کامپیوتر دانشگاه تبریز می‌باشد. زمینه‌های علمی مورد علاقه نام‌برده متنوع بوده و شامل موضوعاتی مانند یادگیری ماشین، شناسایی الگو و تحلیل متن می‌باشد.