الگوریتم DBSCAN بهبودیافته برای داده‌های بزرگ پزشکی

حمید سعادت فر[۱]، نوشین حنفی[۲] و ناهید قلی زاده[۳]
۱استادیارگروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، saadatfar@birjand.ac.ir
۲دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه بیرجند، nooshinhanafi@birjand.ac.ir
۳دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه بیرجند، Nahid.gholizadeh@birjand.ac.ir

چکیده - در دهه اخیر تولید داده با رشد چشمگیری مواجه شده که این داده‌ها از منابع مختلف مانند دستگاه‌های تلفن همراه، شبکه‌های اجتماعی، شبکه‌های حسگر بی‌سیم و غیره در حال تولید می‌باشند. مدیریت این حجم زیاد از داده‌ها به یک چالش بزرگ در عصر حاضر تبدیل شده است. خوشه‌بندی داده‌ها به عنوان یک راه‌حل مطرح می‌شود که داده‌ها را براساس شباهتشان گروه‌بندی می‌کند. دو روش رایج در خوشه-بندی، الگوریتم‌های *DBSCAN* و *K-means* هستند. هر کدام از این روش‌ها مزایا و معایبی دارند که یکی از معایب سرعت اجرای نسبتا کند آن‌ها در داده‌های بزرگ است. ما در این مقاله الگوریتم *DBSCAN*بهبودیافته‌ای را ارائه داده‌ایم که بر مبنای دو روش بالا می‌باشد. نتایج نشان می‌دهد روش پیشنهادی ما در عین حفظ کیفیت از نظر سرعت بهتر از الگوریتم *DBSCAN* عمل می‌کند.

کلید واژه- الگوریتم DBSCAN، الگوریتم K-means، خوشه‌بندی، داده‌های بزرگ

۱-مقدمه

امروزه با رشد اینترنت و شبکه‌های اجتماعی با حجم زیادی از داده‌ها مواجه هستیم، به همین دلیل سیستم‌ها و الگوریتم‌های سنتی نمی‌توانند در زمان‌های قابل قبول پاسخگو باشند. الگوریتم‌های سنتی یادگیری ماشین نیز از این قاعده مستثنی نبوده و قابل اجرا بر روی داده‌های بزرگ با استفاده از یک ماشین تک پردازنده نمی‌باشند. بنابراین روش‌هایی در دنیا وجود دارد که الگوریتم‌ها را بهینه‌تر کرده و سرعت اجرای آن‌ها را بالاتر می‌برد.

یک دسته از روش‌های یادگیری ماشین، خوشه‌بندی[۱] [۲] می‌باشد که هدف آن تشخیص وجود خوشه‌ها و گروه‌ها در یک مجموعه داده است. خوشه‌بندی در زمینه‌های گسترده‌ای از جمله تشخیص الگو [۳]، داده‌کاوی[۴]،فشرده‌سازی داده [۵] [۶] [۷]، آنالیز شبکه‌هایاجتماعی وب[۸] و شبکه‌های حسگر بی-سیم [۹] مورد استفاده قرار می‌گیرد. یک متد خوشه‌بندی خوب، داده‌های مشابه را در یک خوشه یکسان و داده‌های غیر مشابه را در خوشه‌های مختلف قرار می‌دهد.

برای تکنیک خوشه‌بندی، الگوریتم‌های مختلفی ارائه شده است که می‌توان از الگوریتم‌های DBSCAN و K-means به عنوان دو روش رایج در این زمینه نام برد.[۱۰]

DBSCAN برای داده‌های بزرگ بسیار کند است و پیچیدگی بالایی دارد به همین دلیل برای داده‌های بزرگ نمی-توان از آن به تنهایی استفاده کرد. K-means نیز یکی از پرکاربردترین الگوریتم‌های شناخته شده مبتنی بر مرکزیت است که تلاش می‌کند مجموع مربعات فواصل اقلیدسی را از مرکز هر خوشه حداقل کند.

هدف مقاله این است که گروه‌بندی داده‌های بزرگ در کمترین زمان و به بهترین شکل صورت گیرد. از آن جایی که در DBSCAN محلیت و چگالی داده‌ها اهمیت دارد و در ایجاد خوشه‌ها نیز محلیت مهم است بنابراین نیازی نیست که فاصله با نقاط دور محاسبه شود. طی روش پیشنهادی ما ابتدا داده‌ها با استفاده از K-means و با یک تکرار مشخص خوشه‌بندی می-شوند سپس DBSCAN در هر کدام از خوشه‌های ایجاد شده به صورت جداگانه اعمال می‌شود.

آزمایشات نشان می‌دهد زمان اجرای الگوریتم DBSCAN بهبودیافته نسبت به DBSCAN معمولی به طور چشمگیری کاهش پیدا کرده است.

این مقاله به صورت زیر ساختاربندی شده است. بخشی از مقالاتی که در زمینه خوشه‌بندی داده‌های بزرگ با استفاده از روش‌های K-means و DBSCAN نوشته شده در بخش ۲ آورده شده است. بخش ۳ مقاله به بررسی روش پیشنهادی می‌پردازد.

در بخش ۴ نتایج حاصل از آزمایشات مورد بررسی قرار گرفته و بخش پایانی شامل نتیجه‌گیری است.

۲-پیشینه تحقیق

در سال‌های اخیر الگوریتم‌های خوشه‌بندی بی‌شماری برای داده‌های بزرگ ارائه شده است. در حالت کلی این الگوریتم‌ها را می‌توان در دو دسته قرار داد[۱۶]. دسته‌ای از الگوریتم‌ها که بر روی یک ماشین اجرا شده و دسته دیگر که روی چند ماشین اجرا می‌شوند.

H-K-mean[۱۱] جزو الگوریتم‌هایی است که بر روی یک ماشین اجرا می‌شود و در گروه تکنیک‌های کاهش داده قرار می‌گیرد. در این الگوریتم داده‌ها در یک ساختار سلسله مراتبی کاهش پیدا می‌کنند. به این صورت که مراکز هر خوشه به عنوان نماینده داده‌های آن خوشه به سطح بعد انتقال پیدا می‌کنند.

یکی دیگر از مقالاتی که از تکنیک کاهش داده بهره برده است الگوریتم Tri-level K-means می‌باشد[۱۲]. این روش، K-means را با تکرار محدود اجرا می‌کند و سپس با استفاده از یک معیار، کلاسترهای بزرگ را شناسایی کرده و آن ها را به کلاسترهای کوچک‌تر تقسیم می‌کند. به این ترتیب مراکز اولیه هوشمندانه‌تری انتخاب خواهد شد.

ایده کاهش محاسبات در مقاله K^2-means مطرح شده[۱۳] که در آن محاسبات انجام شده توسط الگوریتم K-means کاهش یافته است. به این صورت که در فاز تخصیص الگوریتم K-means، به جای محاسبه فاصله هر نقطه تا کلیه مراکز موردنظر، فقط فاصله نقاط تا مراکز خوشه‌های همسایه محاسبه می‌شود که باعث کاهش محاسبات می‌گردد.

مقاله [۱۴] جزو مقالاتی می‌باشد که از تکنیک مبتنی بر چند ماشین بهره برده است. در این مقاله مجموعه داده به بخش‌های کوچکتری تقسیم شده و بین چندین گره در یک خوشه از ماشین‌ها توزیع می شود. آپاچی Hadoop به عنوان یک پلتفرم مقیاس‌پذیر و قدرتمند برای این منظور مورد استفاده قرار می‌گیرد. K-means روی ماشین‌ها در دو فاز اجرا می‌شود که در فاز اول با انتخاب تعداد خوشه‌های زیاد خوشه‌های اولیه ایجاد می‌شود و در فاز دوم مراکز به دست آمده از فاز اول، طبق فرمول مطرح شده در این مقاله با هم ادغام می‌شوند.

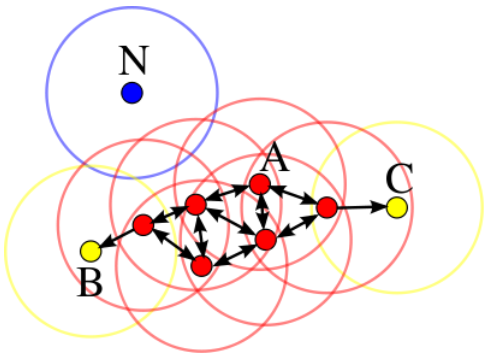
Li Ma و همکاران در مقاله[۱۵]روش MRG-DBSCAN را ارائه کردند که در آن اجرای الگوریتم DBSCAN و تولید نقاط مرکزی با استفاده از تکنیک Map Reduce انجام می‌شود.

۳-روش ارائه شده

خوشه‌بندی یک دیتاست بزرگ توسط الگوریتم‌های داده‌کاوی شناخته شده، زمان‌بر می‌باشد بنابراین با توجه به افزایش حجم داده‌ها و کاهش توان الگوریتم‌ها در پردازش این حجم از داده‌ها، نیاز به ارائه متدهای جدید بیش از پیش احساس می‌شود. الگوریتم بیان شده در این مقاله با هدف افزایش سرعت خوشه‌بندی داده‌های بزرگ در عین حفظ کیفیت خوشه‌بندی ارائه شده است. الگوریتم پیشنهادی در این مقاله بهبود یافته الگوریتم مشهور DBSCAN می‌باشد.

الگوریتم DBSCAN جزو الگوریتم‌های مبتنی بر چگالی است که وجود نویزها در داده‌های اصلی را به خوبی تشخیص می‌دهد. این روش نقاط را به سه گروه طبقه‌بندی می‌کند: نقاط core، density-reachable و نویز. یک نقطه،core درنظرگرفته می‌شود، در صورتی که در فاصله ε از آن به اندازه minpts نقطه وجود داشته باشد (با احتساب خود نقطه core) که این نقاط به صورت مستقیم از طریق نقاط core قابل دستیابی‌اند. سایر نقاط هر خوشهDBSCAN نقطه density-reachable در نظر گرفته می‌شوند که این نقاط به صورت غیرمستقیم به نقاط core متصل هستند. به طبع نقاطی که نتوانند به صورت مستقیم یا غیرمستقیم به نقاط core بپیوندند، خارج از کلاسترها قرار گرفته و به عنوان نویز درنظرگرفته می‌شوند.

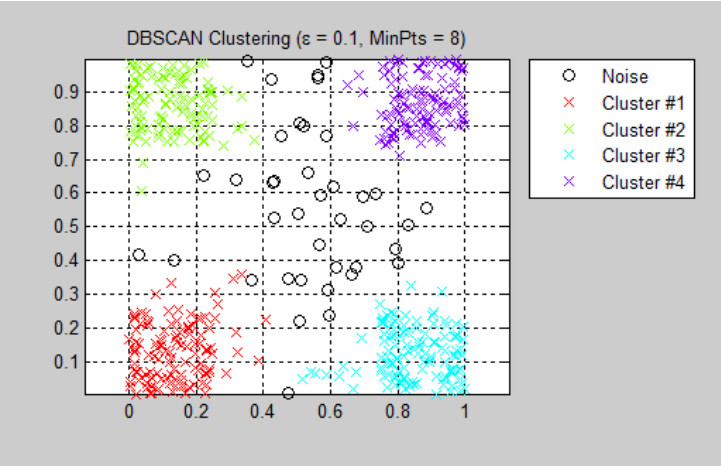
در شکل ۱، minPts=4 است. نقطه A و دیگر نقاط قرمز رنگ اطراف آن به عنوان نقاط core شناخته شده‌اند و نقطه N نیز نویز تشخیص داده شده است.



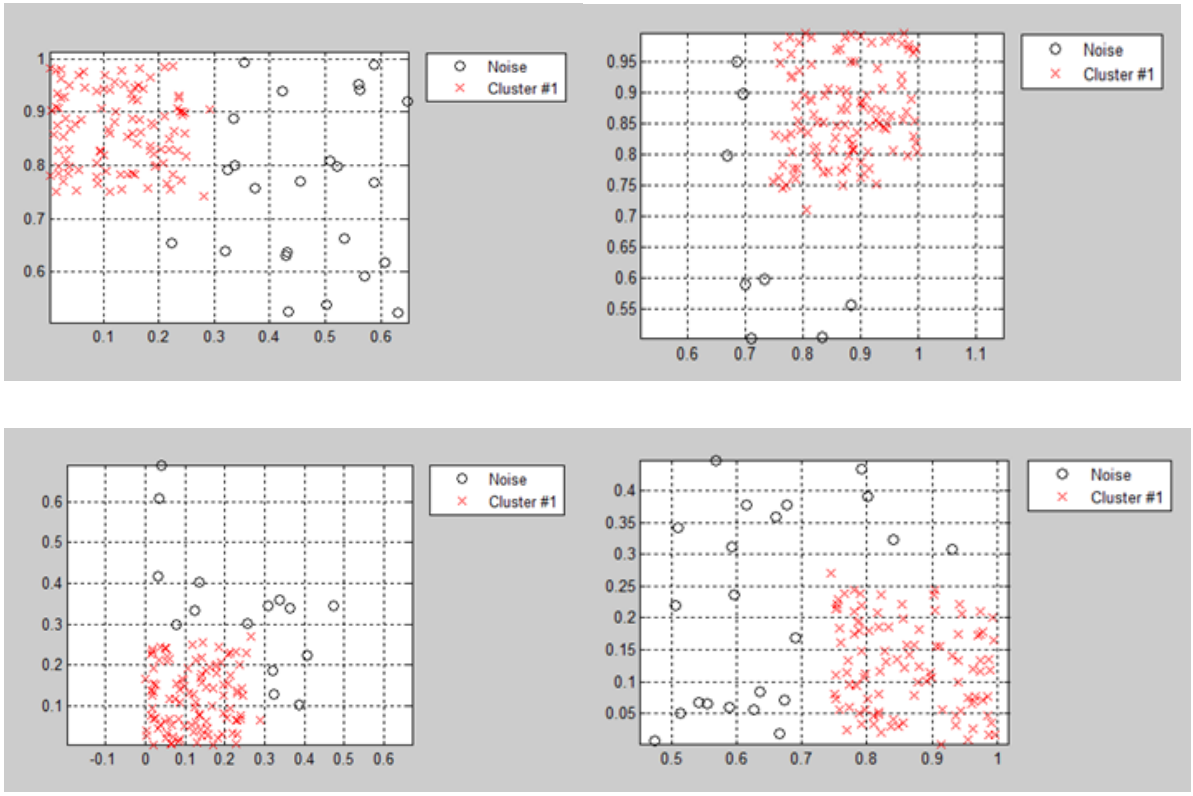
شکل ۱: نمایش خوشه بندی داده ها با روش DBSCSN

این الگوریتم برای شناسایی نقاط core در یک خوشه و نقاط noise، نیازمند محاسبه فاصله هر نقطه تا تمامی نقاط دیگر است که این خود، افزایش حجم محاسبات، زمان اجرا و به دنبال آن کاهش سرعت، به خصوص برای داده‌های بزرگ را به همراه دارد. بنابراین برای افزایش سرعت DBSCAN بر روی داده‌های بزرگ،

از الگوریتم K-means در ابتـدای الگـوریتم DBSCAN اسـتفاده می‌کنیم. هدف از این کار این است که داده‌های نزدیک به هم تا



(a) خوشه‌بندی داده با اعمال روش DBSCAN



(b) ۴ خوشه تولید شده توسط الگوریتم DBSCAN بهبودیافته

شکل ۲: مقایسه نتیجه اجرای DBSCAN و DBSCAN بهبودیافته با دیتاست‌های یکسان، در شکل (a) و اشکال (b)

حد ممکن در یک خوشه قرار گیرند و الگـوریتم DBSCAN نقاط با فواصل دور را در محاسبات خود در نظـر نگیـرد. بـه ایـن ترتیب این الگوریتم در هر خوشه حاصـل شـده از K-means بـه صورت مجزا اجرا شده و نقاط core را تشخیص می‌دهد.

همان طور که در شکل ۲ مشخص اسـت بخـش (a) اجـرای DBSCAN معمـولی را نشـان مـی‌دهـد. امـا در بخـش (b) DBSCANروی خوشه‌های حاصل از K-meansاجرا شده است.

مراحل اجرای الگویتم به صورت گام‌های زیر بیان می‌شود:

گام ۱: انتخاب kنقطه اولیه تصادفی

گام ۲: اعمال الگوریتم K-meansبا تعداد تکرار محدود

گام ۳: اعمال الگوریتم DBSCAN بر روی هر یک از خوشه- های بدست آمده در گام قبل

نام مجموعه داده	تعداد نمونه	زمان حاصل از DBSCAN	زمان حاصل از DBSCAN بهبودیافته
DoctorAUS	۵۰۰۰	۶/۸۳۲۷	۸/۹۵۳۹
HI	۱۰۰۰۰	۱۴/۸۹۹۰	۱۲/۲۳۵۴
Contact with medical doctor	۲۰۰۰۰	۵۴/۲۳۳۱	۲۹/۱۸۷۶
VietNamI	۲۷۰۰۰	۷۶/۸۷۶۲	۴۳/۱۲۱۵

جدول ۱: مقایسه زمان¬های حاصل از اجرای دو الگوریتم DBSCAN و DBSCAN بهبودیافته بر روی مجموعه داده¬های مختلف

۴-تحلیل نتایج

این بخش برای گزارش و بررسی عملکرد الگوریتم بیان شده بر روی چندمجموعه دادهمختلف اختصاص پیدا کرده است. از سایت (https://vincentarelbundock.github.io) برای جمع-آوری مجموعه داده‌ها استفاده شده است. این داده‌ها در رابطه با اطلاعات سلامت پزشکی می‌باشند که به عنوان داده‌های تست در آزمایش مورد استفاده قرار گرفتند.

برای گرفتن نتیجه بهتر و مقایسه راحت‌تر، سعی شده از مجموعه داده‌هایی با تعداد نمونه‌های مختلف استفاده گردد. جدول ۱ شامل ۴ مجموعه داده با تعداد نمونه‌های متفاوت است که زمان اجرای الگوریتم DBSCAN عادی و DBSCAN بهبودیافته در آن درج شده است که به راحتی قابل مقایسه می‌باشند.

مشخصات مجموعه داده‌هایی که در جدول ۱ مورد استفاده قرار گرفته‌اند در زیر خلاصه شده است:

مجموعه‌دادهDoctorAUS درباره تعداد مراجعه‌های بیماران به دکتر در استرالیا می‌باشد که شامل ویژگی‌های درآمد سالانه و تعداد بیماری‌ها در دو هفته گذشته است.

مجموعه‌داده HIکه شامل ویژگی‌هایدرآمدو ساعات کار در هفته است شامل اطلاعات بیمه سلامت افراد می‌باشد.

در مجموعه‌داده Contact with medical doctor دو ویژگی درآمد سالانه خانواده و تعداد بیماری‌های مزمن مدنظر قرار گرفته شده است. این مجموعه داده نیز اطلاعات مراجعه و تماس با پزشک را در اخیار ما قرار می‌دهد.

مجموعهداده VietNamIبیانگر هزینه‌های پزشکی در ویتنام است که ما دو ویژگی جنسیت و سن سرپرست خانه را مدنظر قرار داده‌ایم.

لازم به ذکر است که در محاسبات انجام شده برای هر دو روش DBSCAN و DBSCANبهبودیافته، مقادیر ε و minpts به صورت یکسان در نظر گرفته شده است:

Minpts= 10

$\varepsilon = 0.06$

۵- نتیجه‌گیری

مدیریت کردن حجم زیاد داده‌ها که امروزه با آن روبه‌رو هستیم به یک چالش بزرگ تبدیل شده است. خوشه‌بندی یک متد برای گروه‌بندی داده‌ها است که می‌تواند به عنوان راه‌حلی برای این چالش درنظرگرفته شود.

روش پیشنهادی این مقاله سعی بر این دارد که سرعت خوشه-بندی داده‌های بزرگ را افزایش دهد. برای این منظور یک روش بهبودیافته بر مبنای الگوریتم DBSCAN ارائه شده که به میزان قابل توجهی محاسبات را کاهش می‌دهد. نتایج حاصل از آزمایشات، در بخش قبلی ارائه و به طور کامل مورد بررسی قرار گرفته است. این آزمایشات نشان می‌دهد زمان اجرای الگوریتم DBSCAN بهبودیافته نسبت به DBSCAN رایج به طور قابل ملاحظه‌ای کاهش پیدا کرده است.

مراجع

[1] J. A.K, M. M.N and F. P.J, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, p. 264–323, 1999.

[2] P. Rai and S. Singh, "A survey of clustering techniques," *International Journal of Computer Applications*, vol. 7, no. 12, p. 1–5, 2010.

[3] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Elsevier, 2003.

[4] F. U.M, "Data mining and knowledge discovery: making sense out of data," *IEEE Expert*, vol. 11, no. 5, p. 20–25, 1996.

[5] A. Gersho and R. M.Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[6] J. Z. Lai, Y.-C. Liaw and W. Lo, "Artifact reduction of JPEG coded images using mean-removed classified vector quantization," *Signal Process*, vol. 82, no. 10, p. 1375–1388, 2002.

[7] Y.-C. Liaw, W. Lo and J. Z. Lai, "Image restoration of compressed image using classified vector quantizatio," *Pattern Recognition*, vol. 35, no. 2, p. 329–340, 2002.

[8] S. Qiao, T. Li, H. Li, J. Peng and H. Chen, "A new blockmodeling based hierarchical clustering algorithm for web social networks," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 3, p. 640–647, 2012.

[9] B. Baranidharan and B. Santhi, "DUCF: distributed load balancing unequal clustering in wireless sensor networks using fuzzy approach," *Applied Soft Computing*, vol. 40, p. 495–506, 2016.

[10] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise, Portland, Oregon," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[11] T.-S. Xu, H.-D. Chiang, G.-Y. Liu and C.-W. Tan, "Hierarchical K-means method for clustring large-scale advanced metering infrastructure data," *IEEE Transactions on Power Delivery*, vol. 32, no. 2, pp. 609-616, 2015.

[12] S.-S. Yu, S.-W. Chu, C.-M. Wang and Y.-K. Chan, "Two improved K-means algoritms," *Applied Soft Computing 68*, pp. 747-755, 2018.

[13] E. Agustsson, R. Timofte and L. V. Gool, "k2-means for Fast and Accurate Large Scale Clustering," in *Lecture Notes in Computer Science*, vol. 10535, Springer, Cham, 2017, pp. 775-791.

[14] A. Sinha and P. K. Jana, "A Novel K-Means based Clustering Algorithm for Big Data," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India, 2016.

[15] L. Ma, L. Gu, B. Li, S. Qiao and J. Wang, "MRG-DBSCAN: An Improved DBSCAN Clustering Method Based on Map Reduce and Grid," *International Journal of Database Theory and Application*, vol. 8, no. 2, pp. 119-128, 2015.

[۱۶] ش. عباسی و ب. وزیری، "الگوریتمهای خوشه بندی در داده های عظیم،" در *کنفرانس بین المللی پژوهش های کاربردی در فناوری اطلاعات، کامپیوتر و مخابرات*، ۱۳۹۴.