Journal of Modeling in Engineering Volume Nomber (year) first page-last page

Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

Research Article

An Improved Deep Text Clustering via Local Manifold of an Autoencoder Embedding

Fatemeh Daneshfar [1*], Amin Golzari Oskouei [2], Maryam Dorosti [3] and Mohammad Javad Aghajani [4]

[1] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[2] Department of Computer Engineering, University of Tabriz, Tabriz, Iran
[3] Department of Electrical and Computer Engineering, Kharazmi University, Tehran, Iran

*Corresponding Author:

PAPER INFO

Paper history:

Received:
Revised:
Accepted:

Keywords:

Text clustering; Deep clustering; Deep learning; Manifold learning; Autoencoder.

ABSTRACT

Text clustering is a method for separating specific information from textual data and can even classify text according to topic and sentiment, which has drawn much interest in recent years. Deep clustering methods are especially important among clustering techniques because of their high accuracy. These methods include two main components: dimensionality reduction and clustering. Many earlier efforts have employed autoencoder for dimension reduction; however, they are unable to lower dimensions based on manifold structures, and samples that are like one another are not necessarily placed next to one another in the low dimensional. In the paper, we develop a Deep Text Clustering method based on a local Manifold in the Autoencoder layer (DCTMA) that employs multiple similarity matrices to obtain manifold information, such that this final similarity matrix is obtained from the average of these matrices. The obtained matrix is added to the bottleneck representation layer in the autoencoder. The DCTMA's main goal is to generate similar representations for samples belonging to the same cluster; after dimensionality reduction is achieved with high accuracy, clusters are detected using an end-to-end deep clustering. Experimental results demonstrate that the suggested method performs surprisingly well in comparison to current state-of-the-art methods in text datasets.

DOI: https://doi.org/

How to cite this article:

خوشه‌بندی متن عمیقِ بهبودیافته با استفاده از منیفولد محلی تعبیه‌شده خودرمزگذار

فاطمه دانشفر[۱*]، امین گلزاری اسکویی[۲]، مریم درستی[۳]، محمدجواد آقاجانی[۴]

اطلاعات مقاله

نوع مقاله:
دریافت مقاله:
بازنگری مقاله:
پذیرش مقاله:

واژگان کلیدی:
خوشه‌بندی متن، خوشه‌بندی عمیق، یادگیری عمیق، یادگیری منیفولد، خودرمزگذار

چکیده

خوشه‌بندی متن، روشی برای جداسازی اطلاعات از دادگان متنی است که می‌تواند متن را براساس موضوع و احساس طبقه‌بندی کند که اخیرا مورد توجه بسیاری قرار گرفته است. روش‌های مبتنی برخوشه‌بندی عمیق به دلیل دقت بالا، در میان تکنیک‌های خوشه‌بندی از اهمیت ویژه‌ای برخوردار هستند. این روش‌ها شامل دو جزء اصلی کاهش ابعاد و خوشه‌بندی می‌باشند. بسیاری از روش‌های پیشین عمیق، از خودرمزگذار برای کاهش ابعاد استفاده می‌کردند. این روشها قادر به کاهش ابعاد براساس ساختارهای منیفولد نیستند و در آنها نمونه‌هایی که شبیه یکدیگر هستند لزوماً در ابعاد پایین نیز در کنار یکدیگر قرار نمی‌گیرند. در این مقاله، ما یک روش خوشه‌بندی متن عمیق را براساس یک منیفولد محلی در لایه خودرمزگذار (DCTMA) توسعه می‌دهیم که از ماتریس‌های شباهت متعدد برای درنظرگرفتن جهت، اندازه و معنا استفاده می‌کند، به‌طوری که ماتریس شباهت نهایی از میانگین این ماتریس‌ها به‌دست می‌آید. ماتریس به‌دست آمده به لایه بازنمایی پنهان در خودرمزگذار اضافه می‌شود. هدف اصلی DCTMA تولید بازنمایی‌های مشابه برای نمونه‌های متعلق به یک خوشه است. پس از کاهش ابعاد با دقت بالا، خوشه‌ها با استفاده از خوشه‌بندی عمیق انتها به انتها شناسایی می‌شوند. نتایج تجربی نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های پیشرفته فعلی روی مجموعه دادگان متنی، عملکرد خوبی دارد.

۱-مقدمه[۱]

خوشه‌بندی متن فرآیندی چالش‌برانگیز برای کشف و استخراج گروه‌هایی با عناصر مشابه در مجموعه‌های متنی است [۴]، که در آن اسناد مختلف معمولاً براساس شباهت محتوا دسته‌بندی می‌شوند. در حال حاضر الگوریتم‌های خوشه‌بندی مختلفی معرفی شده‌اند که به چند دسته تقسیم می‌شوند. گروهی از این الگوریتم‌ها مبتنی بر بخش‌بندی هستند، مانند K-means و K-medoids که در آنها دادگان با توجه به کمترین فاصله‌شان به مراکز خوشه‌بندی مختلف تخصیص داده می‌شوند [7]. در این روش‌ها هدف بهینه-سازی فاصله بین نمونه‌ها و مراکز خوشه‌ها می‌باشد. نوع دیگری از روش‌های خوشه‌بندی بر اساس توزیع دادگان است. این الگوریتم‌ها از مجموعه‌ای از توابع توزیع احتمال برای بازنمایی داده‌ها استفاده می‌کنند و فرض بر این است که تمامی نقاطِ هر خوشه به احتمال زیاد از توزیع یکسانی مشتق شده‌اند [7]. این روشها بدون داشتن تابع توزیع مناسب نمی‌توانند به دقت بالایی در خوشه‌بندی دست یابد.

از سوی دیگر، الگوریتم‌های خوشه‌بندی مبتنی بر چگالی، دسته‌ای از خوشه‌بندی‌ها هستند که می‌توانند دادگان با هر توزیع دلخواهی را، خوشه‌بندی کنند [10]. بنابراین می‌توانند به‌طور مشخص خوشه‌ها را با هر توزیع و شکل داده‌ای شناسایی کنند، زیرا که هر خوشه یک منطقه پیوسته از نقاط متراکم است که توسط مناطق به هم پیوسته از دیگران جدا شده است [7]. یکی دیگر از روش‌های رایج، خوشه‌بندی مبتنی بر پیکره است که ابعاد ویژگی‌ها را با ادغام مترادف‌ها کاهش می‌دهد و در نتیجه از مجموعه دادگان بزرگی مانند WordNet با هدف کاهش ابعاد بردار و حذف اطلاعات اضافی استفاده می‌کند [7]. فاکتورسازی ماتریس غیرمنفی[۱] (NMF) و نمایه‌سازی معنایی پنهان (LSI) دو روش معروف خوشه‌بندی مبتنی بر پیکره هستند که در آنها اسناد به فضای ویژگی جدید با ابعاد کوچکتر تبدیل می‌شوند و ویژگی‌ها معمولاً ترکیبی خطی از ویژگی‌های اصلی هستند.

در حوزه خوشه‌بندی، امروزه NMF مورد توجه بسیاری قرار گرفته است و کاربردهای زیادی در تشخیص الگو و متن‌کاوی پیدا کرده است. NMF مدل خطی-جبری برای کاهش ابعاد بردارهای غیرمنفی است. این روش می‌تواند ویژگی‌ها را استخراج کرده و برای توصیف ماتریس اصلی به شکل حاصل‌ضرب یک ماتریس وزنی و یک ماتریس پایه استفاده می‌شود. مهم‌ترین هدفNMF یافتن و استخراج ویژگی‌های کلی با ترکیبی غیرمنفی از ویژگی‌های محلی است. بنابراین، این روش تاکنون به طور گسترده‌ای در پردازش زبان طبیعی، پردازش تصویر و صدا استفاده شده است [12]. با این‌حال ازآنجایی‌که NMF یک راه حل ساده و خطی است، نمی‌تواند روابط غیرخطی پنهان را در دادگان کشف کند و این محدودیت توانایی طبقه‌بندی دادگان با ساختار پیچیده را کاهش می‌دهد [13]. بنابراین روش‌های تجزیه ماتریسی بطور معمول از دو مشکل عمده رنج می‌برند. اولاً آنها نمی‌توانند الگوهای نهفته در ساختار دادگان را به صورت سلسله مراتبی کشف کنند و دوم اینکه نتایج خوشه‌بندی آنها با خوشه‌بندی انسانی فاصله دارد.

امروزه یادگیری عمیق در کشف و استخراج ساختارهای غیرخطی بسیار مورد توجه قرار گرفته است [9, 11] و بسیاری از محققان با استفاده از آن پیشرفت قابل توجهی در استخراج بازنمایی‌های عمیق برای حوزه‌های مختلف داشته‌اند. تکنیک‌های یادگیری عمیق، می-توانند بازنمایی‌ها را به‌صورت سلسله مراتبی و غیرخطی به خوبی یاد گیرند [14]. با این‌حال، یکی از چالش‌های اصلی خوشه‌بندی عمیق زمانی است که دادگان دارای ابعاد بالا و ساختارهای پیچیده هستند یا خوشه‌های زیادی وجود دارد که باعث می‌شود فرآیند خوشه‌بندی زمان‌بر و ناکارآمد باشد [15]. در چنین مواردی بازنمایی دادگان در فضایی با ابعاد کمتر اجتناب‌ناپذیر است و خوشه‌بندی را بهبود می-بخشد. همچنین کاهش ابعاد می‌تواند شباهت معنایی بین متون را به نحو بهتری پیدا کرده و خوشه‌بندی بهتری ایجاد کند. تاکنون روش‌های زیادی برای کاهش ابعاد به‌ویژه در متن استفاده شده است، مانند تحلیل مؤلفه‌های مستقل (ICA)، تجزیه مقدار منفرد (SVD)، تحلیل مؤلفه‌های اصلی (PCA) و غیره [16]. روش‌های عمیق معمولاً از خودرمزگذارها برای بازنمایی دادگان با ابعاد کم استفاده می‌کنند که یکی از روش‌های کاهش ابعاد غیرخطی هستند [17]. با این‌حال، در بیشتر روش‌های خوشه‌بندی مبتنی بر یادگیری عمیق فعلی، این کاهش ابعاد بدون در نظر گرفتن ساختارهای هندسی دادگان انجام می‌شود و دادگان مشابه لزوماً در فضای جدید در کنار یکدیگر قرار نخواهند داشت.

در این مقاله یک روش خوشه‌بندی متن عمیق با یک منیفولد محلی در لایه خودرمزگذار[۲] (DCTMA) معرفی شده است که از

* f.daneshfar@uok.ac.ir
۱. استادیار، دانشکده مهندسی، گروه کامپیوتر، دانشگاه کردستان
۲. دانشکده مهندسی کامپیوتر، دانشگاه تبریز
۳. دانشکده مهندسی کامپیوتر، دانشگاه خوارزمی
٤. دانشکده مهندسی، گروه کامپیوتر، دانشگاه کردستان

How to cite this article:

ماتریس‌های شباهت چندگانه استفاده می‌کند، به‌طوری که ماتریس شباهت نهایی از میانگین این ماتریس‌ها به‌دست می‌آید. این ماتریس شباهت به‌عنوان اطلاعات اضافی، بهمراه بازنمایی دادگان، برای خوشه‌بندی بهتر استفاده می‌شود و به لایه تعبیه‌شده در خودرمزگذار اضافه می‌شود. ایده اصلی این است که نمونه‌های نزدیک به هم در فضای اصلی باید بازنمایی‌های مشابهی نیز در فضای جاسازی[۳] داشته باشند. با این‌کار، دادگان مشابه بیشتری در کنار یکدیگر قرار می‌گیرند و دقت یادگیری بازنمایی[۴] افزایش می‌یابد. در مدل ارائه شده، همراه با کاهش ابعاد به‌دست آمده با دقت بالا، خوشه‌ها با استفاده از یک چارچوب عمیق انتها به انتها شناسایی می‌شوند. در این مدل، ماتریس شباهت به یک شبکه عمیق از طریق یک خودرمزگذار ارائه می‌شود، درنتیجه از شباهت دادگان و بازنمایی بهترِ اسناد توسط یک راه‌حل انتها به انتها بهره می‌برد.

با توجه به اینکه اکثر معیارهای شباهت کنونی براساس شباهت کسینوسی (CS) و فاصله اقلیدسی (ED) می‌باشند و فاصله اقلیدسی تنها اندازه بردارها و فاصله کسینوسی تنها زاویه میان دو بردار را در نظر می‌گیرد، بنابراین این معیارها، روش‌های موثر و مناسبی برای تجزیه و تحلیل متن نمی‌باشند که به‌طور همزمان جهت، اندازه و معنا را درنظر بگیرد [18]. در این مقاله از ترکیب چندین معیار تشابه متون مختلف استفاده شده است که هریک از دیدگاهی متفاوت شباهت بین دو سند را می‌سنجند. سپس یک ماتریس تشابه جامع[۵] برای هر مجموعه داده محاسبه می‌شود که میانگینی از این سه معیار تشابه متن است. این معیار به خوبی شباهت اسناد را پیدا می‌کند و به خوشه‌بندی عمیق کمک می‌کند. مدل پیشنهادی روی سه مجموعه داده مختلف با معیارهای ارزیابی متفاوت مورد بررسی قرار گرفته است. با توجه به نتایج تجربی، روش ارائه شده نتایج بهتری را بر روی همان مجموعه داده نسبت به روش‌های اخیرا منتشر شده به‌دست آورده است.

نوآوری‌های این مقاله به شرح زیر می‌باشد:

• در این مقاله، یک معماری مبتنی بر خوشه‌بندی عمیق انتها به انتها، برای یادگیری توام بازنمایی و برچسب‌های خوشه‌ها ارائه شده است.

• در این مقاله با در نظر گرفتن ماتریس‌های شباهت متفاوت، خوشه‌بندی فضای منیفولد، با در نظر گرفتن ساختار هندسی دادگان در فضای اولیه انجام گرفته است.

• در این مقاله از یک معیار جدید بعنوان تابع زیان[۶] استفاده شده است که نه تنها جهت، بلکه بزرگی و معنا را نیز برای خوشه‌بندی در نظر می‌گیرد.

در ادامه این مقاله، مروری بر ادبیات در بخش ۲ آمده است. سپس، توضیحات مربوط به مدل ارائه شده، در بخش ۳ داده شده است. بخش ۴ نتایج تجربی را ارائه می‌کند و درنهایت، نتیجه‌گیری در بخش ۵ به تفصیل آمده است.

۲-مروری بر ادبیات

اخیراً تحقیقات و تلاش‌های گسترده‌ای برای یادگیری بازنمایی‌های عمیق برای خوشه‌بندی اسناد صورت گرفته است.

دیالو و همکاران [9] از یک روش خوشه‌بندی عمیق برای یادگیری بازنمایی اسناد با معرفی خودرمزگذارهای انقباضی استفاده کرده‌اند، که مشکل حفظ مکان خوشه را با انتقال نقاط داده مجاور به یکدیگر حل می‌کند. در این روش، نرم فروبنیوس به عنوان جریمه علاوه بر تابع هزینه معمول خودرمزگذار برای درک ویژگی‌های اسناد مرتبط استفاده می‌شود. می و همکاران [18] یک فاکتورسازی مفهومی نیمه‌نظارتی را بهمراه ترکیب آن با محدودیت‌های زوجی برای افزایش عملکرد خوشه‌بندی با اطلاعات نظارت‌شده ارائه کرده‌اند. بااستفاده از این روش، نقاط داده‌ای که متعلق به خوشه یکسانی در فضای اولیه هستند در فضای ثانویه نیز در یک خوشه قرار خواهند گرفت.

در [19] یک چارچوب شبکه عصبی برای خوشه‌بندی نیمه‌نظارتی با محدودیت‌های باینری معرفی شده است که دارای دو فاز ساده است: فاز اول از یک جفت شبکه عصبی سیامی برای اتصال جفت‌های بدون برچسب استفاده می‌کند. فاز دوم از جفت مجموعه دادگان برچسب‌دار مرحله اول در یک روش خوشه‌بندی نظارت‌شده استفاده می‌کند. این روش به این دلیل ارائه می‌شود که طبقه‌بندی باینری معمولاً ساده‌تر از خوشه‌بندی چندکلاسه با نظارت جزئی است. در مقاله دیگر، فو و همکاران [20] خوشه‌بندی متون کوتاه را با استفاده از فاکتورسازی ماتریس مجاورت مستقیم انجام داده‌اند. از آنجایی که متون کوتاه نویزی هستند، یافتن شباهت‌های همسایگان در ماتریس همسایگی دشوار است. این تحقیق از فاکتورسازی ماتریس شباهت برای کاهش این مشکل با ترکیب یک ماتریس برای ثبت مستقیم شباهت‌های همسایگان و افزودن منظم‌ساز به ماتریس تخصیص برای حذف خوشه‌بندی سخت استفاده می‌کند.

در رویکرد مبتنی بر یادگیری عمیق ارائه شده توسط شنگ و لیپور [21]، از آموزش مشترک یک شبکه سیامی و خودرمزگذار از طریق محدودیت‌های زوجی، برای یادگیری بدون نظارتِ یک بازنمایی برای خوشه‌بندی و طبقه‌بندی استفاده می‌شود. این چارچوب می‌تواند از یک روش یادگیری به روش دیگر منتقل شود و به‌طور یکپارچه خوشه‌بندی محدود، طبقه‌بندی نیمه‌نظارتی و طبقه‌بندی نظارتی را ادغام کند. ترکیبی از خوشه‌بندی K-means با خودرمزگذارهای پشته‌ای برای خوشه‌بندی متن عمیق توسط حسینی و ورزنه [11] ارائه شده است. در مدل مبتنی بر نمایش توالی ارائه شده توسط

گوان و همکاران [22]، خودرمزگذارهای از پیش‌آموزش‌دیده [۲۳] برای خوشه‌بندی متن بدون نظارت با استفاده از ویژگی‌های عمیق پیشنهاد شده‌اند. در این روش برخلاف روش‌های معمول ارائه متن، بااستفاده از یک رمزگذار متن عمیق از پیش‌آموزش‌دیده، نمایش معنایی متن تهیه می‌شود که می‌تواند مشکل پراکندگی ویژگی‌ها را حل کند. در چارچوب پیشنهادی مرادی‌فرد و همکاران [24]، دادگان و نمایش‌های خوشه‌ای به‌طور مشترک آموخته می‌شوند. در این یادگیری که از طریق پس‌انتشار انجام می‌شود، از محدودیت‌های زوجی برای یادگیری بهتر نمایش اسناد استفاده می‌شود. در [25] روشی مبتنی بر خوشه‌بندی متن نیمه‌نظارتی ارائه شده است که در آن نمونه‌های برچسب‌دار برای خوشه‌بندی نیمه‌نظارتی متضاد عمیق [۲۶] استفاده می‌شوند، که به طور مشترک خوشه‌بندی و یادگیری نمایش را بهینه می‌کند. در مقاله ارائه شده توسط ویلهاگرا و همکاران [27]، خوشه‌بندی عمیق با شبکه سیامی کانولوشن نیز برای یادگیری نمایش داده‌ها با محدودیت‌های زوجی استفاده شده است، و از الگوریتم K-means برای خوشه‌بندی بدون نظارت استفاده می‌شود.

در اکثر کارهای انجام شده تاکنون، شباهت ساختار هندسی دادگان در فضای اولیه تنها از یک دیدگاه در نظرگرفته شده است. در مدل ارائه شده در این مقاله، با درنظر گرفتن ماتریس‌های شباهت متفاوت و با استفاده از یک معیار جدید بعنوان تابع زیان، که نه تنها جهت، بلکه بزرگی و معنا را نیز درنظر می‌گیرد، خوشه‌بندی فضای منیفولد، با دقت بالاتری انجام خواهد گرفت.

جدول ۱ بصورت خلاصه کارهای انجام شده اخیر را در این حوزه نمایش می‌دهد.

۳- مدل ارائه شده

در این بخش مدل DCTMA برای خوشه‌بندی عمیق دادگان متنی توضیح داده شده است. در ابتدا برخی از مفاهیم اصلی معرفی شده و سپس جزییات مدل به تفصیل شرح داده می‌شود.

۳-۱- خودرمزگذار عمیق

خودرمزگذار می‌تواند نمایش‌های معنی‌داری از دادگان ورودی را به شیوه‌ای بدون نظارت بیاموزد [۲۸]. یک خودرمزگذار ساده (AE) شامل یک شبکه عصبی دو لایه است که از یک رمزگذار و یک رمزگشا تشکیل شده است [29]. لایه رمزگذاری، ورودی‌های شبکه را به فضایی با ابعاد کمتر (فضای تعبیه‌شده) تبدیل می‌کند و لایه رمزگشایی مسئول بازسازی شبکه از فضای تعبیه شده است. در ساختار اصلی خودرمزگذار اگر از الگوریتم پس‌انتشار با وزن‌های تصادفی استفاده شود، می‌توان به سرعت آنرا آموزش داد. مراحل رمزگذاری و رمزگشایی به شرح زیر است:

$$h(x_i) = f(W_e x_i + b_1) \quad (۱)$$

$$\hat{x_i} = f(W_d h(x_i) + b_2) \quad (۲)$$

که در آن f یک تابع فعال‌سازی غیرخطی مانند سیگموئید یا ReLU می‌باشد. W_e وزن رمزگذار، W_d وزن رمزگشا و b پارامتر بایاس است. هدف از آموزش AE بهینه‌سازی پارامترهای شبکه $\varphi = (W_e, b_1, W_d, b_2)$ با کمینه کردن تابع خطای زیر است،

$$l_{rec} = \sum_{i=1}^{n} \|X_i - \hat{X}_i\|^2 = \sum_{i=1}^{n} \|f(W_d f((W_e)x_i + b_1) + b_2) - \hat{X}_i\|^2 \quad (۳)$$

برخی از مجموعه دادگان روابط پیچیده‌تری دارند. درنتیجه، استفاده از تنها یک خودرمزگذار کافی نیست. زیرا اندازه ویژگی‌های ورودی ممکن است برای یک خودرمزگذار خودکار منفرد، بزرگ باشد. بنابراین در این مواقع از چندین خودرمزگذار استفاده می‌شود تا یک ساختار عمیق در یک خودرمزگذار پشته‌ای[۷] (SAE) ایجاد کند. عملکرد مهم خودرمزگذار پشته‌ای درعبور لایه به لایه ویژگی‌های ورودی بصورت آموزش بدون نظارت است. اولین لایه می‌تواند به‌عنوان ورودی برای خودرمزگذار مورد استفاده قرار گیرد. الگوریتم پس‌انتشار نیز می‌تواند برای تنظیم دقیق یک شبکه عصبی که قبلا آموزش دیده است استفاده شود. آخرین لایه ممکن است برای طبقه‌بندی سنتی بانظارت استفاده شود.

۳-۲- معیارهای شباهت

اکثر محققان فعلی از معیارهای تشابه بر اساس فاصله اقلیدسی (ED) و شباهت کسینوسی (CS) استفاده کرده‌اند. معیار ED تنها بزرگی بردار و معیار CS تنها زاویه بین دو بردار را درنظر می‌گیرد، بنابراین این دو روش به تنهایی معیارهای موثر و مناسبی برای تحلیل متن نمی‌باشند. در این مقاله از ترکیب چندین معیار تشابه سند استفاده شده است که هر یک از آنها شباهت بین دو سند را از دیدگاهی متفاوت می‌سنجند سپس یک ماتریس تشابه اجماع برای هر مجموعه داده محاسبه می‌شود. ماتریس اجماع، میانگینی از سه معیار مختلف تشابه متن به شرح زیر است:

۳-۲-۱ معیار شباهت DTFSM

در بیشتر معیارهای تشابه مرسوم، از حاصل‌ضرب اسکالر نرمال‌شده دو سند برای نرمال‌سازی حاصل‌ضرب آنها استفاده می‌شود. معیار تشابه مبتنی بر فرکانس-مدت[۸] (DTFSM) اثربخشی خوشه‌بندی اسناد را با در نظر گرفتن تفاوت نرمال‌شده بین تکرار ترم‌ها[۹] بهبود می‌بخشد و پیچیدگی و تعداد عملیات مورد نیاز را برای خوشه‌بندی متن، در مقایسه با سایر معیارها بااستفاده از معادله زیر کاهش می‌دهد [30].

$$\mathrm{DTFSM}(V,W) = 1 - \sqrt{\frac{\sum_{i=1}^{N}|d_{1,i} - d_{2,i}|}{2\sum_{i=1}^{N}|d_{1,i}.d_{2,i}|}} \quad (۴)$$

(V و W هر دو سند متنی هستند و $d_{j,i} > 0$ وزن ترم iام در سند j است)

۳-۲-۲ معیار شباهت TS-SS

با ترکیب چندین معیار هندسی، معیار تشابه مساحت مثلث-مقطع[10] (TS-SS) [18] شباهت بین اسناد مختلف را از منظر بهتری نسبت به معیارهای مشابه هندسی و غیرهندسی محاسبه می‌کند. این روش با ترکیب دو معیار تشابه شباهت مساحت-مثلث (TS) و شباهت مساحت-مقطع (SS) معایب معیارهای ارزیابی معمولی ED و CS را برطرف می‌کند. در روش تشابه مساحت-مثلث، مساحت مثلث بین دو بردار (V و W) به‌عنوان معیار تشابه مورد استفاده قرار می‌گیرد و به این ترتیب، هم بزرگی بردارها و هم زاویه (α) بین آنها به‌صورت زیر در محاسبه لحاظ می‌شود،

$$\mathrm{TS}(V,W) = |W|.|V|.\sin(\alpha)/2 \quad (۵)$$

$$\alpha = \cos^{-1}(\cos(W,V) + 10) \quad (۶)$$

اما از آنجایی که معیار ارزیابی شباهت TS تفاوت بزرگی (MD) دو بردار را اندازه نمی‌گیرد، شباهت مساحت-مقطع (SS) به‌صورت زیر درنظر گرفته می‌شود:

$$\mathrm{SS}(V,W) = \pi.\left(\mathrm{ED}(V,W) + \mathrm{MD}(V,W)\right)^2.\left(\frac{\alpha}{360}\right) \quad (۷)$$

بنابراین معیار تشابه TS-SS از حاصلضرب معیارهای TS و SS به‌دست می‌آید. هنگامی که دو بردار حداکثر شباهت را داشته باشند (از نظر جهت و بزرگی)، معیار TS-SS برابر با صفر خواهد بود. حداکثر مقدار این معیار بی‌نهایت است.

۳-۲-۳ معیار شباهت PDSM

معیار تشابه اسناد زوجی[11] (PDSM) [31] معیاری است که می‌تواند مشابه‌ترین اسناد را در میان اسنادی که دارای درجه یکسانی از تشابه به یک سند باشند، بیابد. در این روش ارزیابی براساس وزن ترم‌ها و تعداد عبارات موجود در یکی از دو سند انجام می‌شود. بنابراین، شباهت دو سند (V و W) با افزایش تعداد ترم‌های (d) به‌کاررفته در هر دو افزایش می‌یابد و با افزایش ترم‌های استفاده شده تنها در یکی از آنها کاهش می‌یابد. روابط این روش در ادامه آورده شده است.

$$\mathrm{PDSM}(V,W) = \frac{V \cap W}{V \cup W} \times \frac{\mathrm{PF}(V,W)+1}{\mathrm{M}-\mathrm{AF}(V,W)+1}$$
$$\text{s.t.}\quad V \cap W = \sum_{i=1}^{N} \min\left(d_{1,i}, d_{2,i}\right),$$
$$V \cup W = \sum_{i=1}^{N} \max\left(d_{1,i}, d_{2,i}\right) \quad (۸)$$

بطوری‌که PF تعداد ترم‌های موجود در هر دو سند را نشان می‌دهد، AF تعداد ترم‌های غایب در یکی از دو سند است، N تعداد کل عبارات و، $d_{i,j}$ وزن ترم iام در سند j است.

۳-۳ شباهت جامع

فرض کنید K ماتریس داریم که هریک نمای متفاوتی از دادگان را نمایش می‌دهند. ماتریس‌ها با استفاده از تابع همجوشی $F: \{M_1, M_2, M_3, \dots M_k\} \to Z$ ترکیب می‌شوند تا ماتریس خروجی Z به‌دست آید. برای سادگی، فرض می‌کنیم که همه ماتریس‌های ورودی و خروجی دارای ابعاد R^{m*n} هستند. ماتریس Z را می‌توان با استفاده از انواع توابع مختلف ترکیب کرد. در زیر به چند مورد از آنها می‌پردازیم.

۱. تابع جمع- هر درایه ماتریس از مجموع درایه‌های ماتریس‌های دیگر به‌دست می‌آید.

$$Z_{ij} = \sum_{m=1}^{k} [M_{ij}]^k \quad (۹)$$

۲. تابع حداکثر- حداکثر مقدار را برای ورودی مورد نظر در ماتریس‌ها برمی‌گرداند.

$$Z_{ij} = \left[{M_{\mathrm{ij}}}^1, {M_{\mathrm{ij}}}^2, {M_{\mathrm{ij}}}^3, \dots, {M_{\mathrm{ij}}}^k \right] \quad (۱۰)$$

۳. تابع الحاق- این تابع نتیجه را با الحاق ماتریس‌های ورودی می‌سازد.

$$Z_{ij} = \left[{M_{\mathrm{ij}}}^1, {M_{\mathrm{ij}}}^2, {M_{\mathrm{ij}}}^3, \dots, {M_{\mathrm{ij}}}^k \right] \quad (۱۱)$$

۴. تابع میانگین- هر درایه از میانگین درایه‌های سایر ماتریس‌ها به‌دست می‌آید.

$$Z_{ij} = \frac{\sum_{m=1}^{k} [M_{ij}]^k}{k} \quad (۱۲)$$

در این مقاله از میان روش‌های پیشنهادی، از روش میانگین برای ترکیب ماتریس‌های شباهت استفاده می‌کنیم. در نتیجه لازم است که ماتریس‌های شباهت به‌دست آمده همسان‌سازی شده و در محدوده یکسانی قرار داشته باشند.

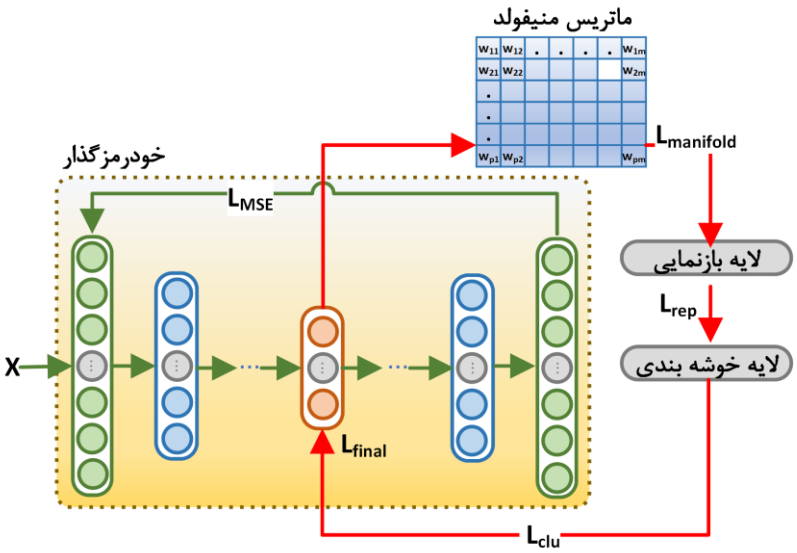
۳-۴ روش ارائه شده

امروزه روش‌های سنتی خوشه‌بندی متن، برای مدیریت دادگان با ابعاد بالا به دلیل نیاز به زمان و حافظه ناکارآمد هستند [32, 33]. رویکردهای کاهش ابعاد با روش‌های مختلف، دادگان را به فضایی با ابعاد بسیار کوچک‌تر منتقل می‌کنند و سپس خوشه‌بندی را در فضای جدید انجام می‌دهند. کارایی خوشه‌بندی به شدت به کیفیت نمایش این دادگان بستگی دارد. امروزه استفاده از شبکه‌های عصبی عمیق (DNN) برای یادگیری بازنمایی مناسب برای خوشه‌بندی، کیفیت خوشه‌بندی را به‌طور قابل توجهی افزایش داده است [34, 35]. این روش خوشه‌بندی عمیق نامیده می‌شود. خوشه‌بندی عمیق به‌طور کلی شامل دو فرآیند اساسی است: کاهش ابعاد و خوشه‌بندی. در روش‌های پیشین معمولا از خودرمزگذار به تنهایی برای کاهش ابعاد استفاده می‌شد که روش دقیقی نیست. در این پژوهش از ماتریس شباهت جدیدی، که از دیدگاه‌های متفاوت شباهت دادگان را در قالب یک منیفولد اطلاعاتی، ارائه می‌دهد استفاده شده است. این ماتریس شباهت، به لایه پنهان خودرمزگذار در قالب یک عبارت

جریمه اضافه می‌شود تا دقت کاهش ابعاد را افزایش دهد. این امر خود باعث می‌شود که دادگانی که شبیه‌تر هستند با دقت بیشتری پس از خوشه‌بندی نیز در کنار یکدیگر قرار گیرند. بعلاوه در روش‌های پیشین، کاهش ابعاد به‌عنوان مرحله پیش‌پردازش برای خوشه‌بندی استفاده شده است، بنابراین فضای بازنمایی ویژگی به‌دست آمده ممکن است برای خوشه‌بندی مناسب نباشد. در روش‌های نوین از تکنیک‌های جدیدتری برای ترکیب کاهش ویژگی و خوشه‌بندی

جدول ۱: تحقیقات انجام شده در حوزه خوشه‌بندی متن

دسته	نویسنده، مرجع	روش
روشهای مبتنی بر فاکتورسازی ماتریس غیرمنفی	اقدم و همکاران [۱]	در این روش با استفاده از فاصله مبتنی بر بردارهای ویژگی، با اعمال جریمه بر جفت بردارهای ویژگی، محدودیت‌ها را تعمیم داده و خوشه‌بندی را انجام می‌دهد.
	سان و همکاران [۲]	از معماری خودرمزگذار متقارن غیرمنفی با اعمال محدودیت‌های تقارن و غیرمنفی برای یادگیری بازنمایی نهفته ورودی‌ها استفاده می‌کند. هدف این روش افزایش تفسیرپذیری و اثربخشی خوشه‌بندی متن است.
	شی و همکاران [۳]	رویکردی جدید برای مدل‌سازی موضوع در متون کوتاه ارائه می‌کند. هدف این روش، استخراج موضوعات معنادار از داده‌های متنی کوتاه بااستفاده از ترکیب فاکتورسازی ماتریس غیرمنفی با همبستگی‌های متن-کلمه محلی، می‌باشد.
	لی و همکارن [۵]	این روش ساختارهای گراف جهانی و محلی را برای ثبت اطلاعات جهانی و محلی در داده‌ها ترکیب می‌کند. از فاکتورسازی مفهومی برای یادگیری یک بازنمایی کم بعد که ساختار ذاتی داده‌ها را حفظ می‌کند، استفاده می‌کند. با ترکیب منظم‌سازی نمودار دوگانه، رویکرد پیشنهادی به طور موثری چالش‌های دادگان با ابعاد بالا را کنترل می‌کند.
	صلاحیان و همکاران [۶]	یک معماری خودرمزگذار عمیق برای فاکاتورسازی ماتریس غیرمنفی، با هدف یافتن روابط میان ویژگی‌های پیچیده در دادگان ارائه می‌کند. برای افزایش کیفیت بازنمایی‌های آموخته شده، نویسندگان یک تنظیم‌کننده متضاد را معرفی می‌کنند که یادگیری ویژگی‌های متمایز را بهبود می‌بخشد.
روشهای مبتنی بر یادگیری عمیق	گوان و همکاران [۸]	از یک خودرمزگذار متن از پیش‌آموزش دیده با چارچوب خوشه‌بندی متن مبتنی بر ویژگی عمیق استفاده می‌کند.
	دیالو و همکاران [۹]	یک خود رمزگذار انقباضی برای نمایش اسناد و یک چارچوب خوشه‌بندی عمیق را پیشنهاد کرده است.
	حسینی و همکاران [۱۱]	یک خودرمزگذار پشته‌ای با خوشه‌بندی K-means برای خوشه‌بندی متن عمیق را ارائه کرده است.
	دانشفر و	این روش خوشه‌بندی متن را با ترکیب یک تنظیم‌کننده الاستیک سازگار با تنظیم گراف برای مدیریت انواع



شکل ۱: یادگیری بازنمایی یکپارچه با خودرمزگذار و خوشه‌بندی

در یک چارچوب استفاده می‌شود که یادگیری این دو رویه همزمان انجام می‌شود. به عبارت دیگر، پارامترهای شبکه عصبی و نحوه اختصاص ویژگی‌های کاهش‌یافته به‌دست آمده به خوشه‌ها، به‌طور مشترک آموخته می‌شوند. ساختار یکپارچه یادگیری خوشه‌بندی عمیق مشترک (یعنی یادگیری بازنمایی با خودرمزگذار و خوشه-بندی بطور همزمان) در شکل ۱ نشان می‌دهد که چگونه می‌توان از یک تابع زیان جدید برای حل مشکل خوشه‌بندی استفاده کرد.

مدل ما شامل یادگیری بازنمایی است که هدف آن یادگیری تعبیه‌های مناسب برای خوشه‌بندی از طریق خودرمزگذار و ماتریس شباهت و همچنین تخمین خوشه‌ها با معیار ارزیابی مناسب است. تابع زیان کلی به شکل زیر است:

$$L_{final} = L_{rep} + L_{clu} \qquad (۱۳)$$

که در آن L_{rep} و L_{clu} به‌ترتیب تابع هدف یادگیری بازنمایی و تابع هدف خوشه‌بندی را نشان می‌دهند. در مرحله یادگیری بازنمایی، از خروجی خودرمزگذار و اطلاعات منیفولد به‌عنوان جریمه استفاده می‌شود. بنابراین، تابع جریمه بازسازی خودرمزگذار به‌صورت زیر تعریف می‌شود:

$$L_{rep} = L_{MSE}(X, X') + L_{manifold}(Z, Z') \qquad (۱۴)$$

هنگامی که نمایش نهفته $H_a^l \in R^{n*d}$ را از متن داده به‌دست آوردیم، می‌توانیم از آن برای انتساب خوشه نرم $Q \in R^{n*d}$ استفاده کنیم، بطوریکه q_{ik} توسط معادله (۱۵) به‌دست می‌آید. این معادله بیانگر احتمال تخصیص نمونه i به خوشه k است. از آنجایی‌که q_{ik} شباهت میان نمایش نمونه z_i و مرکز خوشه‌بندی μ_k را از طریق توزیع t-Student به‌عنوان یک هسته اندازه‌گیری می‌کند، زمانی که نمونه‌ها به مرکز خوشه نزدیک‌تر باشند، نرم احتمال بالا و قابلیت اعتماد بالایی پیدا می‌کنند.

$$q_{ik} = \frac{(1+\|z_i-\mu_k\|^2)^{-1}}{\sum_k (1+\|z_i-\mu_k\|^2)^{-1}} \qquad (۱۵)$$

سپس توزیع هدف P نمونه‌ها از طریق معادله (۱۶) ایجاد می‌شود. این توزیع روی دادگانی که با اطمینان بالاتری به خوشه اختصاص داده شده است، تأکید بیشتری دارد. شایان ذکر است که برای به-دست آوردن مراکز اولیه ابتدا خودرمزگذار را از قبل با به حداقل رساندن جریمه بازسازی[۱۲] آموزش می‌دهیم. پس از آن، K-means را روی بازنمایی‌های آموخته‌شده اجرا کرده و سپس در ادامه تکرار آموزش‌ها، از خوشه‌بندی K-means استفاده نمی‌شود.

$$p_{ik} = \frac{{q_{ik}}^2/\sum_i q_{ik}}{\sum_k ({q_{ik}}^2/\sum_i q_{ik})} \qquad (۱۶)$$

هدف ما این است که تخصیص فعلی Q به توزیع هدف P نزدیک شود. خوشه‌های نهایی را می‌توان از طریق آموزش مشترک خودرمزگذار و تخصیص خوشه به‌دست‌آورد. تابع جریمه خوشه‌بندی با توجه به واگرایی کولبک -لیبلر (KL)[13] بین دو توزیع به‌صورت زیر تعریف می‌شود.

$$L_{clu} = kl(Q\|P) = \sum_i \sum_k p_{ik} \log\left(\frac{q_{ik}}{p_{ik}}\right) \qquad (۱۷)$$

مراحل روش DCTMA در الگوریتم ۱ نشان داده شده است.

۴- آزمایشات

برای ارزیابی کارایی و پیچیدگی محاسباتی چارچوب پیشنهادی در خوشه‌بندی متن، آزمایش‌هایی روی مجموعه دادگان، Reuters-10k، 20Newsgroups و WebKB انجام شده است. این بخش کارایی مدل پیشنهادی را در مقایسه با سایر روش‌ها بااستفاده از این سه مجموعه داده ارزیابی می‌کند. در بخش ۴.۱، هشت روش مقایسه

معروف شامل K-means، NMF_f، NMF_{KL}، خوشه‌بندی سلسله مراتبی، خوشه‌بندی BIRCH، خوشه‌بندی mini-batch K-means، خوشه‌بندی تعبیه‌شده عمیق[۱۴] (DEC)، فاکتورگیری ماتریس غیرمنفی با منظم‌ساز نامتقارن[۱۵] (RANMF)[1] ، AE-NMF [۲]، SeaNMF [۳]، DGLCF [۵]، DANMF-CRFR [۶]، EDA-TEC [۴] بررسی قرار می‌گیرند. مجموعه دادگان استفاده شده، در بخش ۴.۲ بررسی می‌شود. بخش ۴.۳ معیارهای ارزیابی را خلاصه می‌کند و بخش ۴.۴ نتایج ارزیابی عملکرد روش پیشنهادی در سه مجموعه داده مختلف را با چندین روش معروف مقایسه و بر اساس پنج معیار ارزیابی متداول توضیح می‌دهد.

الگوریتم ۱: روش خوشه‌بندی متن عمیق براساس منیفولد محلی در لایه خودرمزگذار

Input:
X: input data.
k: number of clusters.
ε: stopping threshold.
T: target interval,
$MaxIter$: Maximum iterations.
Output:
Cluster representatives *R*, Labels *S;*
1: $\acute{\chi}$ = Compute TF-IDF matrix of input data *X*
2: for $iter \in 0, 1, \dots, Maxiter$
3: if $iter \,\%\, T == 0$
4: Compute the embeddings for all samples
5: Compute target distribution (P) by Eq. (16)
6: Save last label assignments: $s_{old} = s$
7: Compute new label assignments by Eq. (17)
8: if $(sum\ (s_{old} \neq s) / n < \varepsilon)$
9: Stop training
10: Choose a batch of samples $s \in \acute{\chi}$
11: Compute manifold matrix on s by Eq. (12)
12: Update network parameters on s

۴-۱ روش‌های مقایسه

در این بخش، عملکرد روش پیشنهادی با الگوریتم‌های خوشه‌بندی مختلف مقایسه خواهد شد تا نشان داده شود چگونه مدل پیشنهادی می‌تواند عملکرد بهتری نسبت به این روش‌ها داشته باشد. الگوریتمهای خوشه‌بندی مختلف مورد بررسی، به‌طور خلاصه به شرح زیر است:

- روش خوشه‌بندی K-means: الگوریتم تکراری K-means مجموعه دادگان را به K زیرگروه غیرهمپوشان منحصر به فرد (خوشه) تقسیم می‌کند. این الگوریتم نقاط داده را طوری به خوشه‌ها تخصیص می‌دهد که مجموع مربع فواصل بین نقاط داده و مراکز خوشه، که میانگین مقادیر همه نقاط داده خوشه است، کمینه شود. میزان تشابه نقاط داده در داخل یک خوشه با کاهش واریانس درون خوشه افزایش می‌یابد [36].

- روش خوشبندی NMF: در این مطالعه، از دو روش NMF_F و NMF_{KL} استفاده می‌شود [37].

- روش خوشه‌بندی سلسله مراتبی: هدف از این روش ایجاد ساختار درختی از مجموعه‌ای از تکنیک‌های خوشه‌بندی است که به شکل یک درخت سازمان‌یافته ترازبندی می‌شوند. این تکنیک‌ها با تقسیم شیء ورودی به صورت بازگشتی به شیوه بالاگشتی یا پایین‌گشتی خوشه‌بندی را ایجاد می‌کنند. دراستفاده از خوشه‌بندی سلسله مراتبی، نیازی به تعیین تعداد خوشه‌ها از قبل نیست. بااستفاده از این روش، گروه‌های کوچکتری ایجاد خواهند شد که می‌تواند تشابه دادگان را نشان دهد [38].

- روش خوشه‌بندی BIRCH[۱۶]: این الگوریتم، خوشه‌بندی سلسله مراتبی را روی مجموعه دادگان بزرگ اعمال می‌کند. در این الگوریتم مجموعه دادگان به بخش‌های کوچک تقسیم شده و گروه‌بندی می‌شوند. این الگوریتم یک ساختار درختی برای دادگان ارائه شده ایجاد می‌کند که به آن درخت ویژگی کوچکترین مربع‌ها گویند [39].

-روش خوشه‌بندی Mini-batch K-means: این الگوریتم نسخه‌ای از الگوریتم K-means است که برای خوشه‌بندی مجموعه دادگان بزرگ استفاده می‌شود. در هر مرحله از اجرا، این روش از بخشی از دادگان با اندازه ثابت به جای کل مجموعه داده استفاده می‌کند که هزینه محاسباتی را کاهش می‌دهد [40].

-الگوریتم DEC: این الگوریتم به‌طور همزمان بازنمایی ویژگی و تخصیص خوشه را یاد می‌گیرد. این مدل بر اساس مکانیزم خودرمزگذار عمیق طراحی شده است که تبدیل ویژگی و خوشه‌بندی را به‌صورت همزمان بهبود می‌بخشد تا بتواند داده ورودی را به یک فضای تعبیه با بعد کم نگاشت کند. این مدل یک توزیع هدف فرعی و فاصله اطلاعاتی کولبک-لیبلر را محاسبه می‌کند تا خوشه‌بندی را مقاوم کرده و پارامترها را بهینه سازد [41].

-روش RANMF [1]: روش جدیدی است که با استفاده از فاصله مبتنی بر بردارهای ویژگی، با اعمال جریمه بر جفت بردارهای ویژگی، محدودیت‌ها را تعمیم می‌دهد.

-روش AE-NMF [۲]: از معماری خودرمزگذار متقارن غیرمنفی با اعمال محدودیت‌های تقارن و غیرمنفی برای یادگیری بازنمایی نهفته ورودی‌ها استفاده می‌کند. هدف این روش افزایش تفسیرپذیری و اثربخشی خوشه‌بندی متن است.

-روش SeaNMF [۳]: رویکردی جدید برای مدل‌سازی موضوع در متون کوتاه ارائه می‌کند. هدف این روش، استخراج موضوعات

معنادار از داده‌های متنی کوتاه بااستفاده از ترکیب NMF با همبستگی‌های متن-کلمه محلی، می‌باشد.

-روش DGLCF [۵]: رویکردی جدید برای خوشه‌بندی داده‌ها به نام فاکتورسازی مفهومی جهانی و محلی دوگانه پیشنهاد می‌کند. این روش ساختارهای گراف جهانی و محلی را برای ثبت اطلاعات جهانی و محلی در داده‌ها ترکیب می‌کند. از فاکتورسازی مفهومی برای یادگیری یک بازنمایی کم بعد که ساختار ذاتی داده‌ها را حفظ می‌کند، استفاده می‌کند. با ترکیب منظم‌سازی نمودار دوگانه، رویکرد پیشنهادی به طور موثری چالش‌های داده‌های با ابعاد بالا را کنترل می‌کند.

-روش DANMF-CRFR [۶]: یک معماری خودرمزگذار عمیق برای NMF، با هدف یافتن روابط میان ویژگی‌های پیچیده در داده‌ها است. برای افزایش کیفیت بازنمایی‌های آموخته شده، نویسندگان یک تنظیم‌کننده متضاد را معرفی می‌کنند که یادگیری ویژگی‌های متمایز را بهبود می‌بخشد.

-روش EDA-TEC [۴]: این روش خوشه‌بندی متن را با ترکیب یک تنظیم‌کننده الاستیک سازگار با تنظیم گراف برای مدیریت انواع مختلف نویز و حفظ یکپارچگی ساختار داده افزایش می‌دهد.

۴-۲- مجموعه دادگان

در این مقاله، سه مجموعه داده متداول و متفاوت برای ارزیابی عملکرد روش پیشنهادی به شرح زیر استفاده شده است:

۴-۲-۱-مجموعه داده Reuters-10k

این دادگان یک زیرمجموعه ۱۰۰۰۰ نمونه‌ای است که بصورت تصادفی از http://www.daviddlewis.com/resources/testcollection/s/reuters21578) انتخاب شده و ویژگی‌های فراوانی ترم – فراوانی سند معکوس[۱۷] (TF-IDF) روی ۲۰۰۰ کلمه پرتکرار آن اعمال شده است. در این مقاله، تنها از چهار مجموعه اصلی آن به عنوان برچسب استفاده شده است: کسب و کار/صنعت، دولت/اجتماعی، بازار و اقتصاد.

۴-۲-۲-مجموعه داده WebKB

این مجموعه داده http://www.cs.umb.edu/~smimarog/textmining/datas/ets) شامل اطلاعات ۸۳۳۴ صفحه وب است که توسط پروژه پایگاه دانش جهانی گروه یادگیری متن CMU در سال ۱۹۹۷ جمع‌آوری شده است. این دادگان به صورت دستی در هفت دسته: دانشجو، اعضای هیات علمی، درس، پروژه، کارمند، بخش و غیره دسته‌بندی شده است. در این آزمایش، دسته‌بندی‌های بخش، کارمند و دیگر حذف شده‌اند زیرا تعداد نمونه‌های کمی دارند. در ادامه توزیع تعداد سندها در هر خوشه به شرح زیر است: دسته دانشجو شامل ۱۶۴۱ سند (۱۰۹۷ سند برای دادگان آموزشی و ۵۴۴ سند برای دادگان آزمایشی)، دسته اعضای هیات علمی شامل ۱۱۲۴ سند (۷۵۰ سند برای دادگان آموزشی و ۳۷۴ سند برای دادگان آزمایشی)، دسته درس شامل ۹۳۰ سند (۶۲۰ سند برای دادگان آموزشی و ۳۱۰ سند برای دادگان آزمایشی) و دسته پروژه شامل ۵۰۴ سند (۳۳۶ سند برای دادگان آموزشی و ۱۶۸ سند برای دادگان آزمایشی).

۴-۲-۳ مجموعه داده 20Newsgroups

این مجموعه داده شامل ۱۸۸۲۱ سند از مجموعه داده 20Newsgroups است که شامل ۲۰ خوشه مختلف با ۱۰۰۰ سند با دسته‌بندی‌های متوازن بوده که براساس موضوعات مختلف خوشه‌بندی شده‌اند (جدول ۲ را ببینید)

۴-۳ معیارهای ارزیابی

عملکرد خوشه‌بندی با استفاده از پنج معیار رایج، اطلاعات متقابل نرمال‌شده[۱۸] (NMI)، اطلاعات متقابل تنظیم‌شده[۱۹] (AMI)، ضریب سیلوئت[۲۰] (SC)، دقت (ACC) و شاخص تصادفی تنظیم-شده (ARI) ارزیابی خواهد شد.

۴-۳-۱ NMI و AMI

AMI و NMI [42] دو روش مختلف برای ارزیابی اطلاعات متقابل (MI) هستند. در نظریه احتمال و نظریه اطلاعات، روش MI بین دو خوشه کمیتی برای نشان دادن درجه وابستگی دو خوشه به یکدیگر است. این مفهوم به‌طور معمول با آنتروپی یک خوشه مرتبط است که میزان اطلاعات موجود در خوشه دیگر را نشان می‌دهد. با این‌حال، اطلاعات متقابل بین خوشه‌ها (با تعداد ثابتی از عناصر) در هنگام افزایش تعداد خوشه‌ها به‌طور قابل توجهی زیاد می‌شود و مقدار ثابتی نمی‌گیرد. NMI شاخصی است برای مقایسه نتایج بین حداقل اطلاعات متقابل و همبستگی کامل میان آنها. درحالی‌که AMI معیاری است که می‌توان از آن برای مقایسه خوشه‌ها و نرمال‌سازی آنها به نسبت شانس استفاده کرد. در ادامه نحوه محاسبه شاخص AMI توضیح داده شده است.

جدول ۲: توزیع دسته‌های مختلف در مجموعه داده 20Newsgroups

نام دسته	تعداد سندها
alt.athesim	799
comp.graphics	973
comp.os.ms.windows.misc	966
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	963
comp.windows.x	985
misc.forsale	975
rec.autos	989
rec.motorcycles	996
rec.sport.hockey	999
sci.crypt	991
sci.electronics	984
sci.med	990
sci.space	987
Soc.religion.christian	996
talk.politics.guns	909
talk. politics.mideast	940
talk. politics.misc	775
talk.religion.misc	628

مجموعه‌ای از N عنصر را با دو بخش مجزا، S (با C خوشه) و S' (با C' خوشه) درنظر بگیرید. در این‌حالت جدول وابستگی با $C \times C'$ عنصر برای خلاصه‌سازی اطلاعات متقابل میان خوشه‌ها، وجود دارد. بطوریکه،

$$X = [x_{ij}]_{j=1\dots C'}^{i=1\dots C} \quad (۱۸)$$

درصورتی که x_{ij} تعداد نمونه‌هایی باشد که به خوشه‌های S_i و S'_j تعلق دارند، آنگاه احتمال انتخاب داده‌ی تصادفی که به خوشه‌ی S_i تعلق داشته باشد برابر است با:

$$P_S(i) = |S_i|/N \quad (۱۹)$$

و

$$P_{SS'}(i) = |S_i \cap S'_j|/N \quad (۲۰)$$

احتمال تعلق داده به دو خوشه‌ی S_i و S'_j است. بنابراین MI بین دو خوشه‌ی S و S' برابر است با:

$$\mathrm{MI}(S,S') = \sum_{i=1}^{S}\sum_{j=1}^{S'} P_{SS'}(i,j)\log\frac{P_{SS'}(i,j)}{P_S(i)P_{S'}(j)} \quad (۲۱)$$

و با در نظر گرفتن MI و $E\{MI(S,S\text{^}')\}$ مورد انتظار بین دو خوشه‌ی تصادفی S و S' با آنتروپی‌های $H(S)$ و $H(S')$، AMI برابر خواهد بود با:

$$\mathrm{AMI}(S,S') = \frac{\mathrm{MI}(S,S')-\mathrm{E}\{\mathrm{MI}(S,S')\}}{\mathrm{Avg}\{\mathrm{H}(S),\mathrm{H}(S')\}-\mathrm{E}\{\mathrm{MI}(S,S')\}} \quad (۲۲)$$

زمانی‌که مقدار AMI به یک نزدیک باشد، نشان‌دهنده این است که دو خوشه با هم سازگار هستند و زمانی که مقدار AMI برابر با صفر است، بیانگر استقلال خوشه‌ها است.

SC ۲-۳-۴

یکی از مشکلات معیارهای مبتنی بر اطلاعات متقابل، نیاز به برچسب‌گذاری دستی توسط انسان است. ضریب SC یک معیار مستقل از برچسب خوشه‌هاست. این معیار تصویری از طبقه‌بندی اشیاء ارائه می‌دهد. مقدار ضریب SC نشان می‌دهد که یک عنصر به چه میزان به خوشه خود یا خوشه دیگری تعلق دارد. این ضریب مقادیری در بازه $[-۱، ۱]$ اخذ می‌کند [43]. هرچه مقدار ضریب SC بالاتر باشد، خوشه‌بندی بهتر است. برای محاسبه ضریب SC، هر معیار فاصله‌سنجی مانند اقلیدسی یا منهتن می‌تواند برای یافتن فاصله $d(i,j)$ بین دو داده i و j در خوشه S تعیین شود. بنابراین، برای هر نقطه داده i، میانگین فاصله نقطه i و سایر دادگان از رابطه زیر بدست می‌آید:

$$m(i) = \frac{1}{|S|-1}\sum_{j\in S, i\neq j} d(i,j) \quad (۲۳)$$

هر چه $m(i)$ کوچکتر باشد، نشان‌دهنده تخصیص بهتر خوشه است. از سوی دیگر $n(i)$ معیار دیگری است برای محاسبه میانگین عدم شباهت (واگرایی) داده i با داده موجود در خوشه S' به صورت زیر،

$$n(i) = \min_{S\neq S'} \frac{1}{|S'|}\sum_{j\in S'} d(i,j) \quad (۲۴)$$

اندازه $n(i)$ نشان‌دهنده همبستگی بین عناصر خوشه نسبت به سایر خوشه‌ها است. هرچه مقدار $n(i)$ بزرگتر باشد، نشان می‌دهد که خوشه دارای واکنش و واگرایی کمتری با سایر خوشه‌ها است و عناصر داخل خوشه به نسبت عناصر سایر خوشه‌ها فاصله بیشتری دارند. بنابراین، هدف این است که مقدار $n(i)$ را به حداکثر برسانیم تا خوشه‌ها از یکدیگر جدا شوند و تفاوت بین خوشه‌ها بیشتر شود. به‌طورکلی، بهبود $m(i)$ و افزایش $n(i)$ منجر به نتایج بهتر در خوشه‌بندی و نزدیکی عناصر داخل خوشه و فاصله بین خوشه‌ها می‌شود.

خوشه‌ای که کمترین میانگین اختلاف را دارد، مناسب‌ترین خوشه برای همسایه بعدی نقطه i است. بنابراین مقدار SC یک نقطه داده i به‌صورت زیر تعیین می‌شود:

$$SC(i) = \begin{cases} 1-\frac{m(i)}{n(i)}, & \text{if } n(i) > m(i) \\ 0, & \text{if } n(i) = m(i) \\ \frac{n(i)}{m(i)}-1, & \text{if } n(i) < m(i) \end{cases} \quad (۲۵)$$

ARI ۳-۳-۴

در آمار، شاخص تصادفی (RI) در خوشه‌بندی داده، معیاری است که میزان شباهت بین دو خوشه را اندازه‌گیری می‌کند. این شاخص به‌صورت زیر محاسبه می‌شود:

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \quad (۲۶)$$

که در آن TN تعداد منفی‌های واقعی، TP تعداد مثبت‌های واقعی، FN تعداد منفی‌های غلط و FP تعداد مثبت‌های غلط هستند.

شاخص ARI [45] [44] نسخه‌ای از RI است که به شکل زیر تنظیم شده است:

$$ARI = \frac{4-1-4\ RI-E\{RI\}}{max(RI)-E\{RI\}} \quad (۲۷)$$

طبق تعریف بالا، شاخص تصادفی برای برچسب‌گذاری تصادفی و مستقل از تعداد نمونه، مقدار صفر و زمانی‌که خوشه‌ها یکسان هستند مقداری برابر با یک دارد.

۴-۳-۴- دقت

دقت بهترین تطابق بین برچسب‌های واقعی (y) و برچسب‌های پیش‌بینی شده با خوشه‌بندی ($\hat{y}$) [46] را میابد. اگر p تمام جایگشت‌های K خوشه باشد، آنگاه ACC برای n نمونه مختلف برابر است با:

$$ACC(y,\hat{y}) = \max_{perm\in P} \frac{1}{n}\sum_{i=0}^{n-1} 1(perm(\hat{y}_i) = y_i) \quad (۲۸)$$

۴-۴ نتایج و بحث

برای ارزیابی عملکرد روش پیشنهادی، برخی از آزمایشات روی مجموعه داده معرفی شده در بخش ۴.۲ انجام شده است. نتایج بدست آمده با روش‌های متداول و فعلی خوشه‌بندی متن و بااستفاده از معیارهای ارزیابی مختلف مقایسه و ارزیابی شده است. برای این منظور، پیش از پردازش متن، لازم است آن را آماده کنیم. پیش‌پردازش یکی از مراحل اصلی در پردازش متن است. روش‌های معمول پیش‌پردازش شامل توکن‌بندی، حذف کلمات توقف و ریشه‌یابی کلمات است. علاوه بر این، از مدل فضای برداری TF-IDF [47] برای تبدیل متن به اعداد استفاده کردیم. در این مرحله برای هر مجموعه داده، ۲۰۰۰ کلمه پرتکرار انتخاب شده و تعداد تکرار برابر با ۸۰۰۰ بار و اندازه خوشه برابر با ۶۴، در نظر گرفته شده است. آزمایشات با استفاده از چارچوب کراس روی یک رایانه با پردازنده Intel corei7-4700HQ، CPU 2.40 GHz، ۸ گیگابایت رم و کارت گرافیک Nvidia GTX 740 و با استفاده از سه مجموعه داده به شرح زیر انجام شده است.

۴-۴-۱ مجموعه داده Reuters-10k

در جدول ۲، نتایج خوشه‌بندی بر روی مجموعه داده Reuters-10k برای ۲۰۰۰ ویژگی انتخاب‌شده آورده شده است. تمام ارزیابی‌ها برای تعداد خوشه‌های {۲، ۴، ۶، ۸ و ۱۰} انجام شده و نهایتا، میانگین به عنوان نتیجه نهایی در جدول نشان داده شده است. همانطور که مشخص است، الگوریتم پیشنهادی بهتر از روش‌های دیگر عمل کرده است (برای تمامی خوشه‌ها آزمایشات چندین بار اجرا شده است که در بیشتر موارد، نتایج بهتر از میانگین و در برخی موارد، کمتر از میانگین بوده است). نتایج میانگین تمامی خوشه‌ها نشان می‌دهد که روش پیشنهادی نسبت به سایر روش‌های خوشه‌بندی بهترین عملکرد را از نظر پارامترهای AMI و SC داشته است (AMI = 0.5210 و SC = 0.0213). بعلاوه نتایج الگوریتم‌های NMF_F و RANMF نزدیک نتایج بدست آمده از روش پیشنهادی است. بااین‌حال، این دو روش حتی برای تعداد کمتری از خوشه‌ها نتوانستند نسبت به روش مبتنی بر یادگیری عمیقِ پیشنهادی، عملکرد خوشه‌بندی را بهبود بخشند. علاوه براین، از آنجاکه این الگوریتم‌ها قادر به حفظ و انتقال شباهتها بین بردارهای مختلف فضای اصلی به فضای با ابعاد کم نبوده‌اند، نتوانستند مانند روش پیشنهادی از اطلاعات منیفولد برای خوشه‌بندی بهره‌برداری کنند.

جدول ۳: نتایج خوشه‌بندی الگوریتم‌های مختلف روی مجموعه دادگان Reutters-10k

	AMI						SC					
	2	4	6	8	10	میانگین	2	4	6	8	10	میانگین
K-means	0.0644	0.3550	0.4818	0.4972	0.4803	0.3757	0.0086	0.0121	0.0149	0.0182	0.0204	0.0148
NMF_F	0.3230	0.3696	0.3636	0.4718	0.4135	0.3883	0.0133	0.0131	0.0154	0.0184	0.0195	0.0159
NMF_{KL}	0.3645	0.4738	0.4333	0.4204	0.4409	0.4265	0.0132	0.0136	0.0137	0.0152	0.0177	0.0146
Hierarchical clustering	0.0833	0.3609	0.4518	0.4357	0.4156	0.3494	0.0087	0.0098	0.0107	0.0096	0.0132	0.0104
Brich	0.0087	0.3442	0.4105	0.3958	0.3860	0.3090	0.0088	0.0068	0.0104	0.0095	0.0126	0.0096
MiniBatch K-means	0.3515	0.2584	0.4132	0.3971	0.3919	0.3624	0.0136	0.0092	0.0143	0.0155	0.0173	0.0139
DEC	0.3034	0.4976	0.4376	0.4431	0.4123	0.4188	0.0125	0.0131	0.0121	0.0098	0.0190	0.0133
RANMF	0.3668	0.5786	0.4858	0.3924	0.4347	0.4516	0.0134	0.0134	0.0143	0.0146	0.0137	0.0138
AE-NMF	0.2824	0.4554	0.415	0.4661	0.4835	0.4204	**0.027**	**0.0313**	-0.009	0.002	-0.007	0.0088
SeaNMF	0.2927	0.336	0.4007	0.503	0.476	0.4016	0.0271	0.0303	-0.004	-.0004	0.005	0.0116
DGLCF	0.3232	0.3184	0.4126	0.4957	0.4996	0.4099	0.0087	0.0061	0.0059	0.0065	0.0060	0.0066
DANMF-CRFR	0.2909	0.3637	0.4756	0.5116	0.4921	0.4267	0.0089	0.0101	0.0103	0.0181	0.0182	0.0131
EDA-TEC	0.3649	0.5575	0.4218	0.3947	0.4497	0.4377	0.0175	-.0119	-.0468	-.0428	-.0395	-.0246
DCTMA	**0.3938**	**0.6513**	**0.5391**	**0.5462**	**0.5250**	**0.5210**	0.0242	0.0137	**0.0374**	**0.0435**	**0.0367**	**0.0213**

در جدول ۲ و جدول ۳، نتایج ارزیابی روش پیشنهادی با تعداد خوشه‌های مختلف و معیارهای ارزیابی متفاوت آورده شده است. با توجه به نتایج به‌دست آمده، بیشترین مقادیر پارامترهای ACC، NMI، AMI و ARI زمانی به‌دست می‌آیند که چهار خوشه جداگانه وجود داشته باشد. بدترین نتایج پارامترهای NMI، AMI و ARI برای تعداد دو خوشه به‌دست آمده است. همچنین، برای

تعداد خوشه‌های هشت و چهار، بیشترین و بدترین مقادیر SC بدست می‌آیند و بدترین مقدار ACC متناظر با ده خوشه است.

۴-۴-۲ مجموعه داده 20Newsgroup

مجموعه داده 20Newsgroup، مشابه مجموعه دادگان Reuters-10k، نیز دارای عبارات زیادی است (حدود ۶۹،۲۳۶).

جدول ۴: عملکرد الگوریتم پیشنهادی روی مجموعه داده Reuters-10k برای معیارهای مختلف

		2	4	6	8	10
DCTMA	ACC	0.60940	0.84420	0.68240	0.59870	0.55570
	NMI	0.29398	0.65145	0.53933	0.54651	0.52544
	AMI	0.3938	0.65133	0.53910	0.54622	0.52507
	SC	0.02420	0.01370	0.03745	0.04352	0.03672
	ARI	0.29845	0.68203	0.54880	0.48074	0.41673

ما برای کاهش پیچیدگی، پس از پیش‌پردازش، تنها ۲۰۰۰ اصطلاح آن را انتخاب کرده‌ایم و نتایج ارزیابی را در جدول ۴ نشان داده‌ایم. همانطور که مشخص است، مدل پیشنهادی درحالت میانگین بهتر از روش‌های دیگر عمل کرده است (SC=0.0043 و AMI=0.4066) که نشان‌دهنده اهمیت مشارکت اطلاعات منیفولد در خوشه‌بندی دادگان است. پس از روش پیشنهادی، مدل‌های RANMF و DEC بهترین عملکرد را از نظر معیار AMI داشته‌اند و مدل‌های DEC و NMF_F نسبت به سایر روش‌ها عملکرد بهتری از نظر معیار اندازه‌گیری SC داشته‌اند. همچنین، در دو حالت - تعداد خوشه‌ها برابر با چهار و تعداد خوشه‌ها برابر با ۱۶ - روش MiniBatch K-means و NMF_F هر دو عملکرد بهتری نسبت به روش پیشنهادی بدست آورده‌اند.

جدول ۵ سایر معیارهایی که برای ارزیابی عملکرد مدل پیشنهادی در مجموعه داده 20Newsgroup استفاده شده است را نشان می‌دهد. همانطور که مشخص است، با افزایش تعداد خوشه‌ها از چهار خوشه تا بیست خوشه، معیار ACC افزایش یافته است. همچنین، مقادیر NMI، AMI و ARI با افزایش تعداد خوشه‌ها تا ۱۶ به‌طور قابل توجهی بهبود یافته‌اند؛ این معیارها با تعداد خوشه برابر با ۲۰ به مقدار کمی کاهش یافته‌اند.

۴-۴-۳ مجموعه داده WebKB

مجموعه داده WebKB نیز پس از پیش‌پردازش، ۷۶۴۷ عبارت دارد که نتایج خوشه‌بندی آن در جدول ٦ نشان داده شده است. طبق این نتایج، الگوریتم پیشنهادی نسبت به سایر روش‌ها عملکرد بهتری داشته است. در این جدول نتایج به‌ازای ویژگی‌های انتخابی متفاوت آمده است. برای ویژگی‌های انتخاب‌شده از ۵۰۰ تا ۲۰۰۰، بهترین مقدار AMI مربوط به روش پیشنهادی است و برای معیار SC، بهترین مقادیر برای ویژگی‌های انتخاب‌شده ۵۰۰ و ۱۰۰۰ نیز مربوط به روش پیشنهادی است. با این‌حال، بهترین مقدار SC برای ویژگی‌های انتخاب‌شده ۲۰۰۰، مربوط به روش خوشه‌بندی K-means است. سایر معیارهای ارزیابی در جدول ۷ گزارش شده است. با توجه به نتایج ارائه شده، بهترین مقادیر گزارش شده برای ACC و SC مربوط به ۵۰۰ ویژگی انتخابی هستند و بهترین مقادیر برای معیارهای NMI، AMI و ARI مربوط به ۲۰۰۰ ویژگی انتخاب‌شده می‌باشند.

با توجه به نتایج بدست‌آمده، روش پیشنهادی در تمام مجموعه‌های داده بهتر از اکثر روش‌های مقایسه‌شده عمل کرده است. چرا که استفاده از هندسه ذاتی نمونه‌ها و شباهت آنها، کارآمدی خوشه‌بندی را افزایش می‌دهد و نتایج بهتری روی مجموعه دادگان ارائه‌شده به‌دست می‌آورد که قابلیت روش پیشنهادی در خوشه‌بندی داده را نشان می‌دهد. هر مجموعه داده به طور کلی خصوصیات خاصی دارد که شامل تعداد نمونه‌ها، خوشه‌ها و ویژگی‌ها است و هریک از این ویژگی‌ها بر روی نتایج تأثیرگذار است. به‌عنوان مثال، هر چه تعداد نمونه‌ها بیشتر و تعداد کلاس‌ها کمتر باشد، نتایج بهتری بدست می‌آیند، زیرا به دلیل تعداد کمتر کلاس‌ها، نمونه‌ها به دسته‌های بزرگتر تقسیم می‌شوند که نیازمند دقت کمتری می‌باشد. در این بین، مقایسه نتایج حاصل با معیارهای مختلف بر روی سه مجموعه داده ارائه‌شده نشان می‌دهد که نتایج به‌دست آمده روی مجموعه دادگان 20Newsgroup به دلیل تعداد زیاد کلاس‌ها، کمترین مقدار ارزیابی‌ها را نسبت به دو مجموعه دیگر داشته است.

۵- نتیجه‌گیری

در این مقاله، یک روش خوشه‌بندی متن عمیق پیشنهاد شده است که از منیفلد محلی در لایه خودرمزگذار استفاده می‌کند. در این مدل، ابتدا یک ماتریس شباهت که میانگین ماتریس‌های شباهت مختلف از دادگان متنی است ایجاد شده و سپس با خودرمزگذار در فرایند یادگیری بازنمایی ادغام می‌شود. در این معماری از روش واگرایی کولبک-لیبلر بصورت انتها به انتها برای خوشه‌بندی استفاده می‌شود. آزمایش‌های متعدد روی مجموعه دادگان واقعی نشان می‌دهد که الگوریتم پیشنهادی کارایی و پایداری خوبی دارد.

جدول ۵: نتایج خوشه‌بندی الگوریتم‌های مختلف روی مجموعه دادگان 20Newsgroup

	AMI						SC					
	4	**8**	**12**	**16**	**20**	میانگین	**4**	**8**	**12**	**16**	**20**	میانگین
K-means	0.1784	0.2801	0.2829	0.2776	0.3371	0.2712	-.0006	-.0016	0.0023	0.0057	0.0031	0.0025
NMF$_F$	0.2104	0.2688	0.2696	0.3052	0.2820	0.2672	0.0011	0.0024	0.0038	**0.0070**	0.0042	0.0041
NMF$_{KL}$	**0.3235**	0.3006	0.3625	0.3629	0.3661	0.3431	0.0022	0.0013	0.0018	0.0027	0.0010	0.0018
Hierarchical clustering	0.1923	0.2664	0.2967	0.3245	0.3338	0.2827	-.0032	-.0011	-.0015	0.0001	0.0007	-.0010
Brich	0.2216	0.2798	0.3070	0.3361	0.3453	0.2979	-.0038	-.0019	-.0005	0.0003	0.0009	-.0009
MiniBatch K-means	0.1808	0.2013	0.2861	0.2737	0.2905	0.2464	**0.0047**	0.0024	0.0022	0.0027	0.0043	0.0032
DEC	0.3013	0.3923	0.4261	0.4340	0.4330	0.3973	0.0031	0.0041	0.0037	0.0038	0.0040	0.0037
RANMF	0.3010	0.3903	0.4161	0.4241	0.4230	0.3957	0.0030	0.0039	0.0035	0.0036	0.0037	0.0035
AE-NMF	0.1759	0.2885	0.3467	0.3554	0.3748	0.3082	0	0.0008	0.002	0.0033	0.0047	0.0021
SeaNMF	0.2063	0.3489	0.3929	0.3856	0.4179	0.3503	0.0002	0.0012	0.0027	0.004	0.005	0.0026
DGLCF	0.2071	0.3029	0.355	0.3635	0.3937	0.3244	0.0011	0.0014	0.0022	0.0038	**0.0052**	0.0027
DANMF-CRFR	0.2051	0.3031	0.4119	0.4296	0.4403	0.358	0.0023	0.0029	0.0035	0.0041	0.0042	0.0034
EDA-TEC	0.1906	0.2695	0.4285	0.4438	**0.4586**	0.3582	0.0058	0.0029	0.0025	0.0030	0.0040	0.0036
DCTMA	0.3064	**0.4061**	**0.4322**	**0.4457**	0.4427	**0.4066**	0.0045	**0.0043**	**0.0042**	0.0048	0.0051	**0.0043**

جدول ۶: عملکرد الگوریتم پیشنهادی روی مجموعه دادگان 20Newsgroup با معیارهای مختلف

		4	**8**	**12**	**16**	**20**
DCTMA	ACC	0.18725	0.32113	0.41733	0.45453	0.50398
	NMI	0.30689	0.40698	0.43344	0.44721	0.44457
	AMI	0.30640	0.40614	0.43225	0.44571	0.4427
	SC	0.0045	0.00435	0.00425	0.00484	0.0051
	ARI	0.12823	0.23543	0.25005	0.27565	0.27396

جدول ۷: نتایج خوشه‌بندی با الگوریتم‌های مختلف روی مجموعه دادگان WebKB برای ویژگی‌های مختلف

	AMI				SC			
تعداد ویژگی‌ها	**500**	**1000**	**2000**	میانگین	**500**	**1000**	**2000**	میانگین
K-means	0.3521	0.3554	0.3524	0.3533	0.0235	0.0177	**0.0138**	0.0183
NMF$_F$	0.3452	0.3623	0.3680	0.3585	0.0207	0.0153	0.0106	0.0155
NMF$_{KL}$	0.3744	0.3571	0.3684	0.3666	0.0205	0.0134	0.0106	0.0148
Hierarchical clustering	0.2531	0.2538	0.2747	0.2605	0.0156	0.0093	0.0089	0.0112
Brich	0.2601	0.2754	0.2660	0.2671	0.0155	0.0098	0.0087	0.0113
MiniBatch K-means	0.0337	0.2654	0.3377	0.1237	-0.0120	-0.0266	0.0120	-.0088
DEC	0.3541	0.3524	0.3644	0.3569	0.0120	0.0104	0.0105	0.0109
RANMF	0.3314	0.3222	0.3658	0.3398	0.0149	0.0141	0.0111	0.0133
AE-NMF	0.2656	0.2948	0.2811	0.2805	0.0193	0.0163	0.0116	0.0157
SeaNMF	0.31312	0.3491	0.3589	0.3403	0.0214	0.0172	**0.0139**	0.0175
DGLCF	0.3216	0.3213	0.3347	0.3258	0.0215	0.0164	0.0145	0.0174
DANMF-CRFR	0.3149	0.319	0.3269	0.3202	0.0203	0.0164	0.0137	0.0168
EDA-TEC	**0.4462**	**0.397**	0.3465	**0.3965**	0.0137	0.0122	0.0108	0.0122
DCTMA	0.3692	0.3652	**0.3736**	0.3693	**0.0259**	**0.0179**	0.0129	**0.0188**

جدول ۸: نتایج خوشه‌بندی با معیارهای مختلف روی مجموعه دادگان WebKB برای ویژگی‌های مختلف

		500	**1000**	**2000**
DCTMA	ACC	0.63515	0.6213	0.55680
	NMI	0.36975	0.3657	0.37416
	AMI	0.36924	0.3652	0.37365
	SC	0.0259	0.0179	0.01295
	ARI	0.30581	0.2916	0.30889

How to cite this article:

مراجع

.۱ Aghdam, M.H. and M.D. Zanjani, *A novel regularized asymmetric non-negative matrix factorization for text clustering.* Information Processing & Management, 2021. **58**(6): p. 102694.

.۲ Sun, B.-J., et al., *A Non-negative Symmetric Encoder-Decoder Approach for Community Detection*, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, Association for Computing Machinery: Singapore, Singapore. p. 597–606.

.۳ Shi, T., et al. *Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations*. in *Proceedings of the 2018 World Wide Web Conference*. 2018.

.٤ Daneshfar, F., et al., *Elastic deep autoencoder for text embedding clustering by an improved graph regularization.* Expert Systems with Applications, 2024. **238**: p. 121780.

.٥ Li, S., et al., *Semi-supervised bi-orthogonal constraints dual-graph regularized NMF for subspace clustering.* Applied Intelligence, 2022. **52**(3): p. 3227-3248.

.٦ Salahian, N., et al., *Deep autoencoder-like NMF with contrastive regularization and feature relationship preservation.* Expert Systems with Applications, 2023. **214**: p. 119051.

.۷ Wang, S., et al., *Extreme clustering – A clustering method via density extreme points.* Information Sciences, 2021. **542**: p. 24.۳۹-

.۸ Guan, R., et al., *Deep feature-based text clustering and its explanation.* IEEE Transactions on Knowledge and Data Engineering, 2020. **34**(8): p. 3669-3680.

.۹ Diallo, B., et al., *Deep embedding clustering based on contractive autoencoder*. Neurocomputing, 2021. **433**: p. 96-107.

.۱۰ Settipalli, L., G. Gangadharan, and U. Fiore, *Predictive and adaptive drift analysis on decomposed healthcare claims using ART based topological clustering.* Information Processing & Management, 2022. **59**(3): p. 102887.

.۱۱ Hosseini, S. and Z.A. Varzaneh, *Deep text clustering using stacked AutoEncoder.* Multimedia Tools and Applications, 2022. **81**(8): p. 10861-10881.

.۱۲ Ren, Z., W. Zhang, and Z. Zhang, *A deep nonnegative matrix factorization approach via autoencoder for nonlinear fault detection.* IEEE Transactions on Industrial Informatics, 2019. **16**(8): p. 5042-5052.

.۱۳ Behera, G. and N. Nain, *DeepNNMF: deep nonlinear non-negative matrix factorization to address sparsity problem of collaborative recommender system.* International Journal of Information Technology, 2022: p. 1-9.

.۱٤ Wang, J. and X.-L. Zhang, *Deep nmf topic modeling.* Neurocomputing, 2022.

.۱٥ Veiga Simão, A.M., et al., *Prosociality in cyberspace: Developing emotion and behavioral regulation to decrease aggressive communication.* Cognitive Computation, 2021. **13**(3): p. 736-750.

.۱٦ Jiang, Z., et al., *Variational deep embedding: An unsupervised and generative approach to clustering.* arXiv preprint arXiv:1611.05148, 2016.

.۱۷ Curiskis, S.A., et al., *An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit.* Information Processing & Management, 2020. **57**(2): p. 102034.

.۱۸ Diallo, B., et al., *Multi-view document clustering based on geometrical similarity measurement.* International Journal of Machine Learning and Cybernetics, 2022. **13**(3): p. 663-675.

.۱۹ Śmieja, M., Ł. Struski, and M.A. Figueiredo, *A classification-based approach to semi-supervised clustering with pairwise constraints.* Neural Networks, 2020. **127**: p. 193-20.۳

.۲۰ Fu, B., et al., *Anomaly Aware Symmetric Non-negative Matrix Factorization for Short Text Clustering.* 2022.

.۲۱ Sheng, W. and J. Lipor. *A Novel Framework for Deep Learning from Pairwise Constraints*. in *2020 54th Asilomar Conference on Signals, Systems, and Computers*. 2020. IEEE.

.۲۲ Guan, R., et al., *Deep feature-based text clustering and its explanation.* IEEE Transactions on Knowledge and Data Engineering, 2020.

.۲۳ Revathy, V., A.S. Pillai, and F. Daneshfar, *LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model.* Procedia Computer Science, 2023. **218**: p. 1196-1208.

.۲٤ Fard, M.M., T. Thonet, and E. Gaussier. *Pairwise-Constrained Deep Document Clustering*. in *International Conference on Reliability and Statistics in Transportation and Communication*. 2019. Springer.

.۲۵ Wei, F., et al., *Semi-Supervised Clustering with Contrastive Learning for Discovering New Intents.* arXiv preprint arXiv:2201.07604, 2022.

.۲۶ Daneshfar, F., et al., *A survey on semi-supervised graph clustering.* Engineering Applications of Artificial Intelligence, 2024. **133**: p. 108215.

.۲۷ Vilhagra, L.A., E.R. Fernandes, and B.M. Nogueira. *Textcsn: a semi-supervised approach for text clustering using pairwise constraints and convolutional siamese network.* in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 2020.

.۲۸ Berahmand, K., et al., *Autoencoders and their applications in machine learning: a survey.* Artificial Intelligence Review, 2024. **57**(2): p. 28.

.۲۹ Yang, Y., Q.J. Wu, and Y .Wang, *Autoencoder with invertible functions for dimension reduction and image reconstruction.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016. **48**(7): p. 1065-1079.

.۳۰ Lakshmi, R. and S. Baskar, *Efficient text document clustering with new similarity measures.* International Journal of Business Intelligence and Data Mining, 2021. **18**(1): p. 49-72.

.۳۱ Oghbaie, M. and M. Mohammadi Zanjireh, *Pairwise document similarity measure based on present term set.* Journal of Big Data, 2018. **5**(1): p. 1.۲۳-

.۳۲ Jin, D., et al., *A survey of community detection approaches: From statistical modeling to deep learning.* IEEE Transactions on Knowledge and Data Engineering, 2021.

.۳۳ Ahmad, A. and S.S. Khan, *Survey of state-of-the-art mixed data clustering algorithms.* Ieee Access, 2019. **7**: p. 31883-31902.

.۳۴ Su, X., et al., *A comprehensive survey on community detection with deep learning.* IEEE Transactions on Neural Networks and Learning Systems, 2022.

.۳۵ Golzari Oskouei, A., M.A. Balafar, and C. Motamed, *EDCWRN: efficient deep clustering with the weight of representations and the help of neighbors.* Applied Intelligence, 2022: p. 1-23.

.۳۶ Chen, L. and Z. Zhong, *Adaptive and structured graph learning for semi-supervised clustering.* Information Processing & Management, 2022. **59**(4): p. 102949.

.۳۷ Lee, D. and H.S. Seung, *Algorithms for non-negative matrix factorization.* Advances in neural information processing systems, 2000. **13**.

.۳۸ Misztal-Radecka, J. and B. Indurkhya, *Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems.* Information Processing & Management, 2021. **58**(3): p. 102519.

.۳۹ Zhang, T., R. Ramakrishnan, and M. Livny, *BIRCH: an efficient data clustering method for very large databases.* ACM sigmod record, 1996. **25**(2): p. 103-114.

.۴۰ Béjar Alonso, J., *K-means vs Mini Batch K-means: a comparison.* 2013.

.۴۱ Ren, Y., et al., *Semi-supervised deep embedded clustering.* Neurocomputing, 2019. **325**: p. 121-130.

.۴۲ Yang, S., G. Huang, and B. Cai, *Discovering topic representative terms for short text clustering.* IEEE Access, 2019. **7**: p. 92037-92047.

.۴۳ Li, W. and E. Suzuki, *Adaptive and hybrid context-aware fine-grained word sense disambiguation in topic modeling based document representation.* Information Processing & Management, 2021. **58**(4): p. 102592.

.۴۴ Yang, Y., *Chapter 3 - Temporal Data Clustering*, in *Temporal Data Mining Via Unsupervised Ensemble Learning*, Y. Yang, Editor. 2017, Elsevier. p. 19-34.

.۴۵ Hu, D., D. Feng, and Y. Xie, *EGC: A novel event-oriented graph clustering framework for social media text.* Information Processing & Management, 2022. **59**(6): p. 103059.

.۴۶ Wang, R., et al., *Trio-based collaborative multi-view graph clustering with multiple constraints.* Information Processing & Management, 2021. **58**(3): p. 102466.

.۴۷ Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information processing & management, 1988. **24**(5): p. 513-523.

[1] Nonnegative matrix factorization

[2] Deep Text Clustering based on a local Manifold in the Autoencoder
[3] Embedding space
[4] Representation learning
[5] Consensus similarity matrix
[6] Loss function
[7] Stacked autoencoder
[8] Time-frequency-based similarity measure
[9] Term-frequency
[10] Triangle's area similarity-sector's area similarity measure
[11] Pairwise document similarity measure
[12] Reconstruction error
[13] Kullback-leibler distribution
[14] Deep embedding clustering
[15] Regularized asymmetric non-negative matrix factorization
[16] Balanced Iterative Reducing and Clustering using Hierarchies
[17] Term frequency-Inverse document frequency
[18] Normalized mutual information
[19] Adjusted mutual information
[20] Silhouette coefficient