بهینه سازی در الگوریتم K-MEANS مبتنی بر کرنل

حامد ملایی سربیژن

سرشناسه : ملایی سربیژن، حامد، ۱۳۶۴-
عنوان و نام پدیدآور : بهینه‌سازی در الگوریتم K-MEANS مبتنی بر کرنل/ حامد ملایی‌سربیژن.
مشخصات نشر : زرند: نوید دانش دارالفنون، ۱۴۰۰.
مشخصات ظاهری : ۹۵ ص.: جدول، نمودار.؛ ۱۴×۲۱سم.
شابک : ۴۵۰۰۰۰ ریال-7-9-97897-622-978 :
وضعیت فهرست نویسی : فیپا
یادداشت : کتابنامه: ص. ۹۱ - ۹۵.
موضوع : توابع کرنل
موضوع : Kernel functions
موضوع : الگوریتم‌های کامپیوتری
موضوع : Computer algorithms
رده بندی کنگره : QA۳۵۳
رده بندی دیویی : ۵۱۵/۷۳۲
شماره کتابشناسی ملی : ۷۶۲۰۰۲۷
وضعیت رکورد : فیپا

مولف: حامد ملایی سربیژن

عنوان: بهینه سازی در الگوریتم K-MEANS مبتنی بر کرنل/حامد ملایی‌سربیژن

طراح جلد: فهیمه پورحسنخانی

نوبت چاپ: اول بهار ۱۴۰۰

شابک: ۹۷۸-۶۲۲-۹۷۸۹۷-۹-۷

شمار: ۱۰۰۰ نسخه

قیمت: ۴۵۰۰۰ تومان

زرند، خیابان فردوس، نبش فردوس ۲۰، طبقه فوقانی دارالترجمه الفبا

تلفن تماس: ۰۹۲۲۱۱۸۳۴۹۵ ۰۹۱۳۵۱۸۸۷۱۶ ۳۳۴۳۹۹۵۶

فهرست مطالب

مقدمه

K – means یک روش پایه، برای بسیاری از روش های خوشه بندی است .با تحقیقات انجام شده توسط کواترو همکارانش در سال ۲۰۱۲ این روش می تواند در خوشه بندی داده ها در داده کاوی مورداستفاده قرار گیرد و عملکرد مطلوب داشته باشد .الگوریتم K – means را به منظور خوشه بندی به کار خواهیم گرفت .داده های نزدیک یک گروه را باهم می سازند که این گروه یک خوشه در شبکه در نظر گرفته خواهد شد .جاذبه مرکزی خوشه به عنوان مرکز خوشه در نظر گرفته می شود .در موقعیت اولیه اجرا، هر دو داده نزدیک به هم یک گروه یا خوشه را در زمان اولیه می سازند .گره های دیگر به صورت تدریجی به خوشه ها ملحق می شوند و این عمل تا زمانی که تعداد خوشه ها به یک حد مشخصی برسد ادامه پیدا می کند .همه داده ها در ابتدای اجرا دارای انرژی یکسانی هستند . داده ای که نزدیکترین فاصله را با داده های دیگر در خوشه دارد به عنوان مرکز خوشه در آن خوشه انتخاب می شود[۱].

خوشه بندی در واقع تعریف یک مجموعه داده بر اساس تشابهات بین اعضای آن در راستای رسیدن به تشخیص یک مجموعه از دسته بندی ها می باشد .هدف اصلی خوشه بندی،

بزرگنمایی همجنس سازی در هر خوشه و عدم تجانس بین خوشه های مختلف می باشد که وابسته به نوع الگوریتم بکار رفته است .پس درواقع، موجودیت هایی که به یک دسته تعلق دارند دارای اشتراک های بیشتر با یکدیگر می باشند تا موجودیت هایی که متعلق به خوشه های مختلف هستند.

با مسئله سنجش شباهت معمولا به صورت غیرمستقیم برخورد می شود، یعنی مقیاس های مسافت برای تعیین میزان تفاوت بین اشیاء مورداستفاده قرار می گیرد، در چنین روشی اشیاء بسیار مشابه دارای کمترین مقدار تفاوت می باشند .برای اجرای تفکیک، از معیارهای متفاوت بسیاری می توان استفاده کرد .هر معیار ویژگی های خود را داشته و با منفعتها و ایرادهای خود همراه است.

از میان روشهای مختلف شناسایی الگو میتوان به روش های مبتنی بر کرنل به عنوان یکی از حوزه های بسیار فعال در سالهای اخیر اشاره کرد .در روشهای مبتنی بر کرنل می توان به الگوریتم های طبقه بند از قبیل ماشین بردار پشتیبان [1](SVM)، الگوریتم های رگرسیون

[1] Support Vector Machine

مانند رگرسیون بردار پشتیبان (SVR)[2] و الگوریتم های کاهش بعد تعمیم یافته از قبیل آنالیز مؤلفه های ویژه مبتنی بر کرنل (KPCA)[3] و آنالیز تفکیک کننده خطی تعمیم یافته[4] اشاره کرد .تعداد زیادی از این روشهای مبتنی بر کرنل به طور گسترده در کاربردهای مختلف از قبیل فناوری اطلاعات زیستی[5]، بینائی ماشین[6]، بازشناسی گفتار[7]، داده کاوی[8]، بازیابی اطلاعات[9] و شناسایی الگو مورد استفاده قرار می گیرند و این حاکی از این است که روش های مبتنی بر کرنل ابزارهای قوی برای کاربردهای مختلف می باشند .در الگوریتم های مبتنی بر کرنل، داده ها به فضای ویژگی جدیدی با ابعاد بالاتر نگاشت می شوند و سپس در فضای ویژگی جدید فرآیند مورد نظر (به عنوان مثال جستجوی مرزهای خطی بین طبقات دادگان) اجرا می شود .در این الگوریتمها از تابعی بنام تابع کرنل استفاده میشود که میزان شباهت بین دو نمونه در فضای ویژگی جدید را محاسبه می کند .اطلاعات شباهت بین تمام زوج نمونه ها در یک ماتریس به نام ماتریس کرنل جمع آوری می شود که ماتریسی متقارن

[2] Support Vector Regression
[3] Kernel Principal Component Analysis
[4] Generalized Discriminant Analysis
[5] Bioinformatics
[6] Computer Vision
[7] Speech Recognition
[8] Data Mining
[9] Information Retrieval

و نیمه معین مثبت است و همگرایی الگوریتم های آموزشی به دلیل همین ویژگی ماتریس کرنل تضمین می شود[۳]

در روشهای مبتنی بر کرنل ذکر شده، تنها از یک تابع کرنل و یا به طور معادل از یک ماتریس کرنل استفاده میشود و بنابراین کیفیت عملکرد چنین الگوریتم هایی به نوع کرنل و یا پارامترهای کرنل استفاده شده وابسته است .یکی از راهکارهای ارائه شده برای تعیین کرنل مناسب و یا پارامترهای آن، استفاده از روشهای مبتنی بر اعتبار سنجی متقابل[10] است که بسیار زمان بر بوده و از طرفی در این روش ممکن است تعیین مناسب ترین تابع کرنل برای داده های موجود امکان پذیر نباشد ..از طرفی ممکن است برخی از فضاهای ویژگی پایه دارای اهمیت بیشتری در مقایسه با دیگر فضاهای ویژگی باشند و به همین دلیل می توان برای فضاهای ویژگی و در نتیجه کرنل های مورد بررسی در ترکیب کرنلی، وزن هایی را اختصاص داد .در روش های یادگیری کرنل مرکب(MKL)[11]، هدف محاسبه میزان اهمیت هر یک از این فضاهای ویژگی و یا وزن کرنل ها است که در اکثر موارد از یک الگوریتم

[10] Cross Validation
[11] Multiple Kernel Learning

تکراری برای محاسبه وزن کرنل ها و پارامترهای سیستم یادگیرنده استفاده می شود.کرنل ها را می توان به طرق مختلف خطی و غیر خطی ترکیب نمود که ترکیب خطی به دلیل سادگی، بیشتر مورد استقبال محققین بوده و روش های زیادی بر مبنای این ترکیب وزن دار ارائه شده است .با نگاهی عمیق تر می توان برخی از روش های ارائه شده برای MKL را به عنوان روش هایی برای یادگیری ماتریس کرنل تفسیر نمود .به عبارتی می توان روشهای MKLرا این چنین تفسیر نمود که هدف، یادگیری ماتریسی است که به طور غیر صریح داده ها را به فضای جدیدی نگاشت دهد به گونه ای که داده ها در فضای جدید توزیعی متناسب با کاربرد مورد نظر داشته باشند. به عنوان مثال ممکن است هدف طبقه بندی دادگان در فضای جدید با استفاده از روش های خطی باشد. در این دیدگاه، روش های MKL شباهت بسیار زیادی به روش های یادگیری معیار فاصله دارند .از طرفی، مقادیر کرنل و فاصله، هر دو معیاری برای سنجش شباهت میان نمونه ها بوده و از این رو شباهت زیادی در اهداف مورد بررسی در دو روش یادگیری کرنل و یادگیری معیار فاصله وجود دارد . همانطور که نیاز به یافتن کرنل بهینه در روشهای مبتنی بر کرنل منجر به پیدایش روشهای

MKL گردید، نیاز به تعیین تابع فاصله مناسب برای نمایش بهتر دادگان نیز سبب پیدایش و گسترش روش های یادگیری معیار فاصله شده است [۴].

بر این اساس که تفکیک داده ها اصولا ذات غیر نظارتی دارد، به عنوان یکی از سخت ترین و پر چالش ترین پرسش های مجامع علمی در اتوماسیون شمرده می شود .طبیعت غیر نظارتی دلالت بر این امر دارد که ویژگی های ساختاری مسئلهk-means به وضوح برای ما شناخته شده نیستند، مگر اینکه نمونه هایی از دانش پیرامون دامنه به شکل خاص توزیع فضای داده به صورت عدد، حجم، جرم، چگالی، تصویر و جهت خوشه ها در دسترس باشد . حجم زیاد اطلاعات و داده ها به دلیل استفاده از تکنولوژی روز و استفاده از کامپیوتر در تجارت، دانش و خدمات و پیشرفت تکنیک ها و ابزار مورداستفاده در جمع آوری و پردازش دادهها، گسترش روزافزون وب و اینترنت به عنوان یک رسانه جهانی نیاز آشکاری به تکنولوژی ها و ابزارهای جدید و خودکار برای تبدیل هوشمندانه حجم زیاد داده به اطلاعات را ایجاد می کنند .بر این اساس خوشه بندیk-means باعث از بین رفتن جزئیات داده های ورودی خواهد شد اما در نهایت این کار ساده سازی امر استفاده از داده در تحلیل را در پی دارد .با این روش، پایگاه های داده با حجم زیاد بر اساس تعداد کمتری از خوشه

ها، ارائه می شوند، پس در واقع می توان گفت این تکنیک، داده ها را با خوشه هایشان مدل می کند .این رشد در مقدار داده یک مشکل اساسی در به کار گیری علوم و روشهای تجزیه و تحلیل ایجاد کرده است.

در میان این تکنیک ها، تجزیه و تحلیل خوشه بندی یکی از پراستفاده ترین ها است، و به این صورت است که با استفاده از الگوریتم محبوب K – Means انجام می شود .بسیاری از پژوهش های دانش بنیان نشان میدهد که الگوریتم های خوشه بندی در تحقیقات امروز نقش مؤثری دارند.

الگوریتم کرنل

روشهای مبتنی بر کرنل، الگوریتم های قدرتمندی هستند که از حوزه تئوری علم آمار به مسائل کاربردی یادگیری ماشین وارد شده اند .ایده اصلی در روشهای مبتنی بر کرنل، نگاشت مجموعه داده های ورودی به فضایی برداری است که معمولا تعداد ابعاد آن خیلی

زیاد است (فضای ویژگی یا فضای هیلبرت[12]) که پس از آن از الگوریتم های خطی آنالیز الگو برای کشف روابط بین داده های نگاشت یافته استفاده میشود [۴].

یکی از مهم ترین خصوصیات روشهای کرنلی این است که با به کارگیری ترفند کرنل [13] نیازی به محاسبه صریح نگاشت داده ها به فضای ویژگی نیست .به عبارتی اعمال ترفند کرنل منجر به تعمیم الگوریتم های خطی به نسخه ی غیر خطی آنها می شود .از این رو، نوع کرنلی بسیاری از الگوریتم ها ارائه شده است .اما انتخاب کرنل مناسب و پارامترهای آن (به عنوان مثال درجه چندجمله ای[14] در کرنل چند جمله ای ، پارامتر واریانس در کرنل گوسی[15] و...) یکی از مهم ترین مسائل در روشهای کرنلی است چرا که تابع کرنل باید متناسب با داده های ورودی انتخاب شود و نوع کرنل انتخاب شده تأثیر زیادی در کیفیت عملکرد الگوریتم یادگیری مورد بررسی خواهد داشت .بنابراین مهم ترین مسئله در ارتباط با روشهای مبتنی بر کرنل، انتخاب کرنل مناسب است .در سال های اخیر استفاده از ترکیب مناسب تعدادی از توابع کرنل به جای استفاده از یک تابع کرنل پیشنهاد شده است.

12 Hilbert
13 Kernel Trick
14 Polynomial Kernel
15 Gaussian Kernel