مجلـه‌ی علمی پژوهش در علوم رایانه، شماره ۲۱، بهار ۱۴۰۰، ص ۱۰۰ – ۹۰

ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال ششم)
شماره ۲۱ / بهار ۱۴۰۰ / ص ۱۰۰ – ۹۰

خوشه بندی داده های ساختار یافته با استفاده از درخت براساس وزن

محمد حسین زارع[۱]، علیرضا دهقانی[۲]

[۱] دانشجوی کارشناس ارشد مهندسی کامپیوتر گرایش نرم افزار.
[۲] گروه مهندسی کامپیوتر، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران.

نام نویسنده مسئول:
علیرضا دهقانی

تاریخ دریافت: ۱۳۹۹/۱۱/۴
تاریخ پذیرش: ۱۴۰۰/۱/۱۶

چکیده

استفاده از داده کاوی در آموزش و پرورش یک زمینه تحقیق بین رشته ای در حال ظهور است که در سال های اخیر توجه زیادی به آن شده است. باتوجه به الگوریتم های خوشه بندی در داده کاوی می‌توان با استفاده از نمرات و امتیازات دانش آموزان آنها را خوشه بندی و سایر دانش آموزان را باتوجه به داشتن ویژگی های مشابه و مشترک خوشه بندی نمود که این امر در آموزش وپرورش امری ضروری است. باتوجه به مفاهیم داده کاوی هدف این پایان نامه خوشه بندی داده های ساختار یافته با استفاده از درخت وزن دار است که برای این کار باتوجه به دادهای نیمه ساختاریافته ابتدا داده ها تبدیل به ساختاریافته می‌شوند سپس داده ها پاکسازی می‌شوند. در مرحله بعد از الگوریتم نظارت شده ***KNN*** برای خوشه بندی داده ها استفاده شده و با الگوریتم ***RF*** جهت نشان دادن کارایی الگوریتم پایه مقایسه شده است، سپس با استفاده از الگوریتم درخت پوشای کمینه(الگوریتم پریم) خوشه بندی ***KNN*** بهینه شد. جهت ارزیابی از معیارهای دقت، فراخوانی و ***RMSE*** استفاده شده است که نتایج خوبی را به همراه داشته است.

واژگان کلیدی: داده کاوی، الگوریتم خوشه بندی، الگوریتم درخت.

ISSN: 2745-3673
http://www.Rcsj.ir

بیان مساله

داده کاوی[1]، کشف اطلاعات پیش بینی پنهان از مجموعه داده های بزرگ است و این یک فناوری جدید قدرتمند و دارای پتانسیل بسیار بالا برای کمک به شرکت‌ها است تا در مهمترین اطلاعات در انبارهای داده خود تمرکز داشته باشند. نرم افزار داده کاوی یکی از تعدادی ابزار تحلیلی برای تجزیه و تحلیل داده‌ها است. این اجازه می‌دهد تا کاربران داده‌ها را از ابعاد مختلف، طبقه بندی آن و خلاصه روابط شناسایی و تجزیه و تحلیل کنند. از نظر فنی، داده کاوی فرایند یافتن همبستگی‌ها یا الگوهای در بین ده‌ها زمینه در پایگاه داده‌های بزرگ ارتباطی است*[۱و۲]*.

داده کاوی برای استخراج اطلاعات مفید از مجموعه داده های بزرگ و نمایش آن در تصویری آسان برای تفسیر استفاده می شود*[۳]*. داده‌کاوی را برخی علم استخراج اطلاعات از داده‌های موجود در دیتابیس نیز می‌نامند. به طور کلی، با استفاده از داده‌کاوی می‌توان از داده‌های خام موجود ذخیره شده که غالباً تحت عنوان داده بزرگ[2] شناخته می‌شوند، اطلاعات ارزشمند و مفیدی استخراج نموده و به نیازهای کسب‌وکارهایی که بقایشان منوط به دیتا است با سرعت بیشتری پاسخ داد چرا که در غیر این صورت، برای پاسخ‌دهی به چنین نیازهایی می‌بایست در میزان زیادی داده‌های خام جستجو نموده که کاری بسیار زمان‌بر و البته بی‌دقت خواهد بود*[۴]*. داده کاوی همچنین به عنوان کشف دانش در داده ها شناخته می شود، به یافتن یا "استخراج" دانش از مقادیر زیادی از داده ها اطلاق می شود. از تکنیک های داده کاوی برای کار بر روی حجم زیادی از داده ها برای کشف الگوهای پنهان و روابط مفید در تصمیم گیری استفاده می شود. بسیاری از افراد از اصطلاح "کشف دانش در داده[3]" یا KDD برای داده کاوی استفاده می کنند[۵].

یادگیری ماشینی[4] زیر مجموعه ای از داده کاوی است که تکنیک های خودکار برای یادگیری را برای پیش بینی های دقیق براساس مشاهدات گذشته مطالعه می کند. یادگیری ماشین از دو نوع تکنیک استفاده می کند: یادگیری نظارت شده (طبقه بندی و رگرسیون)، که یک مدل را بر روی داده های ورودی و خروجی شناخته شده آموزش می دهد تا بتواند خروجی های آینده را پیش بینی کند، و یادگیری بدون نظارت (طبقه بندی[5]) که الگوهای پنهان یا ساختارهای ذاتی را در ورودی می یابد.

طبقه بندی داده‌ها یکی از مهمترین زمینه های یادگیری ماشین است که تلاش می کند ساختار اساسی یک مجموعه داده را کشف کند. به طور کلی، ساختار به این معنی است که نمونه های مشابه به یک خوشه یکسان اختصاص می یابد در حالی که نمونه های متفاوت به خوشه های مختلف اختصاص داده می‌شوند*[۶]*. عدم دانش قبلی باعث می شود که تجزیه و تحلیل خوشه ای یک مشکل بسیار چالش برانگیز باقی بماند اگرچه بسیاری از الگوریتم‌های خوشه بندی در ادبیات ارائه شده اند. هر الگوریتم خوشه بندی استراتژی خاص خود را برای کشف یک ساختار از یک مجموعه داده دارد. الگوریتم های مختلف یا پارامترهای مختلف برای یک الگوریتم ممکن است منجر به نتایج خوشه‌ای مختلف شود. قضاوت در مورد اینکه کدام ساختار به بهترین وجه با توزیع واقعی مطابقت دارد بدون داشتن اطلاعات نظارت، دشوار است. بنابراین، انتخاب یک الگوریتم مناسب کار دشواری است. برای جلوگیری از این کار، بسیاری از تحقیقات بر ادغام نتایج خوش‌های چندگانه تمرکز دارند، که به عنوان گروه خوشه بندی شناخته می شود [۶]. این گروه خوشه بندی می تواند به طور قابل توجهی استحکام، ثبات و کیفیت یک راه حل خوشه بندی را در مقایسه با یک الگوریتم خوشه بندی واحد بهبود بخشد. تکنیک گروه خوشه بندی به طور موثری برای انجام بسیاری از کارهای خوشه بندی، از جمله داده های طبقه بندی شده، داده ها با ابعاد بالا، داده های نویزدار، داده های زمانی، انتخاب ویژگی ها و غیره مورد استفاده قرار گرفته است*[۶]*.

در این پایان نامه، خوشه بندی داده های ساختار یافته بر پایه خوشه بندی KNN است که پس از آن با استفاده از الگوریتم درخت پوشا(الگوریتم پریم)، KNN بهبود داده می‌شود.

---

[1] Data mining
[2] Big data
[3] knowledge discovery from data
[4] Machin learning
[5] Clustering

مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۱، بهار ۱۴۰۰، ص ۱۰۰ – ۹۰
ISSN: 2745-3673
http://www.Rcsj.ir

اهمیت و ضرورت انجام تحقیق

در یک روش ترکیبی خوشه‌ای باید نتایج حاصل از خوشه بندی‌های متعدد را به یک پارتیشن سازگار که بیشتر شبیه به خوشه های پایه است بدون استناد به مجموعه داده های اصلی ترکیب کند. برخی محققان هم از نتایج خوشه بندی چندگانه و هم از ویژگی های اصلی به عنوان ورودی برای بهبود بیشتر عملکرد خوشه بندی استفاده می کنند [۷،۸]. استفاده ترکیبی از الگوریتم های خوشه بندی باعث کارایی و عملکرد بهتر نسبت به یک الگوریتم خوشه بندی به تنهایی دارد.

برای محاسبات مرکز هر خوشه در این الگوریتم‌ها، از روش‌های جستجوی بهینه استفاده خواهد شد.

اهداف

۱. خوشه بندی داده های ساختار یافته با استفاده از درخت

۲. بهینه سازی اندازه گیری فاصله با استفاده از لبه های وزن دار

داده کاوی یک روش سودمند و کارا برای تمامی سازمان ها وصنایع می باشد که باعث یافتن الگو، استخراج دانش از پایگاه های داده‌های بزرگ، طبقه بندی داده ها و غیره می‌شود.

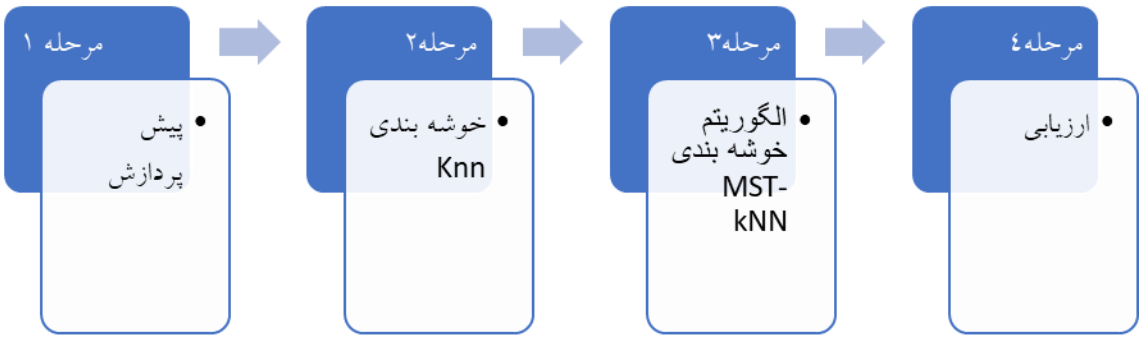
فرضیات:

۱. گره ها دارای وزن نمی‌باشند.

۲. از لبه های وزن دار برای اندازه گیری فاصله استفاده می‌شود.

۳. الگوریتم پایه، الگوریتم خوشه بندی KNN می‌باشد.

باتوجه به اینکه هدف این پایان نامه خوشه بندی داده های با استفاده از درخت براساس وزن است، روش پیشنهادی در 4 مرحله انجام می‌شود. در مرحله اول پیش پردازش جهت پاکسازی داده ها انجام می‌شود سپس داده های جدید جهت طبقه بندی به مرحله دوم ارسال می‌شود. در مرحله سوم از الگوریتم درخت پوشای کمینه جهت بهینه نمودن خوشه ها استفاده می‌شود. در نهایت در مرحله آخر ارزیابی انجام می‌شود که روش پیشنهادی با سایر الگوریتم ها مقایسه می‌شود. چارت پیشنهادی این پژوهش در شکل(۳-۱) نشان داده شده است.



شکل(۳-۱) شماء کلی روش تحقیق

مرحله ۱: پیش‌پردازش داده های خام

پیش پردازش داده ها یکی از مهمترین مراحل در فرآیند کشف دانش پیشرفته است. علیرغم اینکه کمتر از سایر مراحل مانند داده سنجی شناخته شده است، در واقع پیش پردازش داده ها اغلب شامل تلاش بیشتر و زمان در کل فرایند تجزیه و تحلیل داده ها است [۲۳]. داده های خام معمولاً با نقایص زیادی مانند ناسازگاری، ارزش از دست رفته، نویز و/یا افزونگی همراه است. بنابراین اگر آنها با داده های بی کیفیت ارائه شوند، عملکرد ردیف‌های یادگیری بعدی تضعیف می شود. بنابراین با انجام مراحل پیش پردازش مناسب، می توان کیفیت و قابلیت اطمینان اکتشافات و تصمیمات بعدی را تحت تأثیر قرار داد.

آماده سازی داده ها، به عنوان بخشی از پیش پردازش، با هدف تبدیل ورودی خام به ورودی با کیفیت بالا و مناسب فرآیند استخراج انجام می شود. آماده سازی به عنوان یک مرحله اجباری در نظر گرفته می شود و شامل تکنیک های یکپارچه سازی،

ISSN: 2745-3673
http://www.Rcsj.ir

عادی سازی، تمیز کردن و تبدیل است. در حال حاضر، مقدار داده های تولید شده به دنبال ظهور پدیده *Big Data* در حال رشد تصاعدی است. در صورت استفاده از الگوریتم های استاندارد، کاهش پیچیدگی یک مرحله اجباری است. تکنیک های کاهش داده با انتخاب و حذف ویژگی های زائد و پر سر و صدا و یا نمونه ها، یا با گشودن فضاهای پیچیده ویژگی های پیوسته، این ساده سازی را انجام می دهند. این باعث می‌شود تا ساختار اصلی و معنی ورودی را حفظ شود، اما در عین حال می توان اندازه قابل کنترل تری را بدست آورد. آموزش سریعتر و قابلیت های تعمیم یافته بهبود یافته الگوریتم های یادگیری، و همچنین قابل درک بودن و تفسیر بهتر نتایج، از جمله مزایای کاهش داده است.

مرحله ۲: خوشه بندی داده ها با استفاده از الگوریتم نزدیکترین همسایه

الگوریتم نزدیکترین همسایه[۶](*KNN*) به عنوان طبقه بندی و خوشه بندی کننده استفاده می شود*[۲۴و۲۵]*. روش *KNN*، که به آن روش خوشه بندی سریع نیز گفته می شود، در مقایسه با سایر روش های خوشه بندی مانند خوشه بندی سلسله مراتبی یا *OPTICS* (نقاط مرتب سازی برای شناسایی ساختار خوشه بندی) بسیار ساده تر و کارآمدتر است. پیچیدگی الگوریتم می تواند به طور قابل توجهی بیشتر از نظر زمان و مکان کاهش یابد. این امکان را برای روش *KNN* فراهم می کند تا با تجزیه و تحلیل داده های بزرگ یا پردازش سریع انجام شود.

تکنیک *KNN* یک روش یادگیری مبتنی بر شباهت است که می تواند نشان داده شود برای انواع دامنه های بسیار فعال است. برای تعیین خوشه داده آزمایشی مشخص شده الگوریتم *KNN* نزدیکترین همسایگان را در میان داده های یادگیری جستجو می کند و سپس از خوشه های همسایگان k برای تعیین وزن به نامزدهای خوشه استفاده می کند. داده‌ها در یک فضای بعدی ارائه می شوند ، جایی که d تعداد صفات یا خصوصیاتی است که مشاهده شده است که با توجه به شباهت آن با بقیه نقاط داده ذخیره شده در مدل، توسط برخی از اقدامات داده های شبیه به هم طبقه بندی می شود.

الگوریتم سنتی *KNN* از فرمول فاصله اقلیدسی برای خوشه بندی داده ها در یک یا چند کلاس از پیش تعریف شده با توجه به صفات وخصوصیات آنها استفاده می کند. بنابراین، الگوریتم از خصوصیات انتخاب شده که با خصوصیات داده های جدید مقایسه شده اند، استفاده می کند. یعنی اینکه یک مجموعه خوشه بندی شده $\{xi\}$ وجود داشته باشد، و لازم باشد داده جدیدی را خوشه بندی کند، الگوریتم *KNN* به دنبال خصوصیات K در مجموعه است که ممکن است همسایه داده y باشد، اگر داده y شامل خصوصیات K باشد، با استفاده از فاصله اقلیدسی شباهت خصوصیات خوشه K با خصوصیات y سنجیده می‌شود و در صورت داشتن فاصله مناسب در $\{xi\}$ خوشه بندی می‌شود. فرمول فاصله اقلیدسی فاصله بین دو مبحث با طبقه بندی p *(x,x2)* و *q(a,b)*در معادله زیر نشان داده شده است.

$$y = d(p, q) = d(q, p) = \sqrt{(x_1 - a)^2 + (x_2 - b)^2}$$

معادله(۳-۱)

همچنین برای یافتن شباهت بین خوشه ها می‌توان از فاصله جاکارد؛ منهتن و مینکوفسکی استفاده می‌شود که در شکل (۳-۳) نشان داده شده است.

[6] K-Nearest Neighbors

ISSN: 2745-3673
http://www.Rcsj.ir

Manhattan distance	$\lvert(x_{i1}-x_{j1})+(x_{i2}-x_{j2})+\cdots(x_{in}-x_{jn})\rvert$ where $i = x_{i1}, x_{i2},\ldots x_{in}$ and $j = x_{j1}, x_{j2},\ldots x_{jn}$
Minkowski distance	$\left\lvert(x_{i1}-x_{j1})^p+(x_{i2}-x_{j2})^p+\cdots(x_{in}-x_{jn})^{\frac{1}{p}}\right\rvert$ where $i = x_{i1}, x_{i2},\ldots x_{in}$, $j = x_{j1}, x_{j2},\ldots x_{jn}$ and p can be defined as the positive integer
Jaccard coefficient	J(A,B) = (A∩B)/(AUB) where A and B are documents

شکل(۳-۳) توابع فاصله برای اندازه گیری شباهت بین مجموعه ها

خوشه بندی

خوشه بندی یک کار معمول داده کاوی است که معمولاً برای آشکار کردن ساختارهای پنهان شده در مجموعه های بزرگ داده استفاده می شود. مشکل خوشه بندی شامل یافتن گروه هایی از اشیا است، به گونه ای که اشیایی که در یک گروه قرار می گیرند مشابه هستند و اشیا در گروه های مختلف متفاوت نیستند. الگوریتم های خوشه بندی را می توان با توجه به پارامترهای مختلف طبقه بندی کرد. یک نوع خاص از الگوریتم هایی که می توان آنها را تشخیص داد، خوشه بندی مبتنی بر نمودار است. در الگوریتم های خوشه بندی مبتنی بر نمودار، مجموعه داده ها می توانند به صورت نمودار مدل شوند. در چنین گرافیکی، یک گره شی مجموعه داده را نشان می دهد، یک لبه پیوند دهنده بین جفت گره ها است. هر لبه هزینه هایی دارد که مربوط به فاصله بین دو گره است و با استفاده از اندازه گیری فاصله انتخاب شده محاسبه می شود. باتوجه به فصل پیشین در این فصل به پیاده سازی و نتایج آن پرداخته شده است.

مجموعه داده

مجموعه داده انتخاب شده مربوط به امتیازات دانش آموزان سال نهم متوسطه اول می‌باشد. نمونه ای از داده در شکل(۴-۱) نشان داده شده است.

کد دانش اموز	شاخه کاردانش			شاخه فنی و حرفه ای			شاخه نظری					
	رشته های زمینه صنعت	رشته های زمینه خدمات	رشته های زمینه کشاورزی	رشته های زمینه هنر	رشته های زمینه صنعت	رشته های زمینه خدمات	رشته های زمینه کشاورزی	رشته های زمینه هنر	رشته ادبیات و علوم انسانی	رشته علوم و معارف اسلامی	رشته ریاضی فیزیک	رشته علوم تجربی
10												
امتیاز کل	41.97	50.8	44.27	59.48	66.52	77.14	68.56	78.65	86.53	82.22	83.33	87.19
اولویت	12	10	11	9	8	6	7	5	2	4	3	1
20	45.49	54.1	46.44	67.36	68.86	74.15	69.58	80.74	85.44	83.12	85.18	88.75
	12	10	11	9	8	6	7	5	2	4	3	1
30	48.73	65.95	51.38	68.6	57.19	67.12	59.19	71.2	60.28	60.08	62.77	58.9
			3				2			1		
40	54.12	74.52	54.88	79.35	60.3	70.25	60.32	72.17	68.01	67.66	63.81	65.1
									1	2		
50	42.03	55.85	44.29	65.74	54.55	66.34	56.42	65.83	74.03	67.02	64.36	63.42
	10	7	9	5	8	3	6	4	1	2		
60	35.1	42.29	34.59	52.86	62.16	67.18	61.33	68.86	80.74	74.87	83.75	90.4
	11	10	12	9	7	6	8	5	3	4	2	1
70	45.36	56.5	47.09	62.69	61.03	68.91	62.39	71.17	76.24	70.02	73.67	73.9
	10	8	9	5	7	4	6	2	1	3		
80	49.62	75.74	52.69	70.89	50.08	63.41	52.21	64.83	48.94	51.49	50.54	46.67
			1				2					
90	46.45	59.96	48.24	65.76	63.2	75.47	64.62	77.02	78.04	68.5	72.29	7.59
	10	8	9	5	7		6	2	1	4		
91	48.26	71.04	49.35	77.31	42	53.18	42.23	58.85	52.86	51.16	41.85	42.85

شکل(۴-۱) نمونه ای از داده

باتوجه به نمونه مجموع امتیازات از ۱۰۰ است که از مجموع امتیازات دریافت شده از سوی دبیران، والدین، مشاوره و نمرات برای تعین رشته تحصیلی است. هدف خوشه بندی دانش آموزان با امتیازات مشابه در خوشه های تحصیلی است.

مراحل شبیه سازی و ارزیابی نتایج

پیش پردازش

مرحله اول باتوجه به فلوچارت(۳-۱) پیش پردازش است. همانطورکه در شکل(۴-۱) نشان داده شده است رشته تجربی به عنوان گروه ۱، رشته ریاضی گروه ۲، معارف گروه ۳، ادبیات گروه ۴، کاردانش گروه ۵ و فنی و حرفه‌ای گروه ۶ انتخاب شده‌اند.

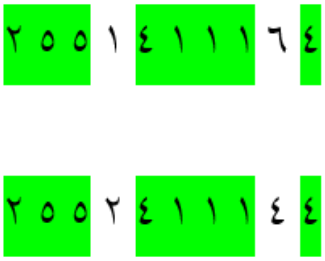
ISSN: 2745-3673
http://www.Rcsj.ir

یکی از مراحل پیش پردازش تمییز کردن داده ها است که کد دانش آموزانی که فاقد هر گونه اطلاعاتی باشد حذف می شود. داده ها پیش پردازش شده برای استفاده با فرمت جدید ذخیره می‌شود.

در مرحله بعد داده‌ها وارد می‌شوند. برای ورود داده‌ها از فایل ارسالی به نام *mdat.csv* که با فرمت *csv.* میباشد و حاوی داده اصلاح شده است استفاده نمایید. برای وارد سازی داده‌ها در محیط *R* مطابق دستور زیر عمل کنید. همچنین در این مرحله باتوجه به الگوریتم *KNN* که یک الگوریتم نظارت شده است داده ها را به دو مجموعه آموزشی[7] و آزمون[8] تقسیم می‌شود که برای این امر ٪۸۰ از مجموعه به عنوان آموزشی و ٪۲۰ باقی مانده به عنوان مجموعه آزمون برای تشخیص خوشه بندی انتخاب می‌شود.

خوشه بندی

در مرحله دوم الگوریتم *KNN* اجرا می‌شود که $K=6$ است و داده ها در ۶خوشه آموزش داده می‌شود. برای محاسبه فاصله بین داده ها از فاصله اقلیدوسی استفاده شده است. پس از آموزش، داده های آزمون جهت خوشه بندی به عنوان ورودی اجرا می‌شوند که نتایج آن در شکل(۴-۲) نشان داده شده است.



شکل(۴-۲) نتایج خروجی خوشه بندی *KNN* برای داده های آزمون

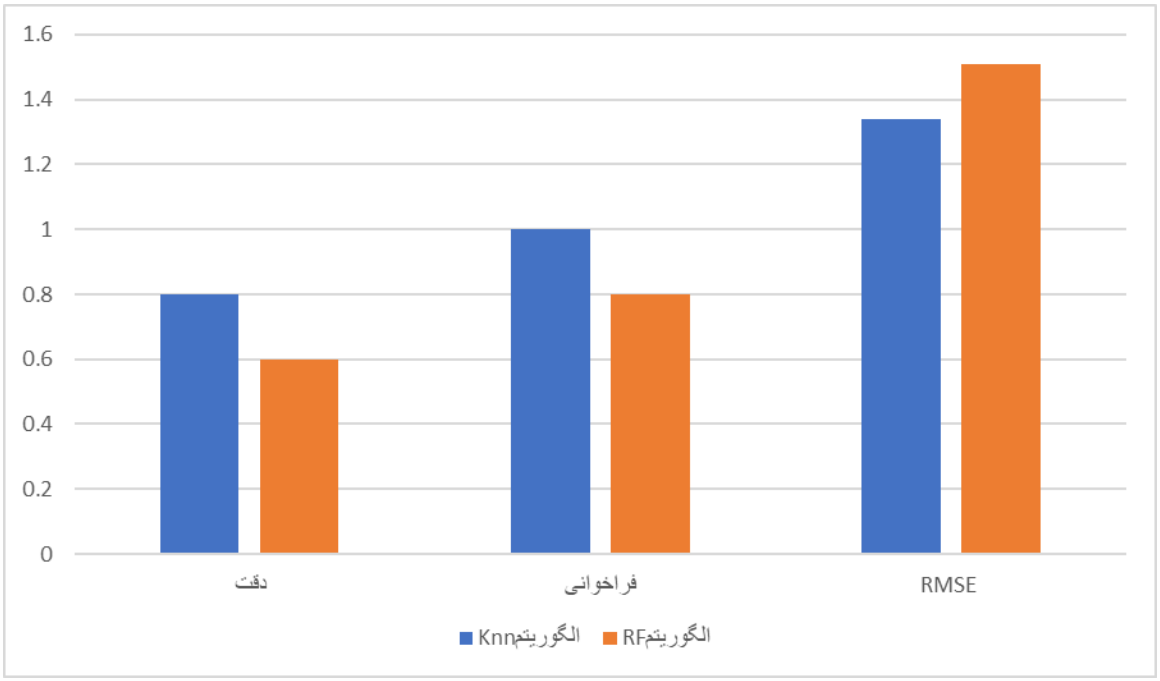
باتوجه به شکل فوق، ردیف اول خروجی پیش بینی خوشه بندی داده های آزمون می باشد که در مقایسه با ردیف دوم که خوشه بندی داده های واقعی است ۸مورد از ۱۰ مورد درست پیش بینی شده است. براین اساس دقت، فراخوانی و *RMSE* محاسبه و با الگوریتم *RF* مقایسه شده است که در جدول(۴-۱) و شکل (۴-۳) نشان داده شده است.

جدول(۴-۱) معیارهای ارزیابی برای الگوریتم خوشه بندی *Knn* و *RF*

	الگوریتم RF	الگوریتم Knn
دقت	0.6	0.8
فراخوانی	0.8	1
RMSE	1.51	1.34

[7] training
[8] test

ISSN: 2745-3673
http://www.Rcsj.ir



شکل(۴-۳) ارزیابی الگوریتم خوشه بندی *Knn* و *RF*

مرحله سوم: بهینه سازی خوشه بندی با *MTS*

در این مرحله باتوجه به مرحله سوم فصل پیشین اجرا می‌شود که در ابتدا باید یک ماتریس مجاورت براساس فاصله بین داده ها که امتیازات دانش آموزان است، ایجاد شود که با کد دستوری شکل(۴-۴) نشان داده شده است.

```
d <- base::as.matrix(stats::dist(mdat[,۱:۱۲], method="euclidean"))
```

شکل(۴-۴) کد دستوری ایجاد ماتریس

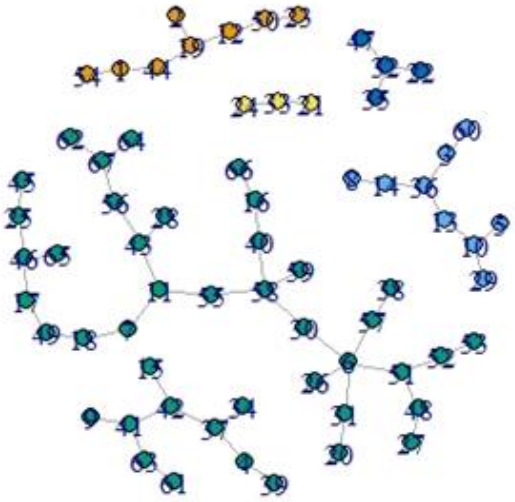
بخشی از خروجی این بخش در به شکل (۴-۵) نشان داده شده است:

۱۰ ۲۰ ۳۰ ۴۰ ۵۰ ۶۰ ۷۰ ۹۰ ۱۰۰ ۱۱۰ ۱۲۰ ۱۳۰ ۱۴۰ ۱۵۰ ۱۶۰ ۱۹۰ ۲۰۰ ۲۱۰ ۲۲۰ ۲۳۰ ۲۴۰ ۲۵۰
۱ ۱ ۲ ۳ ۲ ۳ ۳ ۲ ۳ ۲ ۳ ۱ ۲ ۲ ۳ ۳ ۳ ۳ ۱ ۳ ۴ ۵
۲۶۰ ۲۷۰ ۲۸۰ ۲۹۰ ۳۱۰ ۳۲۰ ۳۵۰ ۳۶۰ ۳۷۰ ۳۸۰ ۳۹۰ ۴۰۰ ۴۱۰ ۴۲۰ ۴۳۰ ۴۵۰ ۴۶۰ ۴۷۰ ۴۸۰ ۵۰۰ ۵۲۰ ۵۳۰
۱ ۴ ۳ ۳ ۳ ۳ ۲ ۳ ۳ ۵ ۴ ۳ ۵ ۲ ۳ ۳ ۳ ۳ ۳ ۳ ۳ ۱
۵۵۰ ۵۷۰ ۵۸۰ ۶۰۰ ۶۱۰ ۶۵۰ ۶۶۰ ۶۷۰ ۶۸۰ ۶۹۰ ۷۰۰ ۷۱۰ ۷۲۰ ۷۳۰ ۷۴۰ ۷۵۰ ۷۶۰ ۷۷۰ ۷۸۰ ۷۹۰ ۸۰۰ ۸۱۰
۳ ۳ ۵ ۳ ۳ ۱ ۳ ۳ ۳ ۱ ۳ ۳ ۳ ۳ ۳ ۲ ۳ ۳ ۳ ۳ ۳ ۳

شکل(۴-۵) ماتریس مجاورت تشکیل شده

خوشه های ایجاد شده توسط الگوریتم *KNN* و *MST* در شکل(۴-۶) نشان داده شده است.

ISSN: 2745-3673
http://www.Rcsj.ir

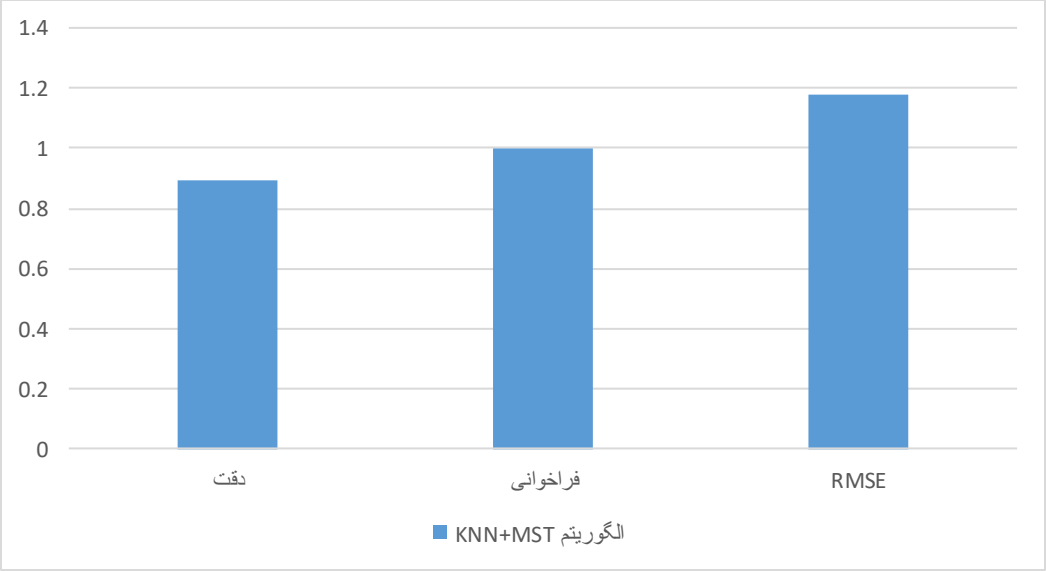


شکل(۴-۶) خوشه ایجاد شده توسط الگوریتم *KNN-MST*

در جدول(۴-۲) و شکل(۴-۷) دقت، فراخوانی و *RMSE* الگوریتم بهینه شده *KNN* با *MST* نشان داده شده است.

جدول(۴-۲) ارزیابی نتایج روش پیشنهادی

	الگوریتم *KNN+MST*
دقت	*0.89*
فراخوانی	*1*
RMSE	*1.18*



شکل(۴-۷) ارزیابی نتایج روش پیشنهادی

باتوجه به جدول فوق دقت و فراخوانی روش پیشنهادی خوب است چرا که توانسته با دقت ۰.۸۹٪ و فراخوانی ۱ و درصد خطای ۱.۱۸ خوشه ها را تشخصی و داده های جدید را با دقت بالایی خوشه بندی نماید.

ISSN: 2745-3673
http://www.Rcsj.ir

نتیجه گیری

میزان داده های نگهداری شده در قالب الکترونیکی در مدت اخیر افزایش چشمگیری داشته است. میزان اطلاعات هر ۲۰ ماه دو برابر می شود و تعداد پایگاه های اطلاعات با سرعت بیشتری افزایش می یابد. جستجو برای تعیین روابط معنی دار بین متغیرها در داده ها به روندی آهسته و ذهنی تبدیل شده است. به عنوان یک راه حل احتمالی برای این مشکل ، مفهوم دانش دانش در پایگاه داده - KDD ظهور کرده است. فرآیند شکل گیری مدل های قابل توجه و ارزیابی در KDD به عنوان داده کاوی شناخته می شود. داده کاوی برای کشف اطلاعات پنهان یا ناشناخته استفاده می شود که مشخص نیست ، اما به طور بالقوه مفید است.

مفهوم داده کاوی در سال‌های اخیر بسیار محبوب شده است. اگرچه درک منحصر به فردی از معنای داده کاوی وجود ندارد، اما به نظر می رسد تعریف زیر بیشتر و بیشتر مورد پذیرش قرار می گیرد: داده کاوی مفهوم همه روش ها و تکنیک ها است که به شما امکان می دهد مجموعه داده های بسیار بزرگ را برای استخراج و کشف ساختارهای قبلی ناشناخته و روابط از انبوه جزئیات زیاد را تجزیه و تحلیل کنید. این اطلاعات فیلتر شده، تهیه و طبقه بندی می شوند تا به عنوان یک کمک ارزشمند برای تصمیمات و استراتژی ها به کار روند. لیستی از تکنیک هایی که می تواند تحت چنین تعریفی در نظر گرفته شود شامل تحلیل لینک/ ارتباطات، الگوهای پی در پی، تجزیه و تحلیل سری های زمانی، طبقه بندی توسط درختان تصمیم گیرنده یا شبکه های عصبی، تجزیه خوشه تا مدل های امتیازدهی است. روش های خوشه بندی مجموعه ای از اشیا را به خوشه ها تقسیم می کند، از این رو اجسام موجود در یک خوشه با توجه به برخی از معیارهای تعریف شده شباهت بیشتری نسبت به سایر خوشه‌های مختلف دارند.

باتوجه به مفاهیم داده کاوی هدف این پایان نامه خوشه بندی داده های ساختار یافته با استفاده از درخت وزن دار است که برای این کار ابتدا ما از الگوریتم نظارت شده KNN برای خوشه بندی داده ها استفاده کردیم سپس با استفاده از الگوریتم درخت پوشای کمینه(الگوریتم پریم) خوشه بندی KNN را بهبود و بهینه کردیم. الگوریتم خوشه بندی MST می تواند خوشه های داده با مرزهای نامنظم را تشخیص دهد. ارزیابی الگوریتم فوق با استفاده از معیارهای دقت، فراخوانی و $RMSE$ انجام شد که برای الگوریتم بهینه نسبت به الگوریتم KNN عملکرد بهتری را داشته است.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] KS, D., & Kamath, A. (2017). Survey on Techniques of Data Mining and its Applications.
[2] Huang, R. J. P., Depari, G. S., Riorini, S. V., & Wang, P. C. (2018). Leveraging Social Media Metrics in Improving Social Media Performances through Organic Reach: A Data Mining Approach. *Review of Economic and Business Studies*, *11*(2), 33-48..
[3] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
[4] Jeffery W. Seifert , Analyst in information science and Technology Policy, ‘ Data
[5] Mining : An Overview ‘ December 2004.
[6] Sondwale, P. P. (2015). Overview of predictive and descriptive data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 262-265.
[7] Li, F., Qian, Y., Wang, J., Dang, C., & Jing, L. (2019). Clustering ensemble based on sample's stability. *Artificial Intelligence*, *273*, 37-55.
[8] Gupta, M. K., & Chandra, P. (2019, March). A comparative study of clustering algorithms. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 801-805). IEEE.
[9] Kim, T., Chen, I. R., Lin, Y., Wang, A. Y. Y., Yang, J. Y. H., & Yang, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in bioinformatics*, *20*(6), 2316-2326.
[10] Chitra, K., & Maheswari, D. (2017). A comparative study of various clustering algorithms in data mining. International Journal of Computer Science and Mobile Computing, 6(8), 109-115.
[11] Pascu, A. I. (2018). DATA MINING. CONCEPTS AND APPLICATIONS IN BANKING SECTOR. Annals of'Constantin Brancusi'University of Targu-Jiu. Economy Series, (1).
[12] Tharwat, A. (2020). Classification assessment methods. Applied Computing and Informatics.
[13] Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. Journal of Business Economics, 87(3), 317-335.
[14] Kumar, B. S. (2019). Data Mining: Clustering. Journal of the Gujarat Research Society, 21(14), 2021-2037.
[15] Harakawa, R., Takimura, S., Ogawa, T., Haseyama, M., & Iwahashi, M. (2019). Consensus Clustering of Tweet Networks via Semantic and Sentiment Similarity Estimation. IEEE Access, 7, 116207-116217.
[16] KADHIM, M. R., TIAN, W., & KHAN, T. (2019, December). Rapid Clustering with Semi-Supervised Ensemble Density Centers. In 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing (pp. 230-235). IEEE.
[17] Cai, Z., Yang, X., Huang, T., & Zhu, W. (2020). A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering. *Information Sciences*, *508*, 173-182.
[18] Gupta, M. K., & Chandra, P. (2019, March). A comparative study of clustering algorithms. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 801-805). IEEE.
[19] Kim, T., Chen, I. R., Lin, Y., Wang, A. Y. Y., Yang, J. Y. H., & Yang, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in bioinformatics*, *20*(6), 2316-2326.
[20] Al Alawi, M., Ray, S., & Gupta, M. (2019). A New Framework for Distance-based Functional Clustering.
[21] Gupta, M. K., & Chandra, P. (2020). An Empirical Evaluation of K-Means Clustering Algorithm Using Different Distance/Similarity Metrics. In *Proceedings of ICETIT 2019* (pp. 884-892). Springer, Cham.

ISSN: 2745-3673
http://www.Rcsj.ir

[22] Gupta, M. K., & Chandra, P. HYBCIM: Hypercube Based Cluster Initialization Method for k-means.
[23] Huang, T., Wang, S., & Zhu, W. (2020). An adaptive kernelized rank-order distance for clustering non-spherical data with high noise. *International Journal of Machine Learning and Cybernetics*, 1-13.
[24] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39-57.
[25] Wang, X., Yang, S., Zhao, Y., & Wang, Y. (2018). Lithology identification using an optimized KNN clustering method based on entropy-weighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. *Journal of Petroleum Science and Engineering*, *166*, 157-174.
[26] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273-292.