مطالعات علوم کاربردی در مهندسی
Online ISSN: ٢٤٧٦-٤٥٠٧
دوره ۳، شماره ۳، پاییز ۱۳۹۶
Print ISSN: ٢٥٣٨-٢٠٥٥
www.irijournals.com
صفحات ۳۱- ۲۳

بررسی و تبیین نقش و جایگاه روش های نوین الگوریتم ژنتیک برای خوشه بندی داده ها

میثم رهنمای فلاح[۱]، مرضیه فریدی ماسوله[۲*]، محمدرضا عسگری پور[۳]

[۱و۲و۳] دانشگاه آزاد اسلامی، واحد الکترونیکی، گروه کامپیوتر، تهران، ایران

[*] نویسنده مسئول مکاتبات

چکیده

خوشه‌بندی یا آنالیز خوشه در آمار و یادگیری ماشینی، یکی از شاخه های یادگیری بی‌نظارت می‌باشد و فرآیندی است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته ها خوشه گفته میشود؛ بنابراین خوشه مجموعه ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه می‌باشند. خوشه بندی تنها روش در یادگیری بدون نظارت است. یک خوشه به مجموعه ای از داده ها گفته می شود که با هم حداقل در یک صفت شباهت داشته باشند. در خوشه بندی سعی می شود تا داده ها به خوشه هایی تقسیم شوند که شباهت بین داده های درون هر خوشه حداکثر و شباهت بین داده های درون خوشه های متفاوت، حداقل شود. در این مقاله یک روش ترکیبی برای خوشه بندی براساس الگوریتم های ژنتیک ارائه کرده ایم، به طوری که الگوریتم پیشنهادی خود تعداد خوشه های بهینه را تشخیص داده و خوشه بندی را انجام دهد. نتایج شبیه سازی نشان می دهد که الگوریتم پیشنهادی با پیدا کردن تعداد خوشه های بهینه منجر به بهبود خوشه بندی داده ها می شود.

واژه‌های کلیدی: روش های نوین، الگوریتم ژنتیک، خوشه بندی، داده ها.

۱- مقدمه

خوشه بندی، تقسیم داده ها به گروه هایی از اشیا مشابه است که هر گروه، خوشه نامیده می شود، خوشه بندی یکی از شاخه های یادگیری بدون نظارت می باشد و فرآیند خودکاری است که در طی آن، نمونه ها به دسته هایی که اعضای آن مشابه یکدیگر می باشند تقسیم می شوند که اشیاء هر خوشه به یکدیگر شبیه بوده و نسبت به اشیاء دیگر خوشه ها شبیه نیستند. برای مشابه بودن می توان معیارهای مختلفی را در نظر گرفت مثلا می توان معیار فاصله را برای خوشه بندی مورد استفاده قرار داد و اشیائی را که به یکدیگر نزدیکتر هستند را بعنوان یک خوشه در نظر گرفت که به این نوع خوشه بندی، خوشه بندی مبتنی بر فاصله نیز گفته می شود (هالکیدی و همکاران[1]، ۲۰۰۱: زالیک، ۲۰۰۸). تکنیک خوشه بندی، تعداد زیادی از اشیاء داده ای را با تعداد کمی خوشه، نمایش می دهد، بنابراین این تکنیک، داده ها را با خوشه هایشان مدل می کند. به لحاظ مدلسازی داده ها ریشه های ایجاد تکنیک خوشه بندی، ریاضیات، آمار و آنالیزعددی می باشد. (راکش[2]، ۲۰۱۰)

خوشه بندی با طبقه بندی متفاوت است. در طبقه بندی نمونه های ورودی برچسب گذاری شده اند ولی در خوشه بندی نمونه های ورودی دارای بر چسب اولیه نمی باشند و در واقع با استفاده از روش های خوشه بندی است که داده های مشابه مشخص و بطور ضمنی برچسب گذاری می شوند. در واقع می توان قبل از عملیات طبقه بندی داده ها یک خوشه بندی روی نمونه ها انجام داد و سپس مراکز خوشه های حاصل را محاسبه کرد و یک بر چسب به مراکز خوشه ها نسبت داد و سپس عملیات طبقه بندی را برای نمونه های ورودی جدید انجام داد (پریتی[3]، ۲۰۱۱: دیوید[4]، ۲۰۰۴: میشل[5]، ۱۹۹۶). هدف خوشه بندی یافتن خوشه های مشابه از اشیاء در بین نمونه های ورودی می باشد اما چگونه می توان گفت که یک خوشه بندی مناسب است و دیگری مناسب نیست؟ می توان نشان داد که هیچ معیار مطلقی برای بهترین خوشه بندی وجود ندارد بلکه این بستگی به مساله و نظر کاربر دارد که باید تصمیم بگیرد که آیا نمونه ها بدرستی خوشه بندی شده اند یا خیر. با این حال معیار های مختلفی برای خوب بودن یک خوشه بندی ارائه شده است که می تواند کاربر را برای رسیدن به یک خوشه بندی مناسب راهنمایی کند. یکی از مسایل مهم در خوشه بندی انتخاب تعداد خوشه ها می باشد (هالکیدی و همکاران، ۲۰۰۱: راکش، ۲۰۱۰). در بعضی از الگوریتم ها تعداد خوشه ها از قبل مشخص شده است و در بعضی دیگر خود الگوریتم تصمیم می گیرد که داده ها به چند خوشه تقسیم شوند. در این مقاله یک روش ترکیبی براساس الگوریتم ژنتیک و الگوریتم K-mean ارائه شده است که خود الگوریتم تعداد خوشه های بهینه را تشخیص می دهد و براین اساس خوشه بندی را انجام می دهد.

۲- طبقه بندی روش های خوشه بندی

خوشه بندی انحصاری و خوشه بندی با هم پوشی

در روش خوشه بندی انحصاری پس از خوشه بندی، هر داده دقیقاً به یک خوشه تعلق می گیرد، مانند روش خوشه بندی K-Means. ولی در خوشه بندی با همپوشی پس از خوشه بندی، به هر داده یک درجه تعلق به ازا هر خوشه نسبت داده می شود. به عبارتی یک داده می تواند با نسبت های متفاوتی به چندین خوشه تعلق داشته باشد. برای مثال خوشه بندی فازی یک نوع خوشه بندی هم پوشی است.

خوشه بندی سلسله مراتبی و خوشه بندی مسطح

در روش خوشه بندی سلسله مراتبی، به خوشه های نهایی بر اساس میزان عمومیت آنها به ساختاری سلسله مراتبی نسبت داده می شود. در خوشه بندی مسطح تمامی خوشه های نهایی دارای یک میزان عمومیت هستند. به ساختار

[1] Halkidi et al.
[2] Rakesh
[3] Preeti
[4] David
[5] Mitchell

سلسله مراتبی حاصل از روشهای خوشه بندی سلسله مراتبی دندوگرام گفته می شود (هالکیدی و همکاران، ۲۰۰۱). با توجه با اینکه روش های خوشه بندی سلسله مراتبی اطلاعات بیشتر و دقیق تری تولید می کنند برای تحلیل داده های با جزئیات زیاد پیشنهاد می شوند ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند برای مجموعه داده های بزرگ روش های خوشه بندی مسطح پیشنهاد می شوند (هالکیدی و همکاران، ۲۰۰۱: راکش، ۲۰۱۰). روش های خوشه بندی بر اساس ساختار سلسله مراتبی تولیدی توسط آنها معمولا به دو دستهٔ زیر تقسیم می شوند (لاسزلو[1]، ۲۰۰۷):

۱. بالا به پایین

۲. پایین له بالا

بالا به پایین[2] یا تقسیم کننده[3] که در این روش ابتدا تمام داده ها به عنوان یک خوشه در نظر گرفته می شوند و سپس در طی یک فرایند تکراری در هر مرحله داده هایی که شباهت کمتری به هم دارند به خوشه های مجزایی شکسته می شوند و این روال تا رسیدن به خوشه هایی که دارای یک عضو هستند ادامه پیدا می کند.

پایین به بالا[4] یا متراکم شونده[5] که در این روش ابتدا هر داده ها به عنوان خوشه ای مجزا در نظر گرفته می شود و در طی فرایندی تکراری در هر مرحله خوشه هایی که شباهت بیشتری با یکدیگر دارند ترکیب می شوند تا در نهایت یک خوشه و یا تعداد مشخصی خوشه حاصل شود. از انواع الگوریتمهای خوشه بندی سلسله مراتبی متراکم شونده رایج می‌توان از الگوریتمهای Single Link، Average-Link و Complete-Link نام برد. تفاوت اصلی در بین تمام این روشها به نحوه محاسبهٔ شباهت بین خوشه ها مربوط می شود.

۳- روش پیشنهادی

در این بخش یک روش خوشه بندی داده براساس الگوریتم ژنتیک ارائه می شود. در ادامه مروری بر الگوریتم های ژنتیک خواهیم داشت و سپس یک الگوریتم ژنتیک برای خوشه بندی داده ها ارائه خواهد شد.

۳-۱- الگوریتم های ژنتیک

الگوریتم ژنتیک تکنیک جستجویی در علم کامپیوتر برای یافتن راه حل تقریبی برای بهینه سازی و مسائل جستجو است.

الگوریتم ژنتیک نوع خاصی از الگوریتم های تکامل است که بر اساس ساختار ژن ها و کروموزوم ها تشکیل شده است. این الگوریتم بر اساس اصل بقای بهترین ها است که طی آنها در یک ساختار تکاملی یک مجموعه تصادفی از جواب های اولیه مساله به تکامل می رسند و در واقع بهینه می شوند. الگوریتم ژنتیک در ابتدا با جمعیتی کاملا تصادفی آغاز می شود و در نسل ها ادامه می یابد. در هر نسل گنجایش تمام جمعیت ارزیابی می شود، چندین فرد مناسب در فرایندی تصادفی از نسل جاری انتخاب می شود (بر اساس شایستگی ها) و برای شکل دادن نسل جدید اصلاح می شود و در تکرار بعدی الگوریتم به نسل جاری تبدیل می شود. این الگوریتم با استفاده از عملگر های انتخاب، ترکیب و جهش ما را به حل بهینه ای می رساند که با روش های دیگر امکان پذیر نیست. این مراحل آنقدر تکرار می شود که به حل بهینه ای از جواب برسیم.

[1] Laszlo
[2] Top-Down
[3] Divisive
[4] Bottom-Up
[5] Agglomerative

دوره ۳، شماره ۳، پاییز ۱۳۹۶، صفحات ۳۱-۲۳

۳-۲- خوشه بندی داده ها براساس الگوریتم ژنتیک

در روش پیشنهادی در ابتدا به تعداد ۱ تا i عدد تصادفی تولید می گردد. در واقع این بدان است که برای هر مقدار تولید شده یک بار خوشه بندی با تعداد خوشه های مشخصی (که به صورت تصادفی تولید شد) انجام می شود. به عبارتی دیگر، با استفاده از الگوریتم K-mean (که یک روش خوشه بندی مشهور است) برای هر مقدار تولید شده یکبار خوشه بندی داده ها انجام می شود.

بنابراین جمعیتی بوجود می آید که هر فرد آن براساس الگوریتم K-mean خوشه بندی شده است. این جمعیت در واقع جمعیت اولیه الگوریتم ژنتیک را تشکیل می دهد.

در الگوریتم ژنتیک پیشنهادی، هر فرد از جمعیت شامل دو بخش است: بخش نمونه ها و بخش خوشه ها. به عنوان مثال در یک راه حل برای یک مسئله خوشه بندی با n نمونه و k خوشه هر فرد به صورت زیر نمایش داده می شود.

$$l_1, l_2, \dots, l_n \mid g_1, g_2, \dots, g_k$$

که ۱ نشان می دهد نمونه مورد نظر به کدام خوشه اختصاص داده شده است. g نیز شماره خوشه ها را مشخص می کند. با یک مثال این موضوع را توضیح می دهیم. فرض کنید ۱۵ نمونه وجود دارد و فردی به صورت زیر برای خوشه بندی وجود دارد:

1 3 2 1 4 1 1 2 3 2 1 3 4 2 1 | 1 2 3 4

در اینصورت چهار خوشه، داده ها را خوشه بندی می کنند که خوشه ها در این فرد شامل نمونه ها به صورت زیر هستند:

خوشه یک {۱,x۴,x۶,x۷,x۱۱,x۱۵x}

خوشه دو {۳,x۸,x۱۰,x۱۴x}

خوشه سه {۲,x۹,x۱۲x}

خوشه چهار {۵,x۱۳x}

در این مقاله ما ابتدا به صورت تصادفی جمعیتی از این افراد را تولید می کنیم، سپس هر یک از این افراد را با استفاده از الگوریتم K-mean حل می کنیم. پس از اعمال الگوریتم K- mean بر روی افراد، جمعیتی جدید بوجود می آید. جمعیت جدید بوجود آمده به عنوان جمعیت اولیه الگوریتم ژنتیک در نظرگرفته می شود.

۳-۳- نحوه ارزیابی افراد

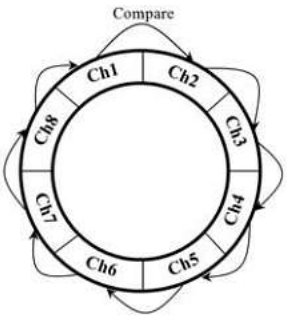
هر فرد بر اساس تابع ارزیابی برای انجام عمگرهای ژنتیکی انتخاب می شود. در واقع تابع ارزیابی مهمترین بخش از الگوریتم ژنتیک است و نحوه تعریف آن بستگی به نوع مسئله دارد (پریتی، ۲۰۱۱: دیوید، ۲۰۰۴: میشل، ۱۹۹۶). در این پروژه ما هر فرد را برساس میزان تراکم درون خوشه ای ارزیابی می کنیم. هر چقدر تراکم درون خوشه ای بیشتر باشد در اینصورت فرد دارای برازندگی بیشتری است و شانس بیشتری برای انتخاب شدن دارد، در غیراینصورت هرچقدر تراکم درون خوشه ای کم باشد در اینصورت برازندگی کمتر می شود و احتمال انتخاب شدن نیز کاهش می یابد.

۳-۴- نحوه انتخاب افراد برای نسل بعدی

در این مقاله ما بر اساس روش تورنومنت[1] افراد را برای نسل بعدی انتخاب می کنیم. این استراتژی به اینصورت پیاده سازی شده است که جمعیت افراد به صورت دوبه دو مقایسه می شوند و از بین دو فرد، فردی که دارای برازندگی بیشتری است انتخاب می شود. در واقع جمعیت افراد به صورت تصادفی در یک آرایه قرار می گیرند که ابتدا و انتهای این آرایه به هم متصل است (آرایه حلقوی). برازندگی هر فرد با برازندگی فرد بعدی مقایسه می شود و فردی انتخاب می شود که دارای برازندگی بیشتری باشد. شکل (۱) نحوه مقایسه دوبه دو افراد را نشان می دهد.

[1] Tournament

دوره ۳، شماره ۳، پاییز ۱۳۹۶، صفحات ۳۱-۲۳



شکل (۱): نحوه مقایسه برازندگی کروموزوم ها برای عملگر انتخاب مبتنی بر استراتژی تورنومنت

۳-۵- عملگر ترکیب

در این مقاله یک عملگر ترکیب جدید ارائه شده است که عملگر ترکیب بر اساس مراحل زیر طراحی شده است:

مرحله اول: ابتدا بر اساس احتمال ترکیب دوفرد از جمعیت انتخاب می شوند و دو نقطه تصادفی در بخش خوشه ها در فرد مشخص می شود (شکل ۲).

Ch1 = [1 3 2 1 4 1 1 2 3 2 1 3 4 2 1 | 1 2 3 4]

Ch2 = [3 1 2 1 3 2 2 1 3 1 2 3 2 2 2 | 1 2 3]

شکل (۲)

مرحله دوم: عناصری که از فرد اول متعلق به خوشه های انتخاب شده هستند در فرزند کپی می شوند (شکل ۳).

Offspring = [- 3 2 - - - - 2 3 2 - 3 - 2 - | 2 3]

شکل (۳)

مرحله سوم: از فرد دوم عناصری که متعلق به خوشه های انتخاب شده هستند در فرزند کپی می شوند (شکل ۴).

Offspring = [- 3 2 1' 1' 2' 2' 2 3 2 2' 3 - 2 2' | 2 3 1' 2']

شکل (۴)

مرحله چهارم: جاهای خالی باقیمانده به صورت تصادفی متناسب با تعداد خوشه ها پر می شوند (شکل ۵).

Offspring = [3 3 2 1' 1' 2' 2' 2 3 2 2' 3 2 2 2' | 2 3 1' 2']

شکل (۵)

مرحله پنجم: برچسب خوشه در فرزند به ترتیب از ۱ تا k مرتب می شود (شکل ۶).

Offspring = [2 2 1 3 3 4 4 1 2 1 4 2 1 1 4 | 1 2 3 4]

شکل (۶)

۳-۶- عملگر جهش

در این مقاله از عملگر جهش جدیدی استفاده شده است که این عملگر به دو صورت در نظرگرفته شده است. جهش براساس تقسیم: در این نوع جهش در بخش خوشه ها یک خوشه جدید اضافه می شود و هر یک از عناصری که در بخش نمونه ها قرار دارند، بر اساس احتمالی به خوشه جدید تعلق پیدا می کنند. فرض کنید در فرزند (offspring) که در مرحله ترکیب ایجاد شد این نوع جهش را اعمال کینم در این صورت فرزند پس از عمگر جهش به صورت شکل (۷) تغییر پیدا می کند.

[2 2 1 3 3 4 4 5 2 1 4 2 5 1 4 | 1 2 3 4 5]

شکل (۷)

جهش براساس ترکیب: در این نوع جهش در بخش خوشه ها یک خوشه به تصادف حذف می شود و تعداد خوشه یک واحد کاهش پیدا می کند. سپس عناصری که متعلق بـه خوشـه حـذف شـده هستند در بخش نمونه ها یک مقدار تصادفی متناسب با تعـداد و شماره خوشه ها به آن ها اختصاص داده می شود. فرض کنید در فرزندی که در مرحله ترکیب ایجاد شد این نوع جهـش را اعمـال کینم در این صورت فرزند پس از عمگر جهش به صـورت شـکل (۸) تغییر پیدا می کند.

[2 2 1 3 3 2 2 1 2 1 2 2 1 1 2 | 1 2 3]

شکل (۸)

۳-۷- انتخاب بازمانده ها

پس از انجام عملگرهای ژنتیکی جمعیت جدیدی تولید می شود، اگر جواب بهینه مورد قبول و مورد نظـر بدسـت بیایـد الگـوریتم خاتمه پیدا می کند در غیر اینصورت الگوریتم از ابتدای الگوریتم ژنتیک تکرار می شود.

به طور خلاصـه الگـوریتم پیشـنهادی، ابتـدا بـا تعـدادی جمعیت اولیه با تعداد خوشه های مشخص در بخـش خوشـه هـا تولید می شود، سپس هر فرد براساس الگوریتم K-mean خوشـه بندی می شود. در مرحله بعدی جمعیتی بوجود می آید کـه هـر فرد آن براساس الگوریتم K-mean خوشه بندی شده اسـت. ایـن جمعیت بوجود آمده به عنوان جمعیـت اولیـه الگـوریتم ژنتیـک پیشنهادی در نظر گرفته می شـود. سـپس عملگرهـای ژنتیکـی معرفی شده برروی ایـن جمعیـت اعمـال مـی شـود تـا براسـاس عمگلرهای ترکیب و جهش پیشنهادی تعداد خوشه هـای بهینـه پیدا شود. الگوریتم ژنتیک تا زمانی که بـه شـرط پایـانی برسـیم تکرار می شود.

۴- شبیه سازی و ارزیابی نتایج

در این مقاله ما روش پیشنهادی را در نرم افزار MATLAB پیاده سازی کرده ایم. شبیه سازی ها براساس تعداد نمونه های مختلف انجام شده است. در این مقاله تعـداد نمونـه هـا در چهـار حالـت ۱۵۰، ۳۰۰، ۴۵۰ و ۶۰۰ نمونه در نظر گرفتـه شـده اسـت. ایـن نمونه ها به صورت تصادفی در یک محیط ۱۰۰×۱۰۰ توزیع مـی شوند. مقادیر مربوط به پارامترهای الگوریتم ژنتیک به صورت زیر در نظر گرفته شده است:

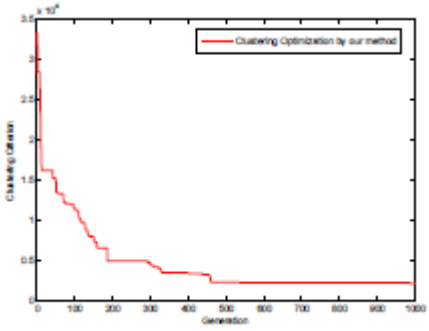
اندازه جمعیت = ۴۰ فرد

احتمال ترکیب = ۹۰

درصد احتمال جهش = ۵ درصد

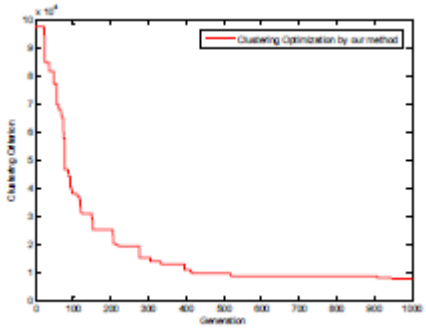
تعداد تکرار نسل ها = ۱۰۰۰ نسل

شکل (۹) بهینه سازی معیار خوشه بندی وقتی که تعـداد نمونـه ها برابر با ۱۵۰ است را نشان می دهد. در این حالت تعداد خوشه بهینه که برای خوشه بندی این نمونه ها پیدا شده است برابـر بـا ۲۵ خوشه است.

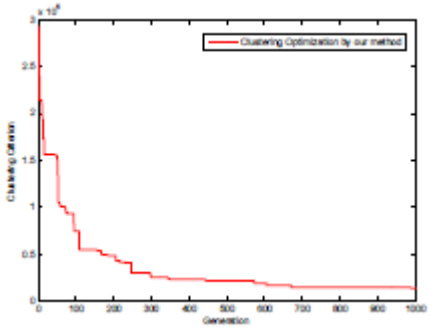


شکل (۹): بهینه سازی معیار خوشه بندی (تعداد نمونه = ۱۵۰)

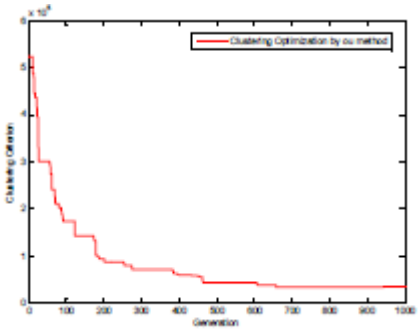
شکل (۱۰) بهینه سازی معیار خوشه بندی وقتی که تعداد نمونه ها برابر با ۳۰۰ است را نشان می دهد. در این حالت تعداد خوشه بهینه که برای خوشه بندی بهینه نمونه ها پیدا شده است برابر با ۲۶ خوشه است.
شکل (۱۱) بهینه سازی معیار خوشه بندی وقتی که تعداد نمونه ها برابر با ۴۵۰ نمونه است را نشان می دهد. در این حالت تعداد خوشه بهینه که برای خوشه بندی بهینه نمونه ها پیدا شده است برابر با ۲۸ خوشه است. شکل (۱۲) بهینه سازی معیار خوشه بندی وقتی که تعداد نمونه ها برابر با ۶۰۰ نمونه است را نشان می دهد. در این حالت تعداد خوشه بهینه که برای خوشه بندی بهینه نمونه ها پیدا شده است برابر با ۲۰ خوشه است.



شکل (۱۰): بهینه سازی معیار خوشه بندی (تعداد نمونه = ۳۰۰)



شکل (۱۱): بهینه سازی معیار خوشه بندی (تعداد نمونه = ۴۵۰)



شکل (۱۲): بهینه سازی معیار خوشه بندی (تعداد نمونه = ۶۰۰)

جدول (۱) تعداد خوشه های بهینه برای خوشه بندی بهینه در تعداد نمونه های مختلف را نشان می دهد. در واقع این جدول این معنی را می دهد که با پیدا کردن تعداد خوشه های بهینه معیار خوشه بندی نیز بهینه می شوند و در حالت کلی خوشه بندی داده ها بهینه می گردد. جدول (۱): تعداد خوشه های بهینه برای خوشه بندی بهینه در تعداد نمونه های مختلف

۵- نتیجه گیری

در این مقاله یک روش ترکیبی برای خوشه بندی براساس الگوریتم های ژنتیک و الگوریتم K-Mean ارائه کرده ایم به طوری که این الگوریتم پیشنهادی خود تعداد خوشه های بهینه را تشخیص داده و خوشه بندی را انجام دهد. همچنین روش های جدیدی را برای عمگرهای ترکیب و جهش ارائه کردیم.

نتایج شبیه سازی نشان می دهد که الگوریتم پیشنهادی با پیدا کردن تعداد خوشه های بهینه معیاری های خوشه بندی را بهینه می کند و روش مناسبی برای خوشه بندی داده ها در مقیاس های بزرگ به حساب می آید.

مراجع

1. David E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Published by Pearson Education, 2004, Page No.60-83.
2. Krishna, K., Murty, M., 1999. Genetic k-means Algorithm. IEEE Transactions on Systems, Man and Cybernetics B Cybernet 29, 433–439.
3. Laszlo, M., Mukherjee, S., 2007. A genetic algorithm that exchanges neighboring centers for k-means clustering. Pattern Recognition Letters 28 (16), 2359–2366.
4. M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering Validation Techniques", Journal of Intelligent Systems, vol. 17:2/3, pp 107-145, 2001.
5. Mitchell, Melanie, An Introduction to Genetic Algorithm, Published Bu MIT Press 1996.
6. Preeti, Vaishali, Genetic algorithm Approach for Optimal CPU Scheduling, IJCST Vol. 2, Issue 2, June 2011.
7. Rakesh, Rajiv, Sanjeev Ashwani, Genetic Algorithm approach to Operating system process scheduling problem, International Journal of Engineering Science and Technology Vol. 2(9), 2010, 4247-4252.
8. Sakai, T., Imiya, A., 2009. Unsupervised cluster discovery using statistics in scale space. Engineering Applications of Artificial Intelligence 22 (1), 92–100.
9. Zalik, K.R., 2008. An efficient k-means clustering algorithm. Pattern Recognition Letters 29, 1385–1391.

مطالعات علوم کاربردی در مهندسی

دوره ۳، شماره ۳، پاییز ۱۳۹۶، صفحات ۳۱-۲۳

Investigating and Explaining the Role and Position of New Genetic Algorithm Approaches for Data Clustering

Meysam Rahnamay Fallah[1], Marzieh Faridi Masouleh[2] *, Mohammad Reza Askaripour[3]

[1, 2, 3] *Islamic Azad University, Electronic Unit, Computer Department, Tehran, Iran*
** Corresponding Author*

Abstract

Clustering or analyzing clusters in machine statistics and learning is one of the branches of unsupervised learning and is a process in which samples are divided into categories with similar members known as clusters. Therefore, a cluster is a set of similar objects different from objects in other clusters. Clustering is the only method in unsupervised earning. A cluster is referred to as a set of data with similarity at least in one attribute. Attempts are made in clustering to divide the data into clusters so that there exists the maximum similarity among the data within each cluster and the minimum similarity among the inter-cluster data. In this paper we present a hybrid method for clustering based on genetic algorithms, so that the proposed algorithm itself detects the number of optimum clusters and performs the clustering. Simulation results show that the proposed algorithm improves data clustering by finding the number of optimum clusters.

Keywords: new approaches, genetic algorithm, clustering, data.