A Systematic Review of K-means Algorithm Improvement Research for Data Clustering

Elham Yalveh
M. Sc. Candidate in Knowledge and Information Science; University of Qom; Qom, Iran Email: elham.yalveh2018@gmail.com

Yaghoub Norouzi*
PhD in Knowledge and Information Science; Associate Professor; Department of Knowledge and Information science; University of Qom; Qom, Iran Email: ynorouzi@gmail.com

Ashkan Khatir
PhD in Information Technology Engineering; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: khatir@students.irandoc.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 2 | pp. 527-556
Winter 2022
https://doi.org/10.52547/jipm.37.2.527

Received: 11, Jan. 2021 **Accepted: 26, May 2021**

Abstract: Clustering as a process to understand the nature and structure of data plays an important role in organizing data in many areas of science and technology. One of the most widely used and simple algorithms for clustering is K-means. The present study was conducted to systematically reviewing research on improving K-means algorithm on data clustering. This research examines the researches conducted in this field and its role in organizing data in the range of 2010 to 2020 with a new strategy based on the shortcomings of the K-means algorithm. For this purpose, the amount of attention of researchers to eliminate any of the shortcomings of this algorithm in order to improve it in recent years has been compiled in the form of research questions. In this study, with the use of a search strategy for refining and extracting articles, 47 related sources were identified and examined. Findings showed that most researches have been done by overcoming the sensitive shortcomings to initial cluster centers to improve the K-means algorithm. Also, out of a total of 47 studies, the improved K-means algorithm has been applied in 35 studies on non-textual data and in 12 studies on textual data. Finally, the results of a review of six studies showed that the amount of data is directly related to the performance of improved K-means algorithm. In other words, this algorithm must be modified in such a way as to perform efficient and accurate clustering by applying it to different amounts of data.

Keywords: Data Clustering, K-means Algorithm, Clustering Improvement, Systematic Review

* Corresponding Author

مروری نظام‌مند بر پژوهش‌های بهبود الگوریتم کا-میانه برای خوشه‌بندی داده‌ها

الهام یلوه
دانشجوی کارشناسی ارشد علم اطلاعات و دانش‌شناسی؛ دانشگاه قم؛ قم، ایران؛
elham.yalveh2018@gmail.com

یعقوب نوروزی
دکتری؛ علم اطلاعات و دانش‌شناسی؛ دانشیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه قم؛ قم، ایران؛ پدیدآور رابط ynorouzi@gmail.com

اشکان خطیر
دکتری؛ مهندسی فناوری اطلاعات؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
khatir@students.irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۱۰/۲۲ | **پذیرش:** ۱۴۰۰/۰۳/۰۵ | مقاله برای اصلاح به مدت ۱۵ روز نزد پدیدآوران بوده است.

چکیده: خوشه‌بندی به‌عنوان یک فرایند جهت شناخت ماهیت و ساختار داده‌ها در بسیاری از حوزه‌های علوم و فناوری‌های مرتبط با آن نقش مهمی در سازماندهی داده‌ها دارد. یکی از الگوریتم‌های پرکاربرد و سادهٔ خوشه‌بندی، کا-میانه است. پژوهش حاضر با هدف مرور نظام‌مند تحقیقات در زمینهٔ بهبود الگوریتم کا-میانه برای خوشه‌بندی داده‌ها صورت گرفته است. این پژوهش با یک راهبرد جدید بر مبنای کاستی‌های الگوریتم کا-میانه به بررسی تحقیقات انجام‌شده در این زمینه و نقش آن در سازماندهی داده‌ها در محدودهٔ سال‌های ۲۰۱۰ تا ۲۰۲۰ می‌پردازد. برای این منظور میزان توجه پژوهشگران به رفع هر یک از کاستی‌های این الگوریتم برای بهبود طی سال‌های مزبور در قالب پرسش‌های پژوهش تدوین شده است. در این پژوهش با استفاده از استراتژی جست‌وجو، پالایش، و استخراج مقاله‌ها در نهایت، ۴۷ منبع مرتبط شناسایی و مورد بررسی قرار گرفت. یافته‌ها نشان داد که بیشترین تحقیقات صورت‌گرفته با غلبه بر کاستی حساس به مراکز خوشهٔ اولیه در جهت بهبود الگوریتم کا-میانه انجام شده است. همچنین، از ۴۷ تحقیق مورد بررسی، الگوریتم بهبودیافتهٔ کا-میانه در ۳۵ تحقیق بر روی داده‌های غیرمتنی و در ۱۲ تحقیق بر روی داده‌های متنی اعمال شده است. سرانجام، نتیجهٔ حاصل از بررسی ۶ تحقیق از تحقیقات صورت‌گرفته نشان داد که حجم داده‌ها رابطه‌ای مستقیم با عملکرد الگوریتم بهبودیافتهٔ کا-میانه

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۳۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۷ | شمارهٔ ۲ | ص‌ص ۵۲۷-۵۵۶
زمستان ۱۴۰۰
https://doi.org/10.52547/jipm.37.2.527

دارد. به‌عبارت دیگر، این الگوریتم باید به‌نوعی اصلاح شود که با اعمال بر روی حجم متفاوت داده‌ها خوشه‌بندی کارآمد و دقیقی انجام دهد.

کلیدواژه‌ها: خوشه‌بندی داده، بهبود الگوریتم کا-میانه، خوشه‌بندی، مرور نظام‌مند

۱. مقدمه

خوشه‌بندی[1] از جملهٔ فنون داده‌کاوی برای تحلیل داده‌هاست که دو هدف را دنبال می‌کند: ۱) داده‌های هر خوشه تا حد ممکن شبیه به‌هم باشند؛ به‌عبارت دیگر، شباهت درون‌خوشه‌ای بالا باشد، و ۲) داده‌های هر خوشه از داده‌های خوشه‌های دیگر متفاوت باشند؛ یعنی شباهت بین خوشه‌ای پایین باشد؛ به معنای دیگر، هر خوشه دارای داده‌های متفاوتی از خوشه‌های دیگر باشد (Mann & Kaur 2013). پژوهش‌هایی که در این حوزه انجام می‌شود، در راستای بهبود هرچه بهتر این فرایند بر روی داده‌هاست. از آنجا که در سال‌های اخیر الگوریتم‌های خوشه‌بندی مورد توجه بسیاری از پژوهشگران قرار گرفته، پژوهش و شناسایی الگوریتم‌های تجزیه و تحلیل خوشه‌بندی به‌طور عمده در دو بعد بهبود الگوریتم‌های خوشه‌بندی سنتی و ارائهٔ مفاهیم الگوریتم جدید ظاهر می‌شود. بنابراین، با بررسی معایب الگوریتم‌های خوشه‌بندی سنتی تحقیقاتی انجام می‌شود و روش‌هایی برای بهبود این الگوریتم‌ها پیشنهاد می‌گردد. تعداد قابل توجهی از پژوهش‌هایی که انجام می‌شود این جنبه را دنبال می‌کنند. با توجه به اینکه الگوریتم‌های خوشه‌بندی اصلی دارای کاستی‌هایی در جریان فرایند خوشه‌بندی هستند، همواره پژوهش‌های متعددی در جهت رفع این کاستی‌ها با هدف بهبود و اصلاح آن‌ها انجام می‌شود.

یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی که در سال ۱۹۶۷ توسط «مک کوئین»[2] ارائه شد، الگوریتم کا-میانه است که یک روش تجزیه و تحلیل خوشه‌ای مبتنی بر افراز است (Aggarwal 2004). علت استفادهٔ گسترده از این الگوریتم سادگی، گروه‌بندی سریع، و کارآمدی آن است. الگوریتم‌های داده‌کاوی[3] باید مقیاس‌پذیری خوبی داشته باشند تا به‌طور مؤثر اطلاعات را از داده‌های حجیم استخراج کنند (JiaweiHan 2005). این الگوریتم مقیاس‌پذیری بالایی دارد و هنگامی که با مجموعهٔ داده‌های بزرگ سروکار دارد،

1. clustering
2. MacQueen
3. data mining

به‌سرعت همگرا می‌شود (Iezzi 2012). با توجه به اینکه تحلیلگر گارتنر[1] عنوان می‌کند که امروزه بیش از ۸۰ درصد داده‌های تولیدشده بدون ساختار و عمدتاً متن هستند (Afzali & Kumar 2019)، توجه به این امر در خوشه‌بندی داده‌های متنی که بیش از دوسوم از داده‌های تولیدشده توسط سازمان‌ها، شبکه‌های اجتماعی، سرویس‌دهنده‌های ایمیل، کتابخانه‌های دیجیتال، و از همه مهم‌تر توسط وب در هر ثانیه در حال رشد است، قابل تأمل است (Han, Kamber & Pei 2012). با این حال، این الگوریتم دارای کاستی‌ها و محدودیت‌هایی است که برای رفع آن‌ها و دستیابی به یک خوشه‌بندی مناسب، پژوهش‌ها و تحقیقاتی در جهت بهبود این الگوریتم انجام شده است. این بدان جهت است که در دنیای واقعی، مجموعهٔ داده‌ها که نیاز به پردازش دارند، همیشه استاندارد نیستند و شکل خوشه‌های آن‌ها ثابت نیست. از این رو، برای پیشنهاد یک روش خوشه‌بندی مناسب برای مجموعه‌های داده، تحقیقات بیشتری مورد نیاز است. تحقیقات بین‌المللی راهبردهای متفاوتی مانند مبنا قرار دادن گام‌های اجرایی را برای رفع کاستی‌های موجود در الگوریتم کا-میانه مد نظر قرار داده‌اند؛ مانند تحقیق «خاندار و الوی» که برخی منابع را که به بهبود الگوریتم کا-میانه پرداخته‌اند، مورد مطالعه قرار داده‌اند. آن‌ها کاستی‌های آن‌ها را خلاصه کرده و به مطالعهٔ فاصله، اعتبار و معیارهای پایداری این الگوریتم پرداختند (Khandare and Alvi 2016). همچنین، در پژوهشی دیگر «بن‌عبدالله، بن‌غبریت و بوهادو» با هدف یافتن الگوریتم‌های مناسب برای مجموعهٔ داده‌های پراکندهٔ صنعتی، به مقایسهٔ الگوریتم‌های خوشه‌بندی عمومی، انعطاف‌پذیر و قابل استفاده در حوزهٔ صنعت پرداختند (Benabdellah, Benghabrit, & Bouhaddou, 2019). پژوهش حاضر با یک راهبرد جدید بر مبنای کاستی‌های الگوریتم کا-میانه به بررسی تحقیقات انجام‌شده در این زمینه و نقش آن در سازماندهی داده‌ها در محدودهٔ سال‌های ۲۰۱۰ تا ۲۰۲۰ می‌پردازد. این امر با تکیه بر کاستی‌های این الگوریتم به‌عنوان وجه تمایز آن با سایر پژوهش‌های مروری در این حوزه در قالب مروری نظام‌مند انجام می‌شود. یافته‌های پژوهش برای محققان و دانش‌پژوهان این امکان را فراهم می‌آورد که با مطالعهٔ تحقیقات انجام‌شده، وارد چرخهٔ تکراری پژوهش نشوند و در عین حال، برای رسیدن به یک نگرش تازه جهت ارائهٔ یک روش بهتر در راستای پژوهش‌های قبلی گام‌های مؤثرتری بردارند. برای این منظور، میزان توجه پژوهشگران به

1. Gartner

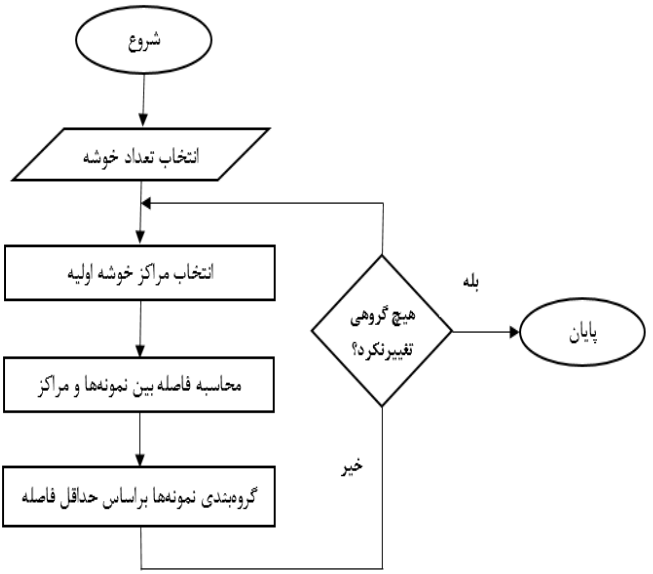
رفع هر یک از کاستی‌های این الگوریتم در جهت بهبود آن طی سال‌های اخیر در قالب سه پرسش تدوین شد:

۱. در تحقیقات صورت‌گرفته کدام‌یک از کاستی‌های الگوریتم کا-میانه بیشتر مورد توجه بوده است؟
۲. در تحقیقات صورت‌گرفته توجه به داده‌های متنی و غیرمتنی به چه میزان بوده است؟
۳. حجم داده‌ها در تحقیقات صورت‌گرفته جهت بهبود الگوریتم کا-میانه در فرایند خوشه‌بندی تا چه حد تأثیرگذار بوده است؟

۲. الگوریتم کا-میانهٔ اصلی

این الگوریتم از جمله متداول‌ترین و ساده‌ترین الگوریتم‌های خوشه‌بندی محسوب می‌شود که مجموعه‌ای از «اشیا داده‌ای»[1] را به تعداد معینی خوشه تقسیم می‌کند. الگوریتم کا-میانه در ابتدا یک مجموعهٔ تصادفی از K خوشه را از اشیا داده تولید می‌کند، و K تعداد نقاط را به‌طور تصادفی انتخاب می‌کند که هر داده با استفاده از تابع شباهت مانند (فاصلهٔ اقلیدسی، فاصلهٔ منهتن، فاصلهٔ ماهانولوبیس[2] به نزدیک‌ترین خوشه تخصیص داده می‌شود (Goswami 2015). پس از اتمام این مرحله که همهٔ اشیای داده گروه‌بندی می‌شوند، برخی از اشیای داده ممکن است از یک خوشه به خوشهٔ دیگر منتقل شوند. سپس، مرکز اولیهٔ هر خوشه بر اساس اشیای دادهٔ جدید در آن خوشه به‌روزرسانی می‌شود تا زمانی که حرکت اشیای داده بین خوشه‌ها متوقف شده و همگرایی برآورده شود (Goswami 2015; Yedla, Pathakota & Srinivasa 2010). شکل ۱، فرایند این الگوریتم را نمایش می‌دهد. الگوریتم کا-میانه همچون سایر الگوریتم‌های خوشه‌بندی، در کنار داشتن نقاط قوت مختص به‌خود کاستی‌هایی نیز دارد. از مزایای این الگوریتم می‌توان به سادگی، آسان بودن قابلیت پیاده‌سازی، سرعت بالا، و مناسب بودن برای مجموعهٔ داده‌های بزرگ اشاره کرد (Fränti & Sieranoja 2019; Saklecha & Raikwal 2017). لزوم تعیین تعداد خوشه، حساس بودن به داده‌های نویزی و دورافتاده، وابستگی نتایج نهایی به مقداردهی مراکز اولیه و تعداد خوشه‌ها، گیر افتادن الگوریتم در بهینهٔ محلی[3] و همگرایی زودرس، و حساس بودن به ابعاد بالای ویژگی (Wang & Su 2011) نیز چالش‌های مورد بحث و بررسی توسط محققان و پژوهشگران است.

1. data objects
2. Mahalanobis
3. local optimum



شکل ۱. فرایند الگوریتم کا-میان Z اصلی (Awawdeh, Edinat & Sleit 2019)

۳. روش پژوهش

مرور نظام‌مند در اصل، برای پاسخگویی به یک سؤالِ پژوهش مبتنی بر ارزیابی بدون سوگیریِ همهٔ مطالعات پژوهشیِ مربوط به آن سؤال طراحی می‌شود. مرور نظام‌مند می‌تواند موجب رفع ابهام در یک موضوع، ایجاد دیدگاه‌های جدید با استفاده از ترکیب نتایج به‌دست‌آمده از مطالعات مختلف، و کاهش تأثیر هرگونه نقصان یا خطا در یک پژوهش خاص شود (Strech & Sofaer 2012). برای مرور نظام‌مند فرایندهای نسبتاً مشابهی توسط نویسندگان مختلف عنوان شده است که به‌طور معمول، از نظر تعداد یا عنوان مراحلی که پیشنهاد شده، از هم متفاوت هستند. برخی از آن مراحل ممکن است در هم ادغام شده باشند. بر همین اساس، در این پژوهش یک ساختار در چهار فاز اصلی که به‌صورت ادغام فرایندهای متداول یک مرور نظام‌مند است، در شکل ۲، ارائه شده است.

فاز اول: فرایند اولیه

۱. طرح سؤال تحقیق

۲. تعیین معیارهای لازم جهت انتخاب پژوهش‌های مرتبط

فازدوم: استراتژی جست‌وجو برای شناسایی مطالعات مرتبط

۱. جست‌وجوی نظام‌مند بر اساس معیارهای تعریف‌شده در فاز اول

۲. شناسایی پایگاه‌های اطلاعاتی متناسب با زمینهٔ پژوهش

۳. انتخاب کلیدواژه‌های متناسب با موضوع پژوهش و معیارهای انتخابی در فاز ۱

۴. گزینش مقاله‌های متناسب با موضوع پژوهش با توجه به معیارهای انتخابی و تأکید بر ذکر واژه‌هایی که به‌معنای بهبوددهنده هستند.

فاز سوم: پالایش و استخراج مقاله‌های مستخرج از فاز ۲

۱. ذکر یکی از کلیدواژه‌هایی که به‌معنای بهبوددهندهٔ الگوریتم کا-میانه است در عنوان مقاله

۲. مطالعه و بررسی چکیده، یافته‌های پژوهش، ارزیابی پژوهش و نتیجه‌گیری

فاز چهارم: یافته‌ها، تجزیه و تحلیل، پاسخ به پرسش‌های پژوهش

شکل ۲. روش اجرای مرور نظام‌مند در پژوهش حاضر

در این پژوهش در فاز اول از معیارهای پیشنهادی توسط Okoli & Schabram (2010) و Fink (2013) استفاده شده است. این معیارها شامل محتوا، طرح تحقیق، زمینه، زبان، تاریخ انتشار و نوع سند است. با توجه به معیارهای اشاره شده، تحقیقاتی گزینش شدند که محتوای آن‌ها با هدف پژوهش حاضر منطبق و به‌عنوان یک پژوهشی در زمینهٔ علوم مهندسی، علوم پزشکی، علوم اجتماعی، علوم گردشگری به زبان انگلیسی بین سال‌های ۲۰۱۰ تا ۲۰۲۰ در نشریات و کنفرانس‌های علمی ارائه شده بودند.

در فاز دوم با هدف جست‌وجوی نظام‌مند جهت شناسایی مقالات مرتبط با در نظر گرفتن معیارهای از پیش تعریف‌شده در فاز اول، پایگاه‌های IEEE، Springer، Science Direct، ACM Digital Library جهت فرایند جست‌وجو انتخاب شدند. علاوه بر این، جهت اطمینان

از دستیابی کامل به مقالات مرتبط، جست‌وجو در پایگاه‌های Scopus و Google Scholar نیز انجام شد. سپس، جست‌وجو در این پایگاه‌ها با ترکیب کلیدواژه‌های Improved، improvement، improving، enhancement، enhanced، new approach، text document، clustering، text mining با کلیدواژهٔ k-means صورت گرفت. برای گزینش مقالات متناسب با هدف[1] پژوهش با بازگشت به فاز ۱ و توجه به معیارهای انتخاب شده، عنوان مقاله و واژگان کلیدی در مقاله معرف نوع کار پژوهشی بود، ملاک گزینش مقاله‌های انتخابی قرار گرفت. در فاز سوم، برای غربالگری و رسیدن به مرتبط‌ترین مقاله‌های کاوش‌شده در فاز ۲، این نکته در نظر گرفته شد که در عنوان مقاله حتماً یکی از کلیدواژه‌هایی باشد که به معنای بهبوددهندهٔ الگوریتم کا-میانه در جست‌وجوی مقاله‌ها مورد استفاده قرار گرفته است و با مطالعه و بررسی چکیده، یافته‌های پژوهش، ارزیابی پژوهش و نتیجه‌گیری تعدادی از مقاله‌ها از چرخهٔ انتخاب خارج شدند. در نهایت، ۴۷ مقاله انتخاب شدند. در فاز چهارم به یافته‌ها، تجزیه و تحلیل، و پاسخ به پرسش‌های پژوهش پرداخته شده است.

٤. یافته‌ها

شناخت و آگاهی از شکاف‌های پژوهشی موجب می‌شود که محققان و پژوهشگران با اشراف بیشتری به پژوهش بپردازند. به همین دلیل، پژوهش حاضر در قالب مروری نظام‌مند، تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه برای خوشه‌بندی داده‌ها را مورد توجه قرار داد. تحقیقات مورد بررسی، الگوریتم کا-میانه را برای رفع یک یا دو کاستی از کاستی‌های آن بهبود دادند. بنابراین، سعی بر آن شد که این تحقیقات بر مبنای این کاستی‌ها مورد بررسی و تحلیل قرار گیرند.

پرسش اول پژوهش: در تحقیقات صورت‌گرفته کدام‌یک از کاستی‌های الگوریتم کا-میانه بیشتر مورد توجه بوده است؟

در پژوهش حاضر ۴۷ منبع منتخب در قالب جداول ۱ تا ۶ آمده است. نحوهٔ تقسیم‌بندی این منابع در جداول بر اساس کاستی‌های مورد توجه در پژوهش‌ها بوده است. در ادامه، در جدول ۱، تحقیقات انجام‌شده با هدف غلبه بر کاستی حساس به مراکز خوشهٔ اولیه ارائه شده است.

---

۱. با توجه به اینکه در مرحلهٔ فیلترینگ منابع مورد پژوهش، اساس فیلتر نوع بهبود الگوریتم کا-میانه در حوزهٔ داده‌کاوی مد نظر بوده است، مقالات مرتبط با هدف پژوهش برگزیده شدند.

جدول ۱. تحقیقات انجام‌شده با هدف غلبه بر کاستی حساس به مراکز خوشهٔ اولیه[1]

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Na, Xumin and Yong (2010)	دو ساختار داده ساده برای حفظ برچسب‌های خوشه و فاصلهٔ همهٔ داده‌های هدف تا نزدیک‌ترین خوشه در طول هر تکرار که می‌تواند در تکرار بعدی استفاده شود، محاسبه می‌شود.	روش بهبودیافته می‌تواند به‌طور مؤثر سرعت خوشه‌بندی و دقت را بهبود بخشد و پیچیدگی محاسباتی الگوریتم کا-میانه را کاهش دهد.
Yedla, Pathakota & Srinivasa (2010)	در الگوریتم پیشنهادی مجموعهٔ داده‌های حاوی ویژگی‌های منفی بررسی شده و سپس، در مجموعهٔ داده شامل ویژگی‌های منفی، تمام نقاط داده در مجموعه با کم کردن هر ویژگی نقطه داده با حداقل مقدار مشخصه در مجموعهٔ داده به مثبت تبدیل شده و ادامهٔ فرایند جهت خوشه‌بندی صورت گرفته است.	الگوریتم پیشنهادی در مقایسه با الگوریتم کا-میانه رایج دقیق و کارآمدتر شده و دارای دقت بالاتری نسبت به آن با زمان محاسباتی کمتر است.
Karegowda et al. (2013)	استفاده از دو روش الگوریتم ژنتیک[1] (GA) و خوشه‌بندی فازی مبتنی بر آنتروپی[2] (EFC) جهت انتخاب مراکز خوشهٔ اولیه	کاهش در خطای طبقه‌بندی و زمان اجرای کا-میانه توسط الگوریتم پیشنهادی در مقایسه با الگوریتم کا-میانه و در نتیجهٔ بهبود عملکرد خوشه‌بندی
Chaturvedi and Rajavat (2013)	در دو فاز به بهبود الگوریتم کا-میانه می‌پردازد؛ به این صورت که در فاز اول به‌طور نظام‌مند مراکز اولیه را تعیین می‌کند، و در فاز دوم از توابع روش خوشه‌بندی استفاده می‌کند.	الگوریتم پیشنهادی نتایج بهتری را برای کاهش زمان محاسباتی و افزایش دقت در مقایسه با الگوریتم پایهٔ کا-میانه به‌دست می‌دهد، بنابراین، برای پویش پایگاه داده‌های بزرگ مفید است.
Shunye (2013)	یک الگوریتم خوشه‌بندی کا-میانه اصلاح‌شده با نامگذاری IKCBD بر اساس عدم تشابه برای اندازه‌گیری شباهت بین هر یک از داده‌ها استفاده می‌کند و از درخت هافمن برای یافتن مراکز اولیه استفاده می‌کند که از ماتریس عدم تشابه برای ساخت استفاده می‌کند.	الگوریتم پیشنهادی در مقایسه با روش‌های سنتی، سرعت و نتایج بهتری دارد. بنابراین، برای مجموعهٔ دادهٔ بزرگ و مجموعهٔ دادهٔ سطح بالا مناسب‌تر است.
Jaganathan and Jaiganesh (2013)	پیشنهاد یک الگوریتم بهبودیافته با نام‌گذاری PSOK که روش ترکیبی جدیدی را با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات[3] (PSO) با الگوریتم بهبودیافته کا-میانه برای خوشه‌بندی اسناد معرفی می‌کند.	روش پیشنهادی نتایج بهتری را در مقایسه با روش‌های دیگر مورد مقایسه تولید می‌کند.

1. genetic algorithm
2. entropy based fuzzy clustering
3. particle swarm optimization

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Ma (2014)	ابتدا، در مدل محاسبهٔ تشابه خوشه‌بندی از یک الگوریتم شباهت معنایی جامع و یکپارچه استفاده شده و فاکتورهای زمینه‌ای و معنایی متن در هر مرحلهٔ محاسباتی ترکیب شده است. و پس از آن الگوریتم کا-میانه اصلاح شده است که از استراتژی اولویت برای تقسیم داده‌ها در ابتدا استفاده شده است.	الگوریتم پیشنهادی نه‌تنها می‌تواند دقت خوشه‌بندی را بهبود بخشد، بلکه پایداری بسیار بالایی نیز دارد.
Goyal and Kumar (2014)	الگوریتم پیشنهادی روشی برای انتخاب سیستماتیک[1] مرکز خوشهٔ اولیه ارائه می‌دهد. ابتدا، نقاط داده[2] در یک فضای دو بعدی رسم می‌شوند. تمام نقاط داده باید دارای ویژگی‌های مثبت باشند. اگر چنین نباشد، ابتدا خصوصیت ارزش منفی باید با تفریق کردن هر خصوصیت نقطه‌ای با حداقل مقدار خصوصیت در مجموعهٔ داده به مثبت تبدیل شود.	الگوریتم پیشنهادی می‌تواند برای انواع مختلف مجموعهٔ داده‌ها کاربرد داشته باشد. مسائل مرتبط با توزیع یکنواخت و همچنین، توزیع غیریکنواخت نقاط داده، بهتر مورد توجه قرار می‌گیرند. همچنین، این الگوریتم تعداد تکرار مورد نیاز برای رسیدن به معیارهای همگرایی را تا حد زیادی کاهش می‌دهد.
Yadav and Singh (2016)	الگوریتم پیشنهادی در دو فاز تخصیص اولیهٔ داده‌ها به نزدیک‌ترین خوشه و تخصیص مجدد اشیا داده انجام می‌شود.	الگوریتم پیشنهادی ارائه‌شده خوشه‌بندی داده‌ها را با حذف خوشه‌های خالی بهبود می‌دهد، کاهش زمان محاسباتی الگوریتم را به همراه دارد، و در نهایت، دقت و کارایی الگوریتم را بهبود می بخشد .
Kant and Ansari (2016)	روش انتخاب مرکز برای الگوریتم کا-میانه را با استفاده از شاخص اتکینسون[3] (AI) همراه با فاصله اقلیدوسی معرفی می‌کند.	خوشه‌بندی در الگوریتم پیشنهادی در مقایسه با الگوریتم کا-میانه دقیق‌تر است.
Xiong et al. (2016)	در روش الگوریتم اصلاح‌شدهٔ پیشنهادی پارامتر چگالی همهٔ اشیا-داده در مجموعهٔ داده را محاسبه و داده‌های دورافتاده را مشخص می‌کند. اگر یک دادهٔ دورافتاده جداسازی شود، از مجموعهٔ داده‌ها حذف خواهد شد.	نتایج آزمایش‌ها نشان می‌دهد که الگوریتم پیشنهادی می‌تواند پایداری و دقت خوشه‌بندی را بهبود دهد.
Vashist and Nath (2016)	تعدادی خوشه برای هر سند متنی بر اساس تولید مرکز ثابت جمع‌آوری می‌شود و تنها کلمات منحصربه‌فرد را از اسناد مختلف جمع‌آوری می‌کند و از معیار شباهت cosine برای قرار دادن اسناد مشابه در خوشه‌های مناسب استفاده می‌کند.	دقت الگوریتم پیشنهادی در مقایسه با الگوریتم موجود از لحاظ معیار انداه‌گیری F، فراخوانی، دقت و پیچیدگی زمانی بالا است.

1. systematically
2. data points
3. Atkinson index

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Choudhary, Sharma and Singh (2016)	روش الگوریتم پیشنهادی مرتب‌سازی نقاط و سپس، تقسیم آن‌ها به k گروه است. به‌جای انجام دسته‌بندی روی تنها یک یا چند خصوصیت، این کار می‌تواند روی کل مجموعهٔ داده‌ها انجام شود. روش پشنهادی ترکیبی از مقداردهی اولیه و نرمال‌سازی مقادیر داده برای بهبود عملکرد الگوریتم است.	کارایی و دقت الگوریتم پیشنهادی از طریق چندین آزمایش اثبات شده و مقایسهٔ آن با الگوریتم استاندارد کا-میانه و الگوریتم‌های مورد نظر نشان از بهبود عملکرد آن دارد.
Raval Unnati and Chaita (2016)	الگوریتم پیشنهادی مبتنی بر دو فاز استخراج مراکز اولیه و تخصیص داده‌ها به نزدیک‌ترین خوشه است.	الگوریتم پیشنهادی بهبودیافته، سرعت و دقت خوشه‌بندی را بهبود بخشیده و پیچیدگی زمانی را کاهش داده است.
Saklecha and Raikwal (2017)	الگوریتم پیشنهادی دو مرحله را برای تعیین مرکز اولیه در نظر می‌گیرد و نقاط داده را به نزدیک‌ترین مرکز ارائه می‌دهد تا دقت و کارایی الگوریتم را بهبود بخشد.	دقت و کارایی در الگوریتم اصلاح‌شده بالاتر از الگوریتم استاندارد کا-میانه است.
Linyao and Jianguo (2018)	در روش الگوریتم پیشنهادی ابتدا دو نقطه دور از نقاط نمونه به‌عنوان نقطهٔ مرکزی اولیه مشخص می‌شود و سپس، نقاط دیگر به خوشه که نزدیک‌ترین نقطه مرکزی به آن تعلق دارد، تقسیم می‌شوند.	الگوریتم پیشنهادی دقت و خطای خوشه‌بندی را در مقایسه با الگوریتم سنتی و دو الگوریتم بهینه‌سازی مراکز اولیه دیگر بهبود می بخشد .
Liu, Bao and Ding (2018)	این الگوریتم رابطهٔ معنایی را در داده‌ها بیان می‌کند، و انتخاب مرکز خوشه‌بندی اولیهٔ الگوریتم کا-میانه را بر اساس چگالی شبکه بهینه می‌کند.	دقت و پیچیدگی زمانی در الگوریتم پیشنهادی در مقایسه با الگوریتم سنتی و الگوریتم کا-میانه++ بهبود یافته است.
Masud et al. (2019)	در الگوریتم پیشنهادی که الگوریتم IK - means نامگذاری شده، دو مرحلهٔ تخمین چگالی به‌عنوان فاز اول و خوشه‌بندی به‌عنوان فاز دوم در نظر گرفته می‌شود. از ساختار داده درخت Kd[1] برای نمایش و نگهداری اشیای داده استفاده شده و تکنیک تخمین تراکم هسته برای تعیین مناطق متراکم (چگال) نقاط داده اعمال شده است.	این الگوریتم خوشه‌بندی را با دقت بهتری انجام می‌دهد و بهبود کیفیت خوشه‌بندی را در مقایسه با الگوریتم کا-میانهٔ متداول تضمین می‌کند.
Fränti and Sieranoja (2019)	در پژوهش صورت‌گرفته مهم‌ترین عواملی که باعث کاهش عملکرد الگوریتم کا-میانه شده، و اینکه چقدر می‌توان با استفاده از دو تکنیک (یکی مقداردهی اولیهٔ بهتر و دیگری با تکرار (شروع مجدد) الگوریتم) بر این عوامل چالشی غلبه کرد، مورد بررسی و آزمایش قرار گرفته.	یافته‌ها نشان داد که وقتی خوشه‌ها با هم همپوشانی[1] داشته باشند، الگوریتم کا-میانه با استفاده از این دو تکنیک به‌طور قابل توجهی بهبود می‌یابد.

1. Kd-tree

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Awawdeh, Edinat & Sleit (2019)	الگوریتم پیشنهادی شامل چهار مرحله است: فاز ۱: استفاده از الگوریتم ژنتیک (GA)؛ فاز ۲: رسیدگی به داده‌هایی با بیش از یک خصوصیت؛ فاز ۳: شامل سه مرحلهٔ مرتب‌سازی، تقسیم لیست مرتب‌شده به k خوشه و یافتن میانگین و مراکز خوشهٔ اولیه برای فاز ۴؛ و فاز ۴: اعمال الگوریتم کا-میانه سنتی بر اساس تعیین مراکز خوشهٔ اولیه در فاز ۳.	روش پیشنهادی توانایی مقابله با داده‌های چند ویژگی را دارد و زمان محاسباتی کمتری دارد. الگوریتم پیشنهادی، نتایج خوشه‌بندی مناسبی را ارائه داده است.
Taihao et al. (2020)	در الگوریتم پیشنهادی بر خلاف الگوریتم سنتی که نقاط دورافتاده را نادیده می‌گیرد، ابتدا نقاط دورافتاده تشخیص داده می‌شوند، و سپس، حذف می‌شوند.	الگوریتم با بهینه‌سازی مرکز خوشه‌بندی اولیه کارایی خوشه‌بندی را افزایش می‌دهد و نسبت به الگوریتم سنتی بهتر عمل می‌کند.
Kim, Kim and Cho (2020)	الگوریتم پیشنهادی در دو مرحله بهبود می‌یابد: ۱) به‌جای انتخاب مراکز اولیهٔ تصادفی، روشی برای انتخاب مراکز اولیه برای داده‌های پراکنده با ابعاد بالا و ۲) روشی برای اعمال پراکندگی جهت حفظ مرکز پراکندگی ارائه می‌شود.	روش پیشنهادی از نظر محاسباتی کارآمدتر از کا-میانه++[2] است. محاسبات سریع و سرعت همگرایی آن، آن را برای خوشه‌بندی تعداد زیادی از اسناد مناسب ساخته است.

با توجه به اینکه در الگوریتم کا-میانه مراکز خوشهٔ اولیه به‌صورت تصادفی انتخاب می‌شوند، خروجی این الگوریتم متأثر از این انتخاب تصادفی مراکز است (Kant & Ansari 2016)، و به‌عنوان یکی از کاستی‌های این الگوریتم تلقی می‌شود. چنان‌که در جدول ۱، مشاهده می‌شود، از مجموع ۴۷ تحقیق مورد بررسی در این پژوهش، ۲۲ تحقیق با هدف غلبه بر کاستی حساس به مراکز خوشهٔ اولیه انجام گرفته و بیشترین توجه پژوهشگران را طی سال‌های ۲۰۱۰ تا ۲۰۲۰ به این مسئله جلب کرده است. از جملهٔ نتایج به‌دست‌آمده افزایش دقت، سرعت، پایداری و کارایی خوشه‌بندی است. این نتایج برای مجموعهٔ داده‌های مختلف و بزرگ حائز اهمیت است. در ادامه، در جدول ۲، پژوهش‌های انجام‌شده با هدف غلبه بر لزوم تعیین خوشه ارائه شده است.

1. overlap
2. K-means++

جدول ۲. تحقیقات انجام‌شده با هدف غلبه بر لزوم تعیین تعداد خوشه

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Zhu and Wang (2010)	استفاده از الگوریتم ژنتیک برای بهینه‌سازی تعداد خوشه‌ها (مقدار k) و بهبود عملکرد خوشه‌بندی	بهبود بیشتر خصوصیات خوشه، و ارتقای معنادار الگوریتم پیشنهادی
Chadha and Kumar (2014)	الگوریتمی برای خوشه‌بندی پیشنهاد داده شده است که به تعداد خوشه‌های K به‌عنوان ورودی نیاز ندارد. در این الگوریتم دو خوشه در ابتدا با انتخاب دو مرکز اولیه که در مجموعهٔ داده‌های دور هستند، ایجاد می‌شوند.	دقت خوش‌بندی توسط الگوریتم پیشنهادی بهتر از الگوریتم کا-میانه اصلی است.
Bide and Shedge (2015)	الگوریتم ارائه‌شده ورودی را به‌عنوان کلید واژه‌ها انتخاب می‌کند و مسئلهٔ خوشه‌بندی را با تقسیم کردن اسناد به گروه‌های کوچک با استفاده از استراتژی تقسیم و غلبه حل می‌کند.	دقت الگوریتم پیشنهادی در مقایسه با الگوریتم موجود از نظر مقیاس اندازه‌گیری F و پیچیدگی زمانی بالاست.
Haraty, Dimishkieh and Masud (2015)	الگوریتم پیشنهادی که G-means نامیده شده است، از یک روش حریصانه برای تولید مراکز اولیه استفاده می‌کند و سپس، k یا کمتر از مجموعهٔ داده‌ها برای تنظیم این نقاط مرکزی استفاده می‌کند.	الگوریتم پیشنهادی از لحاظ آنتروپی و نمرهٔ F از الگوریتم کا-میانه بهتر عمل می‌کند و نتایج بهتری از نظر ضریب واریانس و زمان اجرا به‌دست می‌دهد.
Rajeswa et al. (2015)	ابتدا دو مرکز را از مجموعهٔ داده‌ها انتخاب کنید: پایین‌ترین نقطه مرکزی و بالاترین نقطهٔ مرکزی. پس از انتخاب مراکز، دو خوشه با اعضایی که با هم متفاوت هستند، ایجاد می‌شود.	الگوریتم پیشنهادی در مقایسه با الگوریتم استاندارد کا-میانه از نظر کیفیت و پیچیدگی نتایج بهتری را به همراه داشته است.
Yadav and Dhingra (2016)	الگوریتم پیشنهادی با توجه به گام‌های الگوریتم اقدام به حذف خوشه‌های خالی تولیدشده می‌کند. به این ترتیب که زمانی که شرایط همگرایی برآورده می‌شوند، خوشه‌های تولیدشده دوباره بررسی می‌شوند. خوشه‌هایی که هیچ نقطهٔ داده‌ای به آن اختصاص داده نشده است، در مرحلهٔ تخصیص حذف می‌شوند.	دقت الگوریتم پیشنهادی نسبت به الگوریتم متداول از لحاظ معیار F، فراخوانی، دقت و پیچیدگی زمانی، بالاست. همچنین، داده‌های خوشه‌ای را به‌عنوان فایل‌های متنی روی دیسک ذخیره می‌کند که بتوان آن را در آینده بدون خوشه‌بندی مجدد مورد استفاده قرار داد.
Khatri and Garg 2016))	الگوریتم پیشنهادی برای خوشه‌بندی اسناد به روش دستی استفاده شده است. این الگوریتم از معیار شباهت اقلیدسی برای ایجاد اسناد مشابه در خوشه‌های مناسب استفاده می‌کند.	الگوریتم کا- میانه اصلاح‌شده از لحاظ دقت، معیار اندازه‌گیری F و پیچیدگی زمانی بهتر از الگوریتم موجود عمل می‌کند.

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Bansal, Sharma, and Goel (2017)	الگوریتم خوشه‌بندی کا-میانه به‌گونه‌ای ارائه می‌شود که می‌تواند تعداد خوشه‌ها را به‌صورت خودکار تعریف کرده و خوشهٔ مورد نیاز را به نقاط بدون خوشه اختصاص دهد.	الگوریتم پیشنهادی منجر به بهبود دقت و کاهش زمان محاسباتی خوشه‌بندی می‌شود.
Thilagaraj and Sengottaiyan (2019)	در روش پیشنهادی مرکز ثابت در نظر گرفته می‌شود و از میانگین برای ایجاد خوشه‌های متعادل استفاده می‌شود.	الگوریتم پیشنهادی در مقایسه با الگوریتم خوشه‌بندی کا-میانه اصلی، مرکز ثقل ثابت را پیدا کرده و موفق به ایجادخوشه‌های غیرقابل تغییر شده است.

در الگوریتم کا-میانه تعداد خوشه‌ها باید از قبل مشخص باشد و این مورد به‌عنوان یکی از کاستی‌های آن قابل تأمل و بررسی است (Raval Unnati & Chaita 2016). همان‌گونه که در جدول ۲، قابل مشاهده است، ۹ تحقیق از مجموع ۴۷ تحقیق مورد بررسی با هدف غلبه بر لزوم تعیین تعداد خوشه در الگوریتم کا-میانه در جهت اصلاح آن صورت گرفته است. این تحقیقات نیز پس از تحقیقاتی که بر کاستی حساس به مراکز خوشهٔ اولیه پرداخته‌اند، بیشتر از سایر کاستی‌ها کانون توجه پژوهشگران بوده‌اند. از جمله نتایج حاصل از این تحقیقات در رابطه با رفع این کاستی، دقت و بهبود خوشه‌بندی و کاهش زمان محاسباتی خوشه‌بندی، و ایجاد خوشه‌هایی با کیفیت بهتر و بدون تغییر است (جدول ۳).

جدول ۳. تحقیقات انجام‌شده با هدف غلبه بر کاستی حساس به داده‌های نویزی و دورافتاده

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Wang and Su (2011)	پیش‌پردازش داده‌ها برای حذف داده‌های نویزی قبل از خوشه‌بندی داده‌ها (فیلتر داده‌های نویزی) با استفاده از تشخیص داده‌های دو افتاده مبتنی بر چگالی	الگوریتم پیشنهادی دقت خوشه‌بندی را بهتر می‌کند و برای مجموعهٔ داده‌های کوچک مناسب است و زمان برای مجموعهٔ داده‌های بزرگ به‌دلیل پیمایش بیشتر افزایش خواهد یافت.
Rathore and Shukla (2015)	الگوریتم پیشنهادی ابتدا داده‌ها را جهت افزایش کیفیت پیش‌پردازش کرده و داده‌های دورافتاده را از داده‌های ورودی شناسایی می‌کند. پس از آن داده‌ها با یک توالی از فرایندها و نتایج آن‌ها با استفاده از تکنیک‌های اعتبارسنجی ارزیابی می‌شوند.	الگوریتم پیشنهادی در مقایسه با الگوریتم کا-میانه نتایج مؤثر و بهبود دقت تشکیل خوشه را در مقابل کاهش کارایی نشان می‌دهد. از این رو، برای محیط داده‌های بزرگ با بهبود عملکرد خوشه‌بندی مناسب است.

با توجه به اینکه این الگوریتم به مراکز خوشهٔ اولیه حساس است، در صورتی که تعدادی از داده‌ها دورافتاده و نویزی باشند، این امکان وجود دارد که مراکز خوشهٔ جدید از مراکز واقعی منحرف شده و خروجی خوشه‌بندی را تحت تأثیر قرار دهد (Wang & Su 2011). این یکی دیگر از کاستی‌های این الگوریتم است. شناسایی داده‌های دورافتاده به یافتن خوشه‌های متراکم و واضح کمک می‌کند (Rathore & Shukla 2015). بنابراین، یکی از راه‌های رفع این مشکل شناسایی این داده‌ها با راهکارهای مناسب و در صورت لزوم حذف آن‌هاست. چنانچه در جدول ۳، مشاهده می‌شود، تنها ۲ تحقیق به رفع کاستی حساس به داده‌های نویزی و دورافتاده در جهت بهبود الگوریتم کا-میانه پرداخته‌اند. این تحقیقات نیز در سال‌های ۲۰۱۱ و ۲۰۱۵ انجام شده‌اند که نشان‌دهندهٔ این است که محققان تأثیر این کاستی در ایجاد یک خوشه‌بندی بهینه را کمتر از سایر کاستی‌های الگوریتم کا-میانه دانسته‌اند. آنچه که از یافته‌های این دو تحقیق قابل تأمل است، توجه به عملکرد الگوریتم‌های بهبوددهندهٔ الگوریتم کا-میانه بر روی حجم داده‌هاست. در تحقیقی که توسط Wang & Su (2011) انجام شده، الگوریتم بهبودیافته برای مجموعهٔ داده‌های کوچک مناسب است، اما برای کار با مجموعهٔ داده‌های بزرگ به‌دلیل پیمایش بیشتر، زمانِ بیشتری صرف خوشه‌بندی خواهد شد. و تحقیق بعدی الگوریتم پیشنهادی که Rathore & Shukla (2015) ارائه داده‌اند، عملکرد قابل قبولی در خوشه‌بندی مجموعهٔ داده‌های بزرگ وجود دارد. در ادامه، در جدول ۴، پژوهش‌های انجام‌شده با هدف غلبه بر کاستیِ گیر افتادن در بهینهٔ محلی و همگرایی زودرس ارائه شده است.

جدول ٤. تحقیقات انجام‌شده با هدف غلبه بر کاستیِ گیر افتادن در بهینهٔ محلی و همگرایی زودرس

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Iezzi (2012)	یک نسخهٔ جدید از کا-میانه به نام AIC-k-means که از شاخص مرکزیت اطلاعات[1] (AIC) برای انتخاب مراکز استفاده می‌کند.	روش پیشنهادی با شناسایی مراکز اولیه به‌عنوان نمونه‌هایی از پیکرهٔ زبانی الگوریتم کا-میانه را بهبود می‌بخشد و گروه‌هایی را با چسبندگی داخلی بالا و سطح خوبی از جدایی شناسایی می‌کند.

گیر افتادن در بهینهٔ محلی و همگرایی زودرس یکی دیگر از کاستی‌های این الگوریتم شمرده می‌شود (Larose & Larose 2014). همان‌گونه که در جدول ۴، قابل مشاهده است،

1. actor information centrality

تنها یک تحقیق در سال ۲۰۱۲ با هدف غلبه بر کاستی گیر افتادن الگوریتم کا-میانه در بهینهٔ محلی و همگرایی زودرس انجام شده است. این کاستی نیز کمترین مقبولیت پژوهش را از جانب پژوهشگران داشته است. در ادامه، تحقیقات انجام‌شده با غلبه برکاستیِ حساس به ابعاد بالای ویژگی در جدول ۵، قابل مشاهده است.

جدول ۵. تحقیقات انجام‌شده با غلبه برکاستی حساس به ابعاد بالای ویژگی

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Prabhu and Anbazhagan (2011)	در روش پیشنهادی از تحلیل مؤلفه‌های اصلی[1] (PCA) برای کاهش مجموعهٔ داده‌ها استفاده شده و مراکز خوشهٔ اولیه با میانهٔ داده‌های کاهش‌یافتهٔ تقسیم‌بندی شده استخراج شده‌اند.	روش پیشنهادی جهت اصلاح الگوریتم کا-میانه بهبود دقت خوشه‌بندی را به همراه دارد.
Zhang et al. (2013)	الگوریتم خوشه‌بندی کا-میانه، بر اساس آنالیز تفکیک‌کنندهٔ خطی[2] یعنی الگوریتم LKM پیشنهاد داده شده است. سپس، الگوریتم کا-میانه برای تحلیل خوشه‌بندی اعمال شده است.	الگوریتم پیشنهادی زمان استخراج ویژگی نمونه را کوتاه می‌کند و دقت الگوریتم خوشه‌بندی کا-میانه را افزایش می‌دهد. در نتیجه، تجزیه و تحلیل و پردازش داده‌های گسترده را بهبود می‌بخشد.
Wu et al. (2015)	یک الگوریتم کا-میانه بر مبنای Sim Hash پیشنهاد می‌شود. پس از پیش‌پردازش متن، Sim Hash برای محاسبهٔ بردار ویژگی استخراج شده و سپس اثر انگشت هر متن استفاده می‌شود.	این الگوریتم کیفیت خوشه‌بندی را افزایش می‌دهد، اما اگر طول متن نسبتاً کوچک باشد، دقت آن کمتر است. بنابراین، ممکن است برای مجموعهٔ متون کوتاه برای خوشه‌بندی مناسب نباشد.
Tunali, Bilgin and Camurcu (2016)	در روش پیشنهادی الگوریتمی با نام MCSKM بر مبنای الگوریتم کا-میانه کروی چندخوشه‌ای[3] (SKM) برای خوشه‌بندی مجموعهٔ اسناد با ابعاد بالا و بزرگ با عملکرد و کارایی زیاد توسعه داده می‌شود.	الگوریتم پیشنهادی باعث افزایش قابل توجه کیفیت خوشه‌بندی بدون ایجاد تفاوت قابل توجه در استفاده از CPU در مقایسه با الگوریتم SKM می‌شود.

ابعاد بالای ویژگی یکی از کاستی‌های الگوریتم کا-میانه است که در طی فرایند خوشه‌بندی توسط این الگوریتم منجر به کاهش دقت و کارایی خوشه‌بندی می‌شود. جهت رفع این کاستی انتخاب روش‌هایی در جهت کاهش ابعاد می‌توانند به بهبود عملکرد خوشه‌بندی در برخورد با مجموعهٔ داده‌ها با ابعاد بالا کمک کند (Zhang et al. 2013). بر

1. principal component analysis 2. linear discriminant analysis (LDA) 3. multi-cluster spherical K-Means

اساس آنچه که در جدول ۵، آمده، ۴ تحقیق، حساسیت الگوریتم کا-میانه را با ابعاد بالای ویژگی مورد بررسی قرار داده‌اند. بنابراین، به نظر می‌رسد که این نوع تحقیقات نیز کمتر مورد توجه پژوهشگران قرار گرفته است. یافته‌های Zhang et al. (2013) و Wu et al. (2015) حاکی از آن است که با توجه به اینکه بهبود الگوریتم کا-میانه با رفع این کاستی جهت خوشه‌بندی متأثر از حجم داده‌هاست، برای رسیدن به یک دقت قابل قبول جهت خوشه‌بندی، بهبود الگوریتم کا-میانه باید متناسب با حجم داده‌ها انجام شود. جدول ۶، پژوهش‌های صورت‌گرفته با هدف غلبه بر دو کاستی از کاستی‌های الگوریتم کا-میانه را ارائه می‌دهد.

جدول ٦. تحقیقات انجام‌شده با هدف غلبه بر دو کاستی از کاستی‌های الگوریتم کا-میانه

نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Napoleon and Lakshmi (2010)	در نظر گرفتن دو فاز برای بهبود الگوریتم که مرکز اولیهٔ خوشه‌ها به‌عنوان ورودی فاز دوم در نظر گرفته می‌شود و فاز دوم تعیین هر نقطه داده به خوشه‌های مناسب است.	کاهش پیچیدگی در جهت خوشه‌بندی بهتر و مناسب‌تر بودن الگوریتم کا-میانه بهبودیافته برای مجموعهٔ داده‌های بسیار زیاد
Tajunisha and Saravanan (2011)	روش پیشنهادی تقسیم‌بندی داده‌ها را با تحلیل مؤلفه‌های اصلی (PCA) به‌منظور پیدا کردن مراکز خوشهٔ اولیه برای کا-میانه و برای کاهش ابعاد انجام می‌دهد.	روش پیشنهادی با کاهش پیچیدگی محاسباتی کارایی الگوریتم را بهبود داده است، و زمان اجرا و دقت نتایج خوشه‌بندی را بهبود بخشیده است.
Xinwu (2012)	ترکیب الگوریتم کا-میانه و مدل خود سازمانی[1] (SOM)	الگوریتم ترکیبی خوشه‌بندی پیشنهادی معایب طول مدت همگرایی شبکهٔ مدل خودسازمانی و اثر خوشه‌بندی بد ناشی از انتخاب نامناسب مرکز خوشهٔ اولیه الگوریتم کا-میانه را جبران نموده است.
Kaur, Sahiwal and Kaur (2012)	روش پیشنهادی با استفاده از دو روش استفاده از الگوریتم خوشه‌بندی کا-میانه متداول با ادغام حد آستانه و اعمال روش رتبه‌بندی[2] روی الگوریتم کا-میانه ارائه شده است.	روش پیشنهادی خوشه‌بندی را در زمان اجرای کمتر نسبت به روش خوشه‌بندی با الگوریتم کا-میانه متداول انجام داده است.
Liu et al. (2014)	یک الگوریتم پیشنهادی بر اساس قوانین انجمنی[3] ارائه شده است که در آن کوچک‌ترین قوانینی که مجموعه را به‌عنوان اساس پوشش می‌دهد، پیشنهاد شده است.	روش پیشنهادی دو محدودیت بزرگ الگوریتم کا-میانه یعنی انتخاب صحیح تعداد خوشه و انتخاب تصادفی مرکز اولیه را به‌خوبی حل کرده است.

1. self-organizing model 2. ranking method 3. association rules

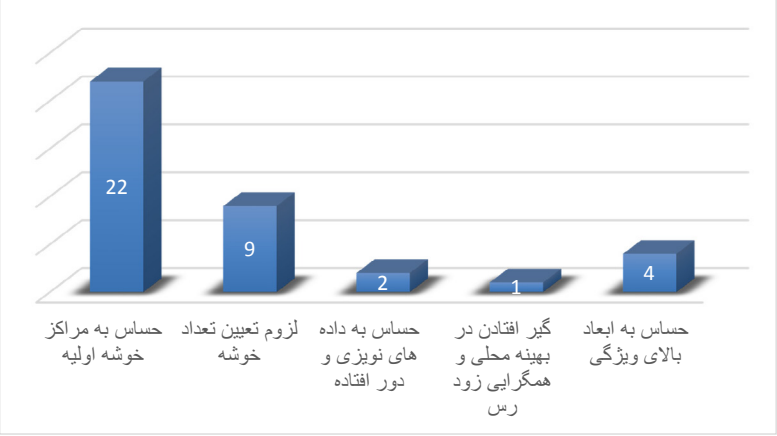
نویسنده و سال انتشار	روش پژوهش	یافته‌ها
Yu et al. (2018)	در روش پیشنهادی یک الگوریتم کا-میانه دو لایه [1] و یک الگوریتم کا-میانه سه‌سطحی [2] پیشنهاد شده است. در عین حال، الگوریتمی مبتنی بر ژنتیک [3] برای استخراج پارامترهای بهینه استفاده‌شده در الگوریتم‌های سه‌سطحی و دوسطحی ارائه گردیده است.	هر دو الگوریتم پیشنهادی می‌توانند دقت بالاتری نسبت به الگوریتم کا-میانه رایج داشته باشند.
Zhang, Zhang and Zhang (2018)	در الگوریتم بهبودیافته پیشنهادی، پارامتر چگالی اضافه می‌شود. چگالی Canopy به‌عنوان روش پیش‌پردازش کا-میانه و نتیجه آن به‌عنوان عدد خوشه و مرکز خوشه‌بندی اولیه الگوریتم کا-میانه مورد استفاده قرار می‌گیرد.	الگوریتم کا-میانه مبتنی بر چگالی Canopy به نتایج خوشه‌بندی بهتری نسبت به الگوریتم کا-میانه سنتی [4]، الگوریتم کا-میانه مبتنی بر Canopy [5]، الگوریتم نیمه نظارتی کا-میانه ++ [6] و الگوریتم کا-میانه-u [7] برای الگوریتم کا-میانه دست می‌یابد.
Xie et al. (2019)	در روش پیشنهادی دو نوع الگوریتم کرم شب‌تاب [8] (FA) به نام‌های IIEFA و CIEFA جهت رفع کاستی حساسیت به مراکز خوشهٔ اولیه و گیر افتادن در بهینهٔ محلی الگوریتم کا-میانه پیشنهاد شده است. برای افزایش قابلیت بهره‌برداری و اکتشاف، پارامترهای جست‌وجو مبتنی بر ماتریس و مکانیسم‌های پراکنده در دو مدل پیشنهادی FA ترکیب می‌شوند.	مدل‌های پیشنهادی FA در مقایسه با خوشه‌بندی کا-میانه، پنج روش جست‌وجوی کلاسیک و پنج نوع پیشرفته FA برتری آماری معناداری را در هر دو مقیاس فاصله و عملکرد برای عملیات خوشه‌بندی نشان می‌دهند.
Zheng (2020)	در روش پیشنهادی دو اصل بهینه‌سازی کاهش تعداد تکرار در فرایند خوشه‌بندی و کاهش مقدار داده در فرایند خوشه‌بندی پیشنهاد شده است. اطلاعات اضافی ایجادشده توسط تغییر پویای اطلاعات به‌منظور کاهش تداخل در فرایند خوشه‌بندی دینامیک حذف می‌شود.	الگوریتم بهبودیافته، بهبود بیشتری در دقت و کارایی نسبت به الگوریتم کا-میانه سنتی دارد، و هرچه مقدار داده بزرگ‌تر باشد، کارایی بالاتر است.

مطابق آنچه که در جدول ۶، آمده، تعداد ۹ تحقیق با هدف غلبه بر دو کاستی، این الگوریتم را بهبود بخشیده‌اند. این موضوع نیز همچون غلبه بر کاستی حساس به مراکز خوشهٔ اولیه تاکنون، یعنی تا سال ۲۰۲۰، توسط پژوهشگران به‌عنوان یک تحقیق قابل بررسی مورد پژوهش قرار گرفته است. از جمله نتایج به‌دست‌آمده از یافته‌های این

1. bi-layer k-means algorithm
2. tri-level k-means algorithm
3. genetic-based algorithm
4. traditional K-means algorithm
5. Canopy-based K-means algorithm
6. supervised K-means++ algorithm
7. K-means-u (corresponding author)
8. firefly algorithm

جدول افزایش دقت، سرعت، کارایی خوشه‌بندی حاصل از بهبود الگوریتم کا-میانه است و تحقیقی که توسط Zheng (2020) انجام شده، نشان می‌دهد که هرچه قدر حجم داده‌ها بزرگ‌تر باشد، الگوریتم بهبودیافتهٔ پیشنهادی آن‌ها دارای کارایی بهتری است.

همچنین، با توجه به بررسی‌های انجام‌شده در راستای هدف پژوهش حاضر، نمودار ۱، توزیع فراوانی تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه را نشان می‌دهد. همان‌طور که مشاهده می‌شود، کاستی حساس به مراکز خوشهٔ اولیه بیشترین تعداد پژوهش را به خود اختصاص داده است. بنابراین، در پاسخ به سؤال اول پژوهش نتایج حاکی از آن است که کارآمدی الگوریتم کا-میانه در غلبه بر کاستی حساس به مراکز خوشهٔ اولیه بیش از سایر تحقیقات در این زمینه مورد توجه پژوهشگران بوده است.



نمودار ۱. توزیع فراوانی تحقیق انجام‌شده در جهت بهبود الگوریتم کا-میانه

پرسش دوم پژوهش: در تحقیقات صورت‌گرفته توجه به داده‌های متنی و غیرمتنی به چه میزان بوده است؟

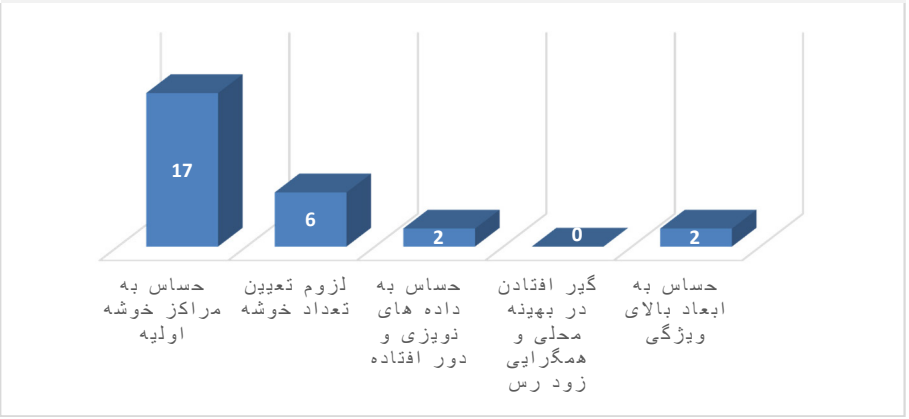
برای پاسخ به سؤال دوم پژوهش، مجموع ۴۷ تحقیق مورد بررسی در قالب جداول ۷ و ۸ به‌ترتیب، بر مبنای داده‌های غیرمتنی و داده‌های متنی آمده است.

جدول ۷. تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه بر روی داده‌های غیرمتنی

مولفان	کاستی‌های مورد بررسی نتایج				
	حساس به مراکز خوشه اولیه	لزوم تعیین تعداد خوشه	حساس به داده‌های نویزی و دور افتاده	گیر افتادن در بهینه محلی و همگرایی زود رس	حساس به ابعاد بالای ویژگی
Na, Xumin & Yong (2010)	✓				
edla, Pathakota & Srinivasa (2010)	✓				
Zhu and Wang (2010)		✓			
Napoleon and Lakshmi (2010)	✓	✓			
Wang and Su (2011)			✓		
Prabhu and Anbazhagan (2011)					✓
Tajunisha and Saravanan (2011)	✓				✓
Xinwu (2012)	✓		✓		
Kaur, Sahiwal & Kaur (2012)	✓	✓			
Karegowda et al. (2013)	✓				
Chaturvedi and Rajavat (2013)	✓				
Shunye (2013)	✓				
Zhang et al. (2013)					✓
Goyal and Kumar (2014)	✓				
Chadha and Kumar (2014)		✓			
Haraty, Dimishkieh & Masud (2015)		✓			
Rajeswari et al. (2015)		✓			
Rathore and Shukla (2015)			✓		
Yadav and Singh (2016)	✓				
Kant and Ansari (2016)	✓				

مولفان	کاستی‌های مورد بررسی نتایج				
	حساس به مراکز خوشه اولیه	لزوم تعیین تعداد خوشه	حساس به داده‌های نویزی و دورافتاده	گیر افتادن در بهینه محلی و همگرایی زود رس	حساس به ابعاد بالای ویژگی
Choudhary, Sharma & Singh (2016)	✓				
Raval Unnati and Chaita (2016)	✓				
Saklecha and Raikwal (2017)	✓				
Bansal, Sharma & Goel (2017)		✓			
Linyao and Jianguo (2018)	✓				
Liu, Bao & Ding (2018)	✓				
Yu et al. (2018)	✓		✓		
Zhang, Zhang & Zhang (2018)	✓	✓			
Masud et al. (2019)	✓				
Fränti and Sieranoja (2019)	✓				
Awawdeh, Edinat & Sleit (2019)	✓				
Thilagaraj and Sengottaiyan (2019)		✓			
Xie et al. (2019)	✓			✓	
Taihao et al. (2020)	✓				
Zheng (2020)	✓	✓			

همان‌گونه که در جدول ۷، قابل مشاهده است، از مجموع ۴۷ تحقیق صورت‌گرفته، ۳۵ تحقیق در جهت بهبود الگوریتم کا-میانه بر روی داده‌های غیرمتنی انجام شده است. از این تعداد، ۲۸ تحقیق یکی از کاستی‌های این الگوریتم را مورد هدف پژوهش خود قرار داده، و ۷ تحقیق به‌طور همزمان دو کاستی را بررسی کرده‌اند. نتایج حاصل در نمودار ۲، قابل مشاهده است.



نمودار ۲. توزیع فراوانی تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه بر روی داده‌های غیرمتنی

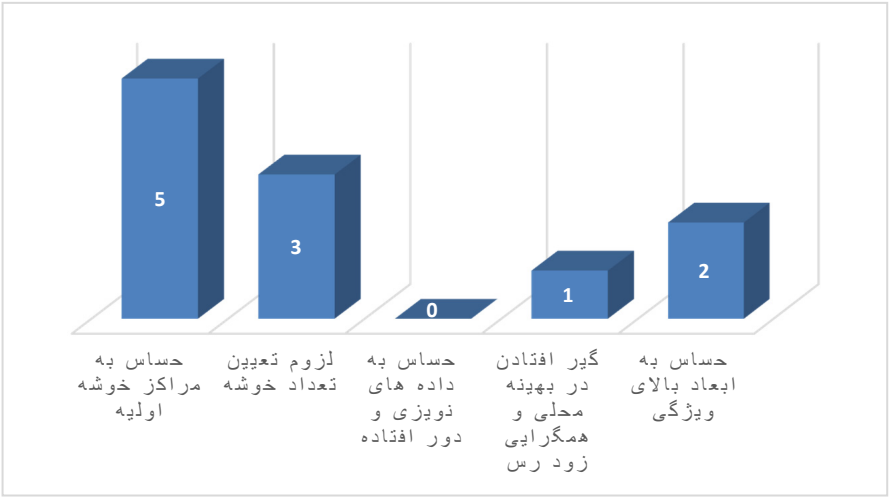
در جدول ۸، به تحقیقاتی اشاره شده است که در آن الگوریتم‌های بهبوددهندهٔ کا-میانه بر روی داده‌های متنی مورد بررسی قرار داده شده است.

جدول ۸. تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه بر روی داده‌های متنی

مؤلفان	کاستی‌های مورد بررسی نتایج				
	حساس به مراکز خوشه اولیه	لزوم تعیین تعداد خوشه	حساس به داده‌های نویزی و دورافتاده	گیر افتادن در بهینه محلی و همگرایی زود رس	حساس به ابعاد بالای ویژگی
Iezzi (2012)				✓	
Jaganathan and Jaiganesh (2013)	✓				
Ma (2014)	✓				
Liu et al. (2014)	✓	✓			
Bide and Shedge (2015)		✓			
Wu et al. (2015)					✓
Yadav and Dhingra (2016)		✓			
Xiong et al. (2016)	✓				
Vashist and Nath (2016)	✓				

مؤلفان	کاستی‌های مورد بررسی نتایج				
	حساس به مراکز خوشه اولیه	لزوم تعیین تعداد خوشه	حساس به داده‌های نویزی و دورافتاده	گیر افتادن در بهینه محلی و همگرایی زود رس	حساس به ابعاد بالای ویژگی
Khatri and Garg (2016)		✓			
Tunali, Bilgin & Camurcu)2016(					✓
Kim, Kim & Cho (2020)	✓				

چنانکه در جدول ۸، مشاهده می‌شود، ۱۲ تحقیق از مجموع ۴۷ تحقیق در جهت بهبود الگوریتم کا-میانه بر روی داده‌های متنی انجام شده است و از این تعداد ۱۱ تحقیق یکی از کاستی‌های این الگوریتم را بر روی داده‌های متنی بررسی می‌کند، و یک تحقیق، به‌طور همزمان دو کاستی را بر روی داده‌های متنی مورد بررسی قرار می‌دهد. همچنین، نتایج حاصل در نمودار ۳، قابل مشاهده است.



نمودار ۳. توزیع فراوانی تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه بر روی داده‌های متنی

بر اساس یافته‌های پژوهش، بهبود الگوریتم کا-میانه با رفع کاستی حساس به مراکز خوشه اولیه بیشترین اولویت پژوهشی را در تحقیقات انجام‌شده داشته است. این نتیجه در تحقیقات انجام‌شده بر روی داده‌های غیرمتنی و داده‌های متنی نیز صدق می‌کند. اما آنچه که قابل تأمل است، اختلاف اندک کاستی حساس به مراکز خوشهٔ اولیه با کاستی

لـزوم تعییـن تعـداد خوشـه در تحقیقاتـی اسـت کـه بـر روی داده‌هـای متنـی انجـام شـده اسـت. ایـن نشـان‌دهندهٔ آن اسـت کـه لـزوم تعییـن تعـداد خوشـه نیـز بـه انـدازهٔ مراکـز خوشـهٔ اولیـه در خوشـه‌بندی داده‌هـای متنـی دارای اهمیـت بیشـتری نسـبت بـه سـایر کاستی‌هاسـت (نمـودار ۲).

یـک خوشـه‌بندی مناسـب در جهـت سـازماندهی داده‌هـا بایـد بتوانـد بـا حداقـل خطـا، خوشـه‌های قابـل قبولـی را از نظـر گروه‌بنـدی داده‌هـای شـبیه‌به‌هم در یـک خوشـه ایجـاد کنـد کـه بـا داده‌هـای خوشـه‌های دیگـر بیشـترین تفـاوت را داشـته باشـد. یافته‌هـای ایـن تحقیقـات در جهـت بهبـود الگوریتـم کا-میانـه در مجمـوع در افزایـش دقـت، سـرعت، کارایـی، پایـداری، و کیفیـت خوشـه‌بندی تأثیـر مثبتـی داشـته‌اند.

پرسـش سـوم پژوهـش: حجـم داده‌هـا در تحقیقـات صورت‌گرفتـه جهـت بهبـود الگوریتـم کا-میانه در فرایند خوشه‌بندی تا چه حد تأثیرگذار بوده است؟

در ادامـه، همان‌طـور کـه در جـدول ۹، قابـل مشـاهده اسـت، بـرای پاسـخ بـه سـؤال سـوم پژوهـش، تعـداد ۸ تحقیـق شناسـایی و بررسـی شـد. هـر یـک از ایـن تحقیقـات بـا مـورد توجـه قـرار دادن حجـم داده‌هـا به‌نوعـی اهمیـت آن را آشـکارتر سـاخته و بـا تأکیـد بـر ایـن نکتـه، پژوهشـگران را بـه توجـه ویـژه بـه آن در پژوهش‌هـای آینـده ترغیـب می‌کننـد.

جدول ۹. تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه با تأثیر حجم داده‌ها بر فرایند خوش‌بندی

مؤلفان و سال انتشار	یافته‌ها
Napoleon and Lakshmi (2010)	الگوریتمـی ارائـه دادنـد کـه بـا کاهـش زمـان اجـرا بـرای مجموعـهٔ داده‌هـای بـزرگ کارآمدتـر اسـت.
Wang and Su (2011)	نتایـج حاصـل از تحقیـق حاکـی از آن اسـت کـه الگوریتـم ارائه‌شـده توسـط آن‌هـا بـرای مجموعـهٔ داده‌هـای کوچـک مناسـب اسـت، ولـی در پیمایـش مجموعـهٔ داده‌هـای بـزرگ، به‌دلیـل افزایـش زمـان نتیجـه مطلوبـی دربـر نخواهـد داشـت.
Shunye (2013)	نتـاج حاصـل از آزمایش‌هـا نشـان داده اسـت کـه الگوریتـم پیشـنهادی آن‌هـا بـرای مجموعـهٔ داده بـزرگ و مجموعـهٔ داده سـطح بـالا مناسـب‌تر اسـت. بـا وجـود ایـن، اعـلام کرده‌انـد کـه ایـن الگوریتـم هنـوز هـم دارای مشـکلاتی اسـت و قابـل بررسـی اسـت، امـا به‌دلیـل محدودیـت در شـرایط پژوهـش بـه آن پرداختـه نشـده اسـت.
Zhang et al. (2013)	بـا ارائـهٔ الگوریتمـی کـه زمـان اسـتخراج ویژگـی نمونـه را کوتـاه می‌کنـد و دقـت الگوریتـم خوشـه‌بندی کا-میانـه را افزایـش می‌دهـد، عملکـرد الگوریتـم خوشـه‌بندی کا-میانـه بـرای تجزیـه و تحلیـل و پـردازش داده‌هـای گسـترده را بهبـود دادنـد.
Wu et al. (2015)	به‌علـت اسـتفاده از Sim Hash بـرای محاسـبهٔ شـباهت متـن و کاهـش ابعـاد ویژگـی بـرای مجموعـهٔ متـون کوتـاه بـرای خوشـه‌بندی مناسـب اعـلام نشـد.

مؤلفان و سال انتشار	یافته‌ها
Tunali, Bilgin & Camurcu (2016)	با توجه به نتایج به‌دست‌آمده، الگوریتم بهبودیافته برای خوشه‌بندی مجموعه‌های بسیار بزرگ اسناد مناسب و قابل قبول است.
Zheng (2020)	در آزمایش‌های خود به این نتیجه رسید که هر چقدر داده‌ها از حجم بیشتری برخوردار باشند، الگوریتم پیشنهادی او از بازده بالاتری برخوردار خواهد بود.
Kim, Kim & Cho (2020)	روش پیشنهادی با زمان محاسبات سریع و سرعت همگرایی برای خوشه‌بندی تعداد زیادی از اسناد مناسب است.

آنچه که از نتایج این تحقیقات حاصل شد، این است که الگوریتم کا-میانه بهبودیافته می‌تواند با تأثیر گرفتن از حجم متغیر داده‌ها دارای عملکردهای متغیری در خوشه‌بندی داده‌ها باشد. به‌عبارت دیگر، می‌توان گفت که حجم داده‌ها رابطه‌ای مستقیم با عملکرد الگوریتم بهبودیافته کا-میانه دارد و این الگوریتم باید به‌نوعی اصلاح شود که با اعمال بر روی حجم متفاوت داده‌ها بتواند خوشه‌بندی دقیق و اثربخشی انجام دهد.

۵. نتیجه‌گیری

شناخت و آگاهی یافتن از شکاف‌های پژوهشی موجب می‌شود که محققان و پژوهشگران در راستای اهداف پژوهشی خود به‌درستی گام بردارند. به همین دلیل، در پژوهش حاضر در قالب مروری نظام‌مند، تحقیقات انجام‌شده در جهت بهبود الگوریتم کا-میانه برای سازماندهی داده‌ها در طی سال‌های ۲۰۱۰ تا ۲۰۲۰ بررسی شدند. از آنجا که رفع کاستی‌های موجود در الگوریتم کا-میانه به بهبود آن در جهت رسیدن به نتایج مؤثرتر و کاراتر کمک به‌سزایی می‌کند، بررسی منابع موجود بر مبنای کاستی‌های الگوریتم کا-میانه شامل حساس به مراکز خوشهٔ اولیه، لزوم تعیین تعداد خوشه‌ها، حساس به داده‌های نویزی و دورافتاده، گیر افتادن در بهینهٔ محلی و همگرایی زودرس حساس به ابعاد بالای ویژگی انجام گرفت. در پاسخ به سؤالات پژوهش، طبق جداول ۱ تا ۶ از مجموع ۴۷ منبع مورد مطالعه، ۲۲ منبع با هدف غلبه بر کاستی حساس به مراکز خوشهٔ اولیه، و ۹ منبع به رفع کاستی لزوم تعیین تعداد خوشه‌ها، و ۹ منبع با هدف غلبه بر رفع دو کاستی این الگوریتم انجام گرفته بودند که به‌ترتیب، بیشترین سهم را در این تحقیقات به خود اختصاص دادند. با توجه به اینکه انتخاب مراکز خوشهٔ اولیه در الگوریتم کا-میانه به‌صورت تصادفی انتخاب می‌شود، به‌کارگیری یک روش مناسب برای انتخاب مراکز خوشهٔ اولیه رابطه‌ای مستقیم با کیفیت خوشه‌بندی دارد. همچنین، تعیین تعداد بهینهٔ خوشه یکی از پارامترهای اولیهٔ الگوریتم کا-میانه است. اگر این پارامتر به‌درستی

تعیین نشود، سبب قرار گرفتن نتایج الگوریتم کا-میانه در دام بهینهٔ محلی می‌شود. با توجه به بررسی منابع مورد مطالعه در این پژوهش می‌توان گفت که نوعی همپوشانی متقابل در غلبه بر کاستی‌های این الگوریتم وجود دارد. این همپوشانی در بهبود الگوریتم حایز اهمیت است. بنابراین، شاید بتوان گفت که این مسئله برای تحقیقات بیشتر در زمینهٔ بهبود الگوریتم کا-میانه توجیه مناسبی بوده است. یافته‌های این تحقیقات حاکی از تأثیرگذاری مثبت اصلاح این الگوریتم در افزایش دقت، سرعت، کارایی، پایداری، و کیفیت خوشه‌بندی است.

بر اساس آنچه پیش‌تر بیان شد، بهبود الگوریتم کا-میانه با رفع نقص حساس به مراکز خوشهٔ اولیه بیشترین اولویت پژوهشی را در تحقیقات انجام‌شده داشته‌اند و این نتیجه در دو دسته‌بندی ایجادشده در این پژوهش که بر مبنای نقایص الگوریتم کا-میانه بر روی داده‌های متنی و غیرمتنی است، نیز صادق است. بر اساس جداول ۷ و ۸ از مجموع ۴۷ تحقیق صورت‌گرفته، ۳۵ تحقیق در جهت بهبود الگوریتم کا-میانه بر روی داده‌های غیرمتنی و ۱۲ تحقیق بر روی داده‌های متنی انجام شده است. با توجه به مطالعهٔ منابع مورد بررسی، زمانی که هدف، خوشه‌بندی داده‌هاست، توجه به حجم داده‌ها که روندی روبه‌رشد دارد، به‌دلیل انتخاب یک الگوریتم مناسب که بتواند خوشه‌بندی مناسبی را برای حجم داده‌های بزرگ ارائه دهد، حایز اهمیت است؛ زیرا ممکن است یک الگوریتم بهبودیافته در جریان خوشه‌بندی برای داده‌های با حجم زیاد مناسب و برای داده‌های با حجم کم نامناسب باشد. از آنجا که حجم داده‌های متنی نسبت به داده‌های غیرمتنی با رشد بیشتر و سریع‌تری همراه است، توجه به حجم داده‌ها و طول متون، روابط معنایی و ویژگی‌هایی از این دست در چگونگی بهبود این الگوریتم به‌گونه‌ای که بتواند داده‌های متنی را به مناسب‌ترین شکل خوشه‌بندی کند نیز مورد توجه است. به‌عنوان مثال، اگر متن کوتاه و یا بلند باشد، بر روی دقت خوشه‌بندی تأثیرگذار است. از این رو، در یافته‌های پژوهش این نتیجه حاصل شد که در تحقیقات انجام‌شده به حجم داده‌های بزرگ و تأثیر آن بر فرایند خوشه‌بندی نیز توجه شده است. در نهایت، می‌توان گفت در صورتی که بهبود الگوریتم کا-میانه در رفع نقایص موجود در آن به‌صورت مناسب و درست انجام شود، می‌تواند یک خوشه‌بندی باکیفیت، کارا، اثربخش، و با دقت قابل قبول را که از اهداف یک خوش‌بندی خوب است، در سازماندهی داده‌های متنی و غیرمتنی در حجم داده‌های بزرگ به همراه داشته باشد.

در ادامـه، جهـت انجـام پژوهش‌هایـی در زمینـهٔ خوشـه‌بندی داده‌هـا در آینـده پیشـنهاداتی ارائـه می‌شـود:

- با توجـه بـه اینکـه در خوشـه‌بندی سـرعت و دقـت دو مؤلفـهٔ مهـم در ایـن فراینـد اسـت، پیشـنهاد می‌شـود پژوهش‌هایـی کـه الگوریتـم کا-میانـه را بـه شـکل اخـص بـا ایـن اهـداف بهبـود بخشـیده‌اند، بررسـی و تحلیـل شـوند؛
- یافته‌هـا نشـان داد کـه رفـع دو کاسـتی از کاسـتی‌های الگوریتـم کا-میانـه به‌صـورت همزمـان بـرای بهبـود ایـن الگوریتـم تاکنـون مـورد توجـه بـوده اسـت. بنابرایـن، پیشـنهاد می‌شـود تحقیقاتـی کـه در ایـن راسـتا انجـام شـده‌اند، بـا هـدف افزایـش کارایـی و کیفیـت خوشـه‌بندی بررسـی شـوند؛
- بـا در نظـر گرفتـن اینکـه غلبه بـر کاسـتی‌های گیرافتـادن در بهینهٔ محلـی و همگرایـی زودرس، حسـاس بـه داده‌هـای نویـزی و دورافتـاده، و حسـاس بـه ابعـاد بـالای ویژگـی به‌ترتیـب، کمتریـن مقبولیـت را در پژوهش‌هـای انجام‌شـده بـه خـود اختصـاص داده اسـت، پیشـنهاد می‌شـود بـه بررسـی دلایـل ایـن عـدم مقبولیـت بـرای هـر یـک از ایـن مـوارد پرداختـه شـود.

References

Afzali, M., & S. Kumar. 2019. *Text Document Clustering: Issues and Challenges.* Paper presented at the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)

Aggarwal, C. C. 2004. A human-computer interactive method for projected clustering. *IEEE transactions on knowledge and data engineering* 16 (4): 448-460.

Awawdeh, S., A. Edinat, & A. Sleit. 2019. An Enhanced K-means Clustering Algorithm for Multi-attributes Data. *International Journal of Computer Science and Information Security (IJCSIS)* 17 (2): 1-6.

Bansal, A., M. Sharma, & S. Goel. 2017. Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining. *International Journal of Computer Applications* 157 (6): 0975-8887.

Benabdellah, A. C., A. Benghabrit, & I. Bouhaddou. 2019. A survey of clustering algorithms for an industrial context. *Procedia Computer Science* 148: 291-302.

Bide, P., & R. Shedge. 2015. *Improved Document Clustering using k-means algorithm.* Paper presented at the 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). Coimbatore, India.

Chadha, A., & S. Kumar. 2014. *An improved K-means clustering algorithm: a step forward for removal of dependency on K.* Paper presented at the 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). Faridabad, India.

Chaturvedi, E. N., & E. A. Rajavat. 2013. An improvement in K-mean clustering algorithm using better time and accuracy. *International Journal of Programming Languages and Applications* 3 (4): 13-19.

Choudhary, A., P. Sharma, & M. Singh. 2016. *Improving K-means through better initialization and normalization.* Paper presented at the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Jaipur, India.

Fink, A. 2013. *Conducting research literature reviews: from the internet to paper*. *SAGE Publications*.

Fränti, P., & S. Sieranoja. 2019. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition* 93: 95-112.

Goswami, J. 2015. A Comparative Study on Clustering and Classification Algorithms. *International Journal of Scientific engineering and Applied Science (IJSEAS)* 1 (3): 2395-3470.

Goyal, M., & S. Kumar. 2014. Improving the initial centroids of K-means clustering algorithm to generalize its applicability. *Journal of the Institution of Engineers (India): Series B,* 95 (4): 345-350.

Han, J., M. Kamber, & J. Pei. 2012. Data mining: concepts and techniques. Waltham, MA: Morgan Kaufman Publishers, 10, 978-971.

Haraty, R. A., M. Dimishkieh, & M. Masud. 2015. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of distributed sensor networks* 11 (6): 615740.

Hotho, A., A. Nürnberger, & G. Paaß. 2005. *A brief survey of text mining.* Paper presented at the Ldv Forum.

Iezzi, D. F. 2012. A new method for adapting the k-means algorithm to text mining. *Italian Journal of Applied Statistics* 22 (1): 69-80.

Jaganathan, P., & S. Jaiganesh. 2013. *An improved K-means algorithm combined with particle swarm optimization approach for efficient web document clustering.* Paper presented at the 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE). Chennai, India.

Kant, S., & I. A. Ansari. 2016. An improved K means clustering with Atkinson index to classify liver patient dataset. *International Journal of System Assurance Engineering and Management* 7 (1): 222-228.

Karegowda, A. G., T. Vidya, M. Jayaram, & A. Manjunath. 2013. *Improving performance of k-means clustering by initializing cluster centers using genetic algorithm and entropy based fuzzy clustering for categorization of diabetic patients.* Paper presented at the Proceedings of International Conference on Advances in Computing. New Delhi, India.

Kaur, N., J. K. Sahiwal, & N. Kaur. 2012. Efficient k-means clustering algorithm using ranking method in data mining. *International Journal of Advanced Research in Computer Engineering & Technology* 1 (3): 85-91.

Khandare, A., & A. Alvi. 2016. Survey of Improved k-means Clustering Algorithms: Improvements, Shortcomings and Scope for Further Enhancement and Scalability. In *Information Systems Design and Intelligent Applications* (pp. 495-503) New Delhi, India: Springer.

Khatri, S. & K. Garg. 2016. Document Clustering Using Improved K-Means Algorithm. *International Journal of Engineering Research and General Science* 4 (3): 787-793.

Kim, H., H. K. Kim, & S. Cho. 2020. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications, 150*, 113288. doi: https://doi.org/10.1016/j.eswa.2020.113288.

Larose, D. T., & C. D. Larose. 2014. *Discovering knowledge in data: an introduction to data mining* (Vol. 4). Canada: John Wiley & Sons.

Linyao, X., & W. Jianguo. 2018. *Improved K-means Algorithm Based on optimizing Initial Cluster Centers and Its Application.* Paper presented at the 2018 Second International Conference of Sensor Network and Computer Engineering (ICSNCE 2018). Xi'an, China.

Liu, G., S. Huang, C. Lu, & Y. Du. 2014. An improved k-means algorithm based on association rules. *International Journal of Computer Theory and Engineering* 6 (2): 146.

Liu, Z., J. Bao, & F. Ding. 2018. *An Improved K-Means Clustering Algorithm Based on Semantic Model.* Paper presented at the Proceedings of the International Conference on Information Technology and Electrical Engineering 2018. Xiamen Fujian, China.

Ma, J. 2014. Improved K-Means Algorithm in Text Semantic Clustering. *The Open Cybernetics & Systemics Journal* 8 (1): 530-534.

Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. *Global Journal of Computer Science and Technology*.

Masud, M. A., M. M. Rahman, S. Bhadra, & S. Saha. 2019. *Improved k-means Algorithm using Density Estimation.* Paper presented at the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). India.

Na, S., L. Xumin, & G. Yong. (2010. *Research on k-means clustering algorithm: An improved k-means clustering algorithm.* Paper presented at the 2010 Third International Symposium on intelligent information technology and security informatics. Jian, China.

Napoleon, D., & P. G. Lakshmi. 2010. An enhanced k-means algorithm to improve the efficiency using normal distribution data points. *International Journal on Computer Science and Engineering* 2 (7): 2409-2413.

Okoli, C., & K. Schabram. 2010. A guide to conducting a systematic literature review of information systems research.https://dx.doi.org/10.2139/ssrn.1954824

Prabhu, P., & N. Anbazhagan. 2011. Improving the performance of k-means clustering for high dimensional data set. *International Journal on Computer Science and Engineering* 3 (6): 2317-2322.

Rajeswari, K., O. Acharya, M. Sharma, M. Kopnar, & K. Karandikar. 2015. *Improvement in K-means clustering algorithm using data clustering.* Paper presented at the 2015 International Conference on Computing Communication Control and Automation. Pune, India.

Rathore, P., & D. Shukla. 2015. *Analysis and performance improvement of K-means clustering in big data environment.* Paper presented at the 2015 International Conference on Communication Networks (ICCN).

Raval Unnati, R., & Chaita, J. (2016). Implementing & Improvisation of K-means Clustering Algorithm. *International Journal of Computer Science and Mobile Computing* 5: 191-203.

Saklecha, A., & J. Raikwal. 2017. Enhanced K-Means Clustering Algorithm Using Collaborative Filtering Approach. *Oriental Ournal of Computer Science & Technology*.10 (2): 474-479.

Shunye, W. 2013. *An improved k-means clustering algorithm based on dissimilarity.* Paper presented at the Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC). Shenyang, China.

Strech, D., & N. Sofaer. 2012. How to write a systematic review of reasons. *Journal of Medical Ethics* 38 (2): 121-126.

Taihao, L., N. Tuya, Z. Jianshe, R. Fuji, & L. Shupeng. 2020. An Improved K-Means Algorithm Based on Initial Clustering Center Optimization. *ZTE Communications* 15 (S2): 43-46.

Tajunisha, N., & V. Saravanan. 2011. An efficient method to improve the clustering performance for high dimensional data by principal component analysis and modified K-means. *Intl Journal of Database Mgt System* 3: 196-205.

Thilagaraj, T., & N. Sengottaiyan. 2019. Implementation of an Improved K-Means Clustering Algorithm for Balanced Clusters. *Pramana Research Journal* 9 (6): 352-360.

Tunali, V., T. Bilgin, & A. Camurcu. 2016. An Improved Clustering Algorithm for Text Mining: Multi-Cluster Spherical K-Means. *International Arab Journal of Information Technology (IAJIT)* 13 (1): 12-19.

Vashist, A., & R. Nath. 2016. Document Clustering using Improved K-means Algorithm. *International Journal of Research in Social Sciences* 6 (9): 193-204.

Wang, J., & X. Su. 2011. *An improved K-Means clustering algorithm.* Paper presented at the 2011 IEEE 3rd International Conference on Communication Software and Networks. Xi'an, China.

Wu, G., H. Lin, E. Fu, & L. Wang. 2015. *An improved k-means algorithm for document clustering.* Paper presented at the 2015 international conference on computer science and mechanical automation (CSMA). Hangzhou, China.

Xie, H., L. Zhang, C. P. Lim, Y. Yu, C. Liu, H. Liu, & J. Walters. 2019. Improving K-means clustering with enhanced firefly algorithms. *Applied Soft Computin,* 84: 105763.

Xinwu, L. 2012. A new text clustering algorithm based on improved K-means. *Journal of Software* 7 (1): 95-101.

Xiong, C., Z. Hua, K. Lv, & X. Li. 2016. *An Improved K-means text clustering algorithm By Optimizing initial cluster centers.* Paper presented at the 2016 7th International Conference on Cloud Computing and Big Data (CCBD). Macau, China.

Yadav, A., & S. Dhingra. 2016. An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters. *International Journal of Engineering Development and Research (IJEDR)* 4 (4): 901-907.

_____, A., & S. K. Singh. 2016. An Improved K-Means Clustering Algorithm. *International Journal of Computing* 5 (2): 88-103.

Yedla, M., S. R. Pathakota, & T. Srinivasa. 2010. Enhancing K-means clustering algorithm with improved initial center. *International Journal of computer science and information technologies* 1 (2): 121-125.

Yu, S.-S., S.-W. Chu, C.-M. Wang, Y.-K. Chan, & T.-C. Chang. 2018. Two improved k-means algorithms. *Applied Soft Computing* 68: 747-755.

Zhang, G., C. Zhang, & H. Zhang. 2018. Improved K-means algorithm based on density Canopy. *Knowledge-based systems* 145: 289-297.

Zhang, Y., K. Wang, H. Lu, H. Guo, & L. Xu. 2013. *An improved k-means clustering algorithm over data accumulation in Delay Tolerant Mobile Sensor Network.* Paper presented at the 2013 8th International Conference on Communications and Networking in China (CHINACOM). Guilin, China.

Zheng, L. 2020. Improved K-Means Clustering Algorithm Based on Dynamic Clustering. *International Journal of Advanced Research in Big Data Management System* 4: 17-26.

Zhu, J., & H. Wang. 2010. *An improved K-means clustering algorithm.* Paper presented at the 2010 2nd IEEE International Conference on Information Management and Engineering. Chengdu, China.

الهام یلوه

دانشجوی کارشناسی ارشد علم اطلاعات و دانش‌شناسی از دانشگاه قم است.

داده‌کاوی، متن‌کاوی و علم‌سنجی از جمله علاقه پژوهشی وی است.

یعقوب نوروزی

متولد سال ۱۳۵۱ دارای مدرک تحصیلی دکتری علوم کتابداری و اطلاع‌رسانی از دانشگاه آزاد واحد علوم و تحقیقات است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه قم است. کتابخانه‌های دیجیتالی، سازماندهی اطلاعات، نرم‌افزارهای کتابخانه‌ای و اطلاع‌رسانی از جمله علایق پژوهشی وی است.

اشکان خطیر

متولد ۱۳۶۴، دارای مدرک تحصیلی دکتری در رشته مهندسی فناوری اطلاعات از پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. تحلیل روند، متن‌کاوی و داده‌کاوی از جمله علایق پژوهشی وی است.