بسم الله الرحمن الرحیم

سال ۱۳۰۰

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان نامه ی کارشناسی ارشد رشته ی مهندسی کامپیوتر گرایش نرم افزار

بهبود روش های دسته بندی در داده کاوی با استفاده از دانش گذشته

استاد راهنما:

دکتر احمد براآنی

پژوهشگر:

مهدی مرادیان

خرداد ماه ۱۳۸۸

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان نامه ی کارشناسی ارشد رشته ی مهندسی کامپیوتر گرایش نرم افزار

آقای مهدی مرادیان تحت عنوان

بهبود روش های دسته بندی در داده کاوی با استفاده از دانش گذشته

در تاریخ ۱۳۸۸/۳/۲۵ توسط هیأت داوران زیر بررسی و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان نامه	دکتر احمد براآنی	با مرتبه ی علمی استادیار	امضا
۲- استاد داور داخل گروه	دکتر بهروز ترک لادانی	با مرتبه ی علمی استادیار	امضا
۳- استاد داور خارج از گروه	دکتر مازیار پالهنگ	با مرتبه ی علمی استادیار	امضا

امضای مدیر گروه

با تشکر از

استاد ارجمندم جناب آقای دکتر براآنی که همواره از مساعدت های ایشان در به انجام رساندن این تحقیق بهره مند بوده ام.

و

فرهیختگان محترم هیئت داوران و اساتید گرانقدر که داشته هایم از دانش، مرهون زحمات و راهنمایی های ایشان است.

تقدیم به پیشگاه مقدس حضرت ولی عصر(عج)

تقدیم به پدر ارجمندم
که بزرگوارانه در حفظ و تربیت من کوشید و عزت و همت والای خود را سرلوحه دفتر زندگیم قرار داد و اقتدار من حاصل زحمت بیکران اوست.

تقدیم به مادر عزیزم
که خورشید مهربانیش هیچگاه غروب نمی کند و باران رحمتش همیشگی است.او که دریای عشق و ایثار است و از دعای اوست که امروز به اینجا رسیده ام.

و

همسر

عزیزم

چکیده

الگوریتم KNN یکی از بهترین و پرکاربردترین الگوریتم های دسته بندی است که از آن استفاده ی گسترده ای در کاربردهای مختلف می شود. یکی از مشکلات این الگوریتم، تأثیر یکسان همه ی خصیصه ها در محاسبه ی فاصله ی رکورد جدید با رکوردهای موجود در پایگاه داده های آموزشی می باشد، در صورتی که برخی از این خصیصه ها برای عمل دسته بندی کم اهمیت ترند. این امر باعث گمراهی روند دسته بندی و کاهش دقت الگوریتم دسته بندی می شود.

در این تحقیق به استخراج نوع خاصی از قوانین وابستگی می پردازیم که سمت چپ آن ها فقط یک قلم وجود دارد و سمت راست آن ها برچسب دسته وجود دارد. سپس با بررسی و آنالیز این قوانین وابستگی و ترکیب آن ها با الگوریتم KNN دو الگوریتم دسته بندی جدید پیشنهاد می کنیم. در الگوریتم اول یعنی الگوریتم نزدیک ترین k تایی همسایگی مبتنی بر وابستگی پویا[1] بر اساس مقادیر ویژگی های رکورد جدید به صورت پویا به ویژگی های مختلف وزن اختصاص می دهیم، یعنی هر بار که یک رکورد جدید وارد می شود، وزن ویژگی ها برای محاسبه ی فاصله تغییر می کند. اما در الگوریتم دوم یعنی الگوریتم نزدیک ترین k تایی همسایگی مبتنی بر وابستگی ایستا[2] بدون توجه به مقادیر ویژگی های رکورد جدید و به صورت ایستا به ویژگی های مختلف وزن اختصاص می دهیم. در الگوریتم دوم وزن همه ی ویژگی ها برای همه ی رکوردهای جدیدی که وارد سیستم می شوند، ثابت است. بعد با توجه به وزن خصیصه های مختلف و با استفاده از فرمول محاسبه ی فاصله ی مانهتن بین رکوردها به دسته بندی بر اساس الگوریتم KNN می پردازیم و با این عمل دقت دسته بندی الگوریتم KNN را افزایش می دهیم. مقایسه ی نتایج ارزیابی این الگوریتم با هفت الگوریتم دیگر دسته بندی بر روی پایگاه داده ی مختلف، بهبود قابل توجه دقت دسته بندی توسط این الگوریتم را نشان می دهد.

کلیدواژه: داده کاوی، دسته بندی، الگوریتم نزدیک ترین k تایی همسایگی، قوانین وابستگی، وزن دهی به خصیصه ها

---

[1] D_KNNBA (Dynamic-K-Nearest-Neighbor)
[2] S_KNNBA (Static-K-Nearest-Neighbor)

فهرست مطالب

عنوان صفحه

عنوان

صفحه

عنوان | صفحه

فهرست شکل ها

عنوان صفحه

فهرست جدول ها

عنوان صفحه

فصل اول

کلیات مسئله

۱-۱ شرح مسئله

استفاده از داده کاوی به دلیل حجم زیاد داده ها و همچنین میزان کم دانش و اطلاعاتی که بشر با توجه به این حجم داده در اختیار دارد ، ضروری به نظر می رسد [۱] - [۲].داده کاوی فرایندی است که با استفاده از تکنیکهای هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند.در داده کاوی از سه روش اصلی استخراج قوانین وابستگی[1] ، خوشه بندی[2] و دسته بندی[3] استفاده می شود[۱].

دسته بندی نوعی از آنالیز داده ها می باشد که رکوردها را به وسیله ی الگوهای از پیش استخراج شده [4] ، بین دسته های از پیش تعریف شده تقسیم می کند و هر رکورد را با توجه به خصیصه هایش[5] در یکی از این دسته ها قرار می دهد.دسته ها بر اساس بررسی داده های موجود و با توجه به شرایط و نیاز محیط (تعریف مسئله) تعریف می شوند وبه هر دسته معنای خاص و برچسب دسته ی خاص اختصاص داده می شود . برای مثال مشتری های یک بانک را می توان به دو دسته ی مشتری های خوش حساب و مشتری های بد حساب تقسیم کرد که برچسب دسته ی اول، مشتری خوش حساب و معنای آن برای استفاده کننده از سیستم در بانک

---

[1] Association rules
[2] Clustering
[3] Classification

[4] در ساده ترین حالت این الگوها تعدادی قوانین اگر...آنگاه..(.....if....then)هستند.

[5] Attribute

مشخص است[۱]. دسته بندی کاربردهای متعددی دارد ، به عنوان مثال از دسته بندی برای تشخیص بیماری ها[۶][۳] ، بازاریابی انتخابی[۷] [۴]، تصویب اعتبار[۸][۱] و غیره استفاده می شود. الگوها توسط الگوریتم های مختلف دسته بندی از جمله الگوریتم های شبکه عصبی ، درخت تصمیم ، ژنتیک و غیره [۱] - [۵] - [۶] - [۷] - [۸] - [۹] - [۱۰] - [۱۱] استخراج می شوند. از الگوهای به دست آمده برای فهم و تشریح داده های موجود و برای پیش بینی نحوه ی رفتار موارد جدید استفاده می شود.

البته دسته ی دیگری از الگوریتم های دسته بندی نیز وجود دارند، که به جای استخراج الگو به منظور دسته بندی تا زمان ورود یک داده ی جدید به سیستم کاری انجام نمی دهند، هر بار پس از ورود داده ی جدید با مقایسه ی آن داده و داده های موجود در سیستم به دسته بندی داده ی جدید می پردازند.

الگوریتم نزدیک ترینk تایی همسایگی[۹] یکی از این الگوریتم ها است که پس از ورود رکورد جدید به سیستم، k تا از مشابه ترین رکوردهای موجود را به عنوان نزدیک ترین همسایه های این رکورد تشخیص داده و بر اساس برچسب دسته ی این همسایه ها به رکورد جدید یک برچسب دسته اختصاص می دهد. در الگوریتم نزدیک ترین K تایی همسایگی، هر بار پس از ورود یک رکورد جدید به سیستم، آن را در یک فضای n بعدی (n تعداد خصیصه های رکورد است.) با رکوردهای موجود مقایسه می کند و براساس معیار فاصله ی اقلیدسی k تا از نزدیک ترین رکوردها به این رکورد جدید را به عنوان همسایه هایش انتخاب می کند. سپس در ساده ترین حالت بر چسب دسته ای که اکثریت همسایه ها متعلق به آن دسته باشند را استخراج می کند. استفاده از این الگوریتم در پایگاه داده های بزرگ به دلیل حجم زیاد عملیات، کارایی زیادی ندارد ولی در پایگاه داده های کوچک یکی از بهترین الگوریتم های دسته بندی است.

یکی از مشکلات این الگوریتم این است که وقتی رکورد جدید را با رکوردهای موجود در فضای n بعدی مقایسه می کنیم به همه ی ویژگی ها به یک اندازه توجه می کنیم، در صورتی که ویژگی های مختلف برای دسته بندی رکوردها اهمیت یکسان ندارند. برخی از ویژگی ها هیچ نقشی در دسته بندی ندارند، برخی اهمیت

---

[6] با توجه به علائم یک بیماری ، الگویی استخراج می کند و با آن الگو افراد را به دو دسته ، افراد دارای بیماری و افراد فاقد بیماری تقسیم می کند.

[7] با توجه به ویژگی های مشتری ها ، الگویی استخراج می کند و با آن الگو افراد را به دسته هایی با سلایق خرید مختلف تقسیم می کند و در هنگام ارسال آگهی های تبلیغاتی برای هر دسته مشتری آگهی های مناسب را ارسال می کند.

[8] مشتری های یک بانک را با توجه به ویژگی هایشان از نظر گردش مالی ، خوش حسابی و غیره به چند تقسیم می کند و به هر دسته تا سقف مشخصی اعتبار اختصاص می دهد.

[9] k-Nearest-Neighbors

کمتر و برخی دیگر اهمیت زیادی در دسته بندی آن رکورد جدید دارند. به همین دلیل برخی از تحقیق هایی که سعی در بهبود الگوریتم نزدیک ترین k تایی همسایگی داشته اند، به جای ارزش دهی یکسان به همه ی ویژگی ها، به ویژگی ها، وزن های مختلف نسبت می دهند و به جای استفاده از فاصله ی اقلیدسی از فاصله ی مانهتن استفاده می کنند. این کار باعث کاهش تأثیر منفی ویژگی های غیر مرتبط با روند دسته بندی در محاسبه ی فاصله ی بین رکوردها می شود و درعمل دقت دسته بندی را افزایش می دهد.

یکی دیگر از روش های داده کاوی، استخراج قوانین وابستگی است. در این روش، الگوهای بین داده ها به صورت قوانینی بیان می شود. به عنوان مثال در پایگاه داده های مشتری های بانک ممکن است قانونی به شکل زیر قابل استخراج باشد:

Credit = good AND Age= middle → Credit – rating = good

یک قانون وابستگی یک گزاره ی منطقی به شکل A→ B است که در آن A و B دو مجموعه قلم هستند. که اشتراکشان تهی است. هر قلم ارتباطی بین یک ویژگی ومقدارش را بیان می کند.

A پیش فرض و B نتیجه ی قانون است. در این قانون بیان می شود که مشتری های میان سال که حقوق آنها خوب است، مشتری های خوبی برای بانک هستند. کارایی هر قانون وابستگی توسط دو پارامتر پشتیبانی وضریب اطمینان سنجیده می شود. پشتیبانی قانون A→ B، برابر درصدی از رکوردها است که A و B با هم در آن ها وجود داشته باشند و ضریب اطمینان این قانون، برابر درصدی از رکوردهای شامل A است که شامل B نیز هستند.

در این تحقیق می خواهیم با استفاده از قوانین وابستگی برای هر ویژگی درالگوریتم KNN وزن به دست آوریم و دقت الگوریتم KNN را افزایش دهیم.

۱-۲- الگوریتم ارائه شده

در این تحقیق کلیه ی قوانین وابستگی به صورت A→ B که در آنها A تنها شامل یک قلم و B بیانگر یک برچسب دسته است، استخراج می شوند. سپس ضریب اطمینان هر قانون و پشتیبانی قلم A را در آن قانون استخراج می کنیم. حال کلیه ی قوانینی که قلم A در آنها مربوط به یک ویژگی یکسان می شود را در یک گروه قرار می دهیم. واضح است که با این کار سمت چپ کلیه ی قوانین هر گروه مثل G_i فقط اقلام مرتبط با یک ویژگی مثل A_i وجود دارد و بنابراین هر گروه نماینده ی یکی از ویژگی ها است.

سپس برای هر گروه دو پارامتر ضریب اطمینان گروه برابر حداکثر مقدار ضرایب اطمینان قوانین مختلف موجود در یک گروه و پشتیبانی برابر حداکثر مقدار پشتیبانی قلم های مختلف موجود در سمت چپ قوانین آن گروه را تعریف می کنیم.

ایده ی اصلی برای تشخیص اهمیت هر ویژگی در دسته بندیّ این است که چقدر مقادیر آن ویژگی تکرار شده اند و از طرف دیگر تکرار مقادیر آن ویژگی چقدر با تعیین بر چسب کلاس برای رکوردهای آن پایگاه داده ها مرتبط است. بر همین اساس مقادیر دو پارامتر ضریب اطمینان گروه و پشتیبانی گروه را با مقادیر حد آستانه ی تعریف شده در هر پایگاه داده مقایسه می کنیم، ویژگی هایی که مقادیر حداقل یکی از این دو پارامتردر آن ها از حد آستانه ی تعریف شده پایین تر باشد، وزن ϕ خواهند داشت. ($w_i = \phi$) و بقیه ی ویژگی ها براساس پشتیبانی گروه آنها مانند فرمول ۱-۱ وزن می گیرند.

فرمول ۱-۱

$$w_i = \frac{1}{1-(G-SUP_i)}$$

حال با توجه به این وزن ها و با استفاده از فرمول محاسبه ی فاصله ی ۱-۲ به جای فرمول فاصله ی اقلیدسی، الگوریتم KNN را اجرا می کنیم. این وزن های جدید، باعث بهبود الگوریتم KNN می شوند.

فرمول ۱-۲

$$x_1 = (x_{11}, x_{12}, \ldots, x_{1n})$$
$$x_2 = (x_{21}, x_{22}, \ldots, x_{2n})$$
$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} w_i * (x_{1i} - x_{2i})^2}$$

۱-۳- نتایج به دست آمده:

الگوریتم ارائه شده از زوایای مختلفی با الگوریتم KNN بر روی پایگاه داده های موجود در UCI [۱۲] مورد ارزیابی قرار گرفت. در فصل ۴ ابتدا به بررسی تاثیر اعداد K ی مختلف بر روی دقت الگوریتم ارائه شده و الگوریتم KNN می پردازیم و مشاهده می کنیم که برای مقادیر مختلف K الگوریتم ارائه شده بهبود دقت را نسبت به الگوریتم KNN نشان می دهد. از طرف دیگر برای مقادیر مختلف n ، روش ارزیابی n-Cross-fold–validation را در دو الگوریتم KNN و الگوریتم ارائه شده اجرا کرده و مشاهده می کنیم که مقادیر مختلف n نیز نمی تواند تاثیر چندانی در کیفیت عملکرد الگوریتم ارائه شده نسبت به الگوریتم KNN داشته باشد و در پایان الگوریتم ارائه شده را با ۷ الگوریتم دیگر دسته بندی در نرم افزار Weka [۱۳] بر روی ۱۵

پایگاه داده از UCI اجرا کرده و مشاهده می کنیم که این الگوریتم، متوسط دقت بهتری را نسبت به دیگر الگوریتم ها ارائه می کند.

در فصل ۲ به بررسی اجمالی مفهوم داده کاوی، روش های مختلف آن وغیره می پردازیم. در فصل ۳ به بیان روش های مختلف دسته بندی و مسائل پیرامون آن می پردازیم. در فصل ۴ ابتدا الگوریتم KNN را با ذکر یک مثال بیان کرده و الگوریتم های مختلفی که قصد در بهبود این الگوریتم داشته اند را بیان می کنیم و در ادامه به بیان مفهوم قوانین وابستگی و نحوه ی استخراج آن ها از داده ها می پردازیم، سپس دو الگوریتم جدید براساس استفاده از نوع خاصی از قوانین وابستگی در الگوریتم KNN پیشنهاد می کنیم. در فصل ۵ به ارزیابی الگوریتم های ارائه شده در این تحقیق از جنبه های مختلف و مقایسه ی آن ها با دیگر الگوریتم های دسته بندی می پردازیم.

فصل دوم

داده کاوی

۲-۱مقدمه

از سال ۱۹۶۰ که سیستمهای پایگاه داده ها ، جای سیستمهای پردازش فایل را گرفتند[۱] ، با سرعت این سیستمها رو به پیشرفت بودند و از طرف دیگر کاربرد آنها درعرصه های مختلف صنعت و اقتصاد و سیاست و غیره با سرعت شگفت انگیزی رو به افزایش بود. این رشد روز افزون داده ها و از طرف دیگر ، عدم توانایی بشر برای درک و بررسی همهٔ داده ها باعث شد که بیشتر داده ها به آرشیو سپرده شوند و درتصمیم گیریها فقط از داده های کمی (که به نظر مدیر مهم تر بودند) استفاده شود، این درحالی بود که شاید داده های موجود درآرشیو دارای اطلاعات مفید زیادی بودند.

چون نیاز ما دراختراع است .محققان شروع به کار برای به دست آوردن فرایندی کردند که بتواند از میان داده های زیاد، دانش ارزشمند را به دست آورد. آنها به دنبال یک مکانیزم خودکار بودند که بدون دخالت بشر بتواند در میان حجم عظیم داده ها الگوهای مفید، جدید و قابل فهم را استخراج کند. اولین بار دکتر شاپیرو[1] در سال ۱۹۸۹ یک کارگاه کوچک[2] برای داده کاوی[3] ایجاد کرد ولی از آن سال به بعد به مرور زمان آن کارگاه کوچک تبدیل به کنفرانسهای مهم و بزرگ کشف دانش در پایگاه داده ها[4] شد. داده کاوی تا آنجا گسترش

---

[1] Shapiro
[2] KDD-89: IJCAI-89 workshop on Knowledge Discovery in Databases August 20, 1989, Detroit MI, USA
[3] Datamining
[4] KDD

یافت که اگر امروز در موتور جستجوی گوگل، داده کاوی را جستجو کنیم بیش از ۱۰۰۰۰۰۰۰ صفحه را پیدا می کند که نشان دهنده گسترش سریع و زیاد داده کاوی در صنعت، تجارت و غیره است.

۲-۲داده کاوی چیست؟

داده کاوی استخراج دانش ارزشمند از بین حجم زیادی داده است[۱]. اما به نظر می رسد که این نام چندان مناسب نیست. وقتی طلا را از سنگ آن استخراج می کنیم آنگاه می گوییم به کاویدن طلا پرداختیم نه کاویدن سنگ، بنابراین باید از دانش کاوی[۱] به جای داده کاوی استفاده کنیم. محققان مختلف برای فرایند داده کاوی نام های مختلفی را استفاده کرده اند، از جمله عصاره گیری دانش[۲] ، آنالیز داده / الگو[۳] ، باستان شناسی داده[۴] و غیره. داده کاوی را می توان به درختی تشبیه کرد که درعلوم مختلف ریشه دارد. از جمله این علوم می توان به تکنولوژی پایگاه داده ها، یادگیری ماشین ، آمار، شخصی سازی[۵] و بصری سازی[۶] اشاره کرد(شکل۲-۱). در ادامه جای پای این علوم را در فرایند داده کاوی مشاهده خواهیم کرد.

برخی از محققان کل فرایند کشف دانش از داده ها[۷] را داده کاوی می گویند ولی برخی دیگر داده کاوی را فقط یکی از مراحل فرایند کشف دانش می دانند.

برای طراحی یک دوره حیات[۸] برای فرایند داده کاوی شرکتهای مختلف پیشنهادهای متفاوتی داده اند. معروف ترین آنها CRISP-DM [۲] از شرکت SPSS و SEMMA از شرکت SAS است ولی در این قسمت به بررسی فرایندی می پردازیم که در بیشتر کتاب های آکادمیک از این فرایند استفاده می شود.

---

[1] Knowledge mining
[2] Knowledge extraction
[3] data/pattern Analysis
[4] Data Archaeology
[5] privacy
[6] visualization
[7] Knowledge Discovery from Data(KDD)
[8] life cycle