بسم الله الرحمن الرحيم

دانشگاه پیام نور واحد شهرضا

پایان نامه ی مقطع کارشناسی رشته مهندسی فناوری اطلاعات

داده کاوی وکاربرد آن درتشخیص بیماری ها (دیابت)

ارائه دهنده:

زینب

تابستان 92

سپاسگزاری

شکر خدا که هر چه طلب کردم از خدا بر منتهای همت خود کامران شدم

وبا سپاس فراوان از راهنمائی ها وزحمات استاد محترم وصبورم

چکیده

امروزه در دانش پزشکی جمع آوری داده های فراوان در مورد بیماری های مختلف از اهمیت فراوانی برخوردار است. مراکز پزشکی با مقاصد گوناگونی به جمع آوری این داده ها می پردازند . تحقیق روی این داده ها و به دست آوردن نتایج و الگوهای مفید در رابطه با بیماری ها ،یکی از اهداف استفاده از این داده ها است . حجم زیاد این داده ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می شود.

بنابراین از داده کاوی برای غلبه بر این مشکل و به دست آوردن روابط مفید بین عوامل خطر زا در بیماری ها استفاده می شود.

این مقاله به معرفی داده کاوی وکاربردآن در صنعت پزشکی (پیش بینی بیماری) با استفاده از الگوریتم های داده کاوی به همراه نرم افزارهای مرتبط با آن پرداخته است.

کلمات کلیدی

داده کاوی ،درخت تصمیم،پیش بینی بیماری،دیابت

فهرست مطالب

عنوان صفحه

فهرست اشکال

عنوان صحفه

فصل اول

مقدمه

1-1مقدمه

از سال 1950 به بعد که رایانه ، در تحلیل و ذخیره سازی داده ها به کار رفت ، حجم اطلاعات ذخیره شده درآن پس از حدود 20 سال دو برابر شد و همزمان با پیشرفت فناوری اطلاعات ، حجم داده ها در پایگاه داده ها هر دو سال یک بار ، دو برابر شد و همچنان باسرعت بیش تری نسبت به گذشته حجم اطلاعات ذخیره شده بیش تروبیش تر می شود . با وجود شبکه جهانی وب ، سیستم های یکپارچه اطلاعاتی ، سیستم های یکپارچه بانکی ، تجارت الکترونیکی و ... لحظه به لحظه به حجم داده ها در پایگاه داده ها اضافه شده و باعث به وجود آمدن انبارهای (توده های) عظیمی از داده ها شده است ، به طوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده ها را بیش از پیش نمایان کرده است .

شدت رقابت ها در عرصه های علمی ، اجتماعی ، اقتصادی ، سیاسی و نظامی نیز اهمیت سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است . بنا براین نیاز به طراحی سیستم هایی که قادر به اکتشاف سریع اطلاعات مورد علاقه کاربران با تاکید بر حداقل مداخله انسانی باشند از یک سو و روی آوردن به روش های تحلیل متناسب با حجم داده های حجیم ازسوی دیگر ، به خوبـــــــی احســـاس می شود . در حال حاضر ، داده کاوی مهم ترین فناوری برای بهره وری موثر ، صحیح و سریع ازداده های حجیم است و اهمیت آن رو به فزونی است

داده کاوی پل ارتباطی میان علم آمار ، علم کامپیوتر ، هوش مصنوعی ، الگوشناسی ، فراگیری ماشین داده می باشد. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می باشد، به طریقی که این الگو ها و مدلها برای انسانها قابل درک باشند.

داده کاوی به صورت یک محصول قابل خریداری نمی باشد، بلکه یک رشته علمی و فرآیندی است که بایستی به صورت یک پروژه پیاده سازی شود .

داده ها اغلب حجیم می باشند و به تنهایی قابل استفاده نیستند،اما دانش نهفته در داده ها قابل استفاده می باشد.

بنابراین بهره گیری از قدرت فرآیند داده کاوی جهت شناسایی الگوها و مدلها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده ها و نهایتا تبدیل داده به اطلاعات، روز به روز ضروری تر می شود.

در داده کاوی معمولا به کشف الگوهای مفید از میان داده هااشاره می شود . منظور از الگوی مفید ، مدلی در داده ها است که ارتباط میان یک زیر مجموعه از داده ها را توصیف می کند و معتبر ، ساده ، قابل فهم و جدید است .

1-2 شرح وبیان مسئله:

امروزه علوم پزشکی وپزشکان با حجم زیاد داده ها روبه رو می باشند. از آنجایی که تشخیص بیماری همواره کار آسانی نیست بنابراین پزشک برای اتخاذ یک تصمیم مناسب،باید نتیجه ی آزمایش های بیمار وتصمیم هایی که در گذشته برای بیماران با وضعیت مشابه گرفته است، را بررسی کند .به عبارت دیگر پزشک نیازمند دانش و تجربه خواهد بود .ولی به دلیل تعداد زیاد بیماران و آزمایشهای متعدد هر بیمار، نیاز به یک ابزار خودکار برای کاوش در میان بیماران دیابتی احساس میشود .یکی از این روشهای مهم که برای استنتاج داده ها استفاده میشود، داده کاوی است.

1-3 هدف تحقیق:

به طور کلی می توان دلیل انجام این تحقیق را اینگونه بیان کرد:

- آشنایی با مفهوم داده کاوی
- آشنایی با کاربرد داده کاوی در تشخیص بیماری
- معرفی الگوریتم های داده کاوی در پیش بینی بیماری
- معرفی نرم افزار های داده کاوی
- استفاده ازاین روش در صنعت پزشکی

1-4 اهمیت وکاربرد نتایج تحقیق:

- استفاده از روش های مختلف داده کاوی در پزشکی
- آشنایی با الگوریتم های داده کاوی در پیش بینی بیماری وکاربرد آن در صنعت پزشکی
- استفاده از نرم افزار های داده کاوی در این صنعت

1-5 محدودیت:

- کمبود کتابهاومنابع در زمینه داده کاوی
- کمبود مقالاتی در زمینه پیش بینی بیماری بخصوص در تشخیص بیماری دیابت
- مقالاتی که در این زمینه وجود دارد بیشتر جنبه ی آماری دارد لذا نمی توان با چگونگی روش های داده کاوی در پیش بینی بیماری آشنایی پیدا کرد
- کمبود اطلاعات درباره ی نرم افزارهای داده کاوی وچگونگی کار با این نرم افزارها
- کمبود اطلاعات پزشکی بیماران دیابتی

1-6 تعریف عملیاتی واژگان:

داده کاوی: عبارت است از فرایند استخراج اطلاعات معتبر ، از پیش ناشناخته ، قابل فهم و قابل اعتماد از پایگاه داده های بزرگ

انبارداده: یک بانک اطلاعاتی بزرگ می‌باشد که از طریق آن کلیه داده‌های حال و گذشته یک سازمان جهت انجام عملیات گزارش گیری و آنالیز در دسترس مدیران قرار می‌گیرد.

خوشه بندی(Clustering):کشف و مستند سازی مجموعه‌ای از حقایق ناشناخته

دسته بندی(Classification): شناسایی الگوهای جدید

پیش بینی(Forecasting):کشف الگوهایی که بر اساس آنها پیش بینی قابل قبولی از رویدادهای آتی ارایه می‌شود،

شبکه عصبی:این تکنیک مدل های پیش بینی غیرخطی تولید می کند که یاد می دهد چگونه یک الگو با یک پروفایل خاص قابل تطبیق است،اما درباره ی علت رسیدن به این نتیجه ی خاص توضیحی ارایه نمی کند.

درخت تصمیم(Decision tree): یک ابزار برای پشتیبانی از تصمیم است که از درختان برای مدل کردن استفاده می‌کند. درخت تصمیم به طور معمول در تحقیق درعملیات استفاده می‌شود، به طور خاص در آنالیز تصمیم، برای مشخص کردن استراتژی که با بیشترین احتمال به هدف برسد بکار، می‌رود. استفاده دیگر درختان تصمیم، توصیف محاسبات احتمال شرطی است.

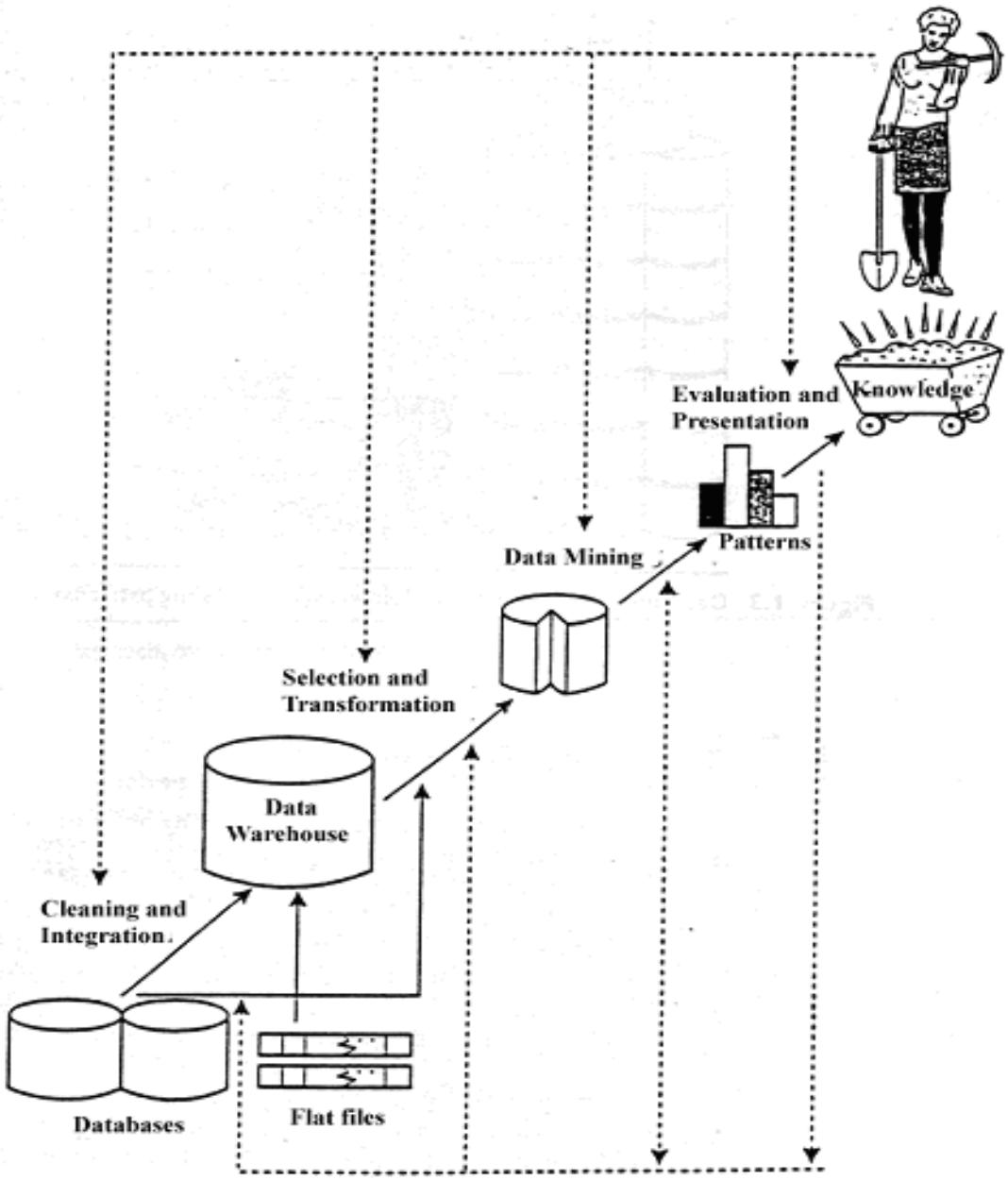
فصل دوم

مفاهیم داده کاوی

2-1تاریخچه[1]

با توجه به وجود اطلاعات ارزشمند در پایگاه های داده ای در اواخر دهه 80 میلادی ، تلاش برای استخراج و استفاده از اطلاعات پایگاه های داده ای شروع شد . داده کاوی فرایندی است که در آغاز دهه 90 پا به عرصه ظهور گذاشته و با نگرشی نو ، به مسئله استخراج اطلاعـــات از پایگـــاه داده ها می پردازد . در سال 1989 و 1991 کارگاه های کشف دانش از پایگاه داده ها توسط پیاتتسکی و همکارانش و در فاصله سال های 1991 تا 1994 کارگاه های فوق ، توسط فایاد و پیا تتسکی و دیگران برگزار شد . به طور رسمی اصطلاح داده کاوی برای اولین بار توسط « فیاض » در اولیـن کنفرانس بین المللی « کشف دانش و داده کاوی » در سال 1995 مطرح شد . از سال 1995 داده کاوی به صورت جدی وارد مباحث آمار شد.و در سال 1996 ، اولین شماره مجله کشف دانش از پایگاه داده ها منتشر شد .

امروزه کنفرانس های مختلفی دراین زمینه در سراسر دنیا برگزار می شود . داده کاوی حاصل تحول تدریجی در طول تاریخ بوده و از اوایل دهه 90 همزمان با همه گیر شدن استفاده از پایگاه های داده ای به عنوان یک علم مطرح شده است.



شکل2-1: داده کاوی به عنوان یک مرحله از فرآیند کشف دانش

2-2موضوع داده کاوی چیست؟

موضوع داده کاوی شناخت چیزهای جدید و با ارزش ، بالقوه مفید ، رابطه های منطقی و الگوهای موجود در داده ها است در جوامع مختلف یافتن الگو های مفید در داده ها با عناوین متعددی (مانند داده کاوی) بیان می شود . برای مثال از عنوان هایی نظیر استخراج دانش ، کشف اطلاعات ، برداشت اطلاعات ، پردازش الگوهای داده ها می توان نام برد .

عبارت « داده کاوی » توسط آمار شناسان ، محققان پایگاه های داده ها و سیستم های اطلاعات مدیریتی و جوامع بازرگانی به کار برده می شود . عبارت کشف دانش در پایگاه داده ها عموما برای اشاره به فرایند کلی کشف دانش مفید از داده هایی که داده کاوی گام مهمی دراین فرایند است ، مورد استفاده قرار می گیرد گام های دیگری در فرایند کشف دانش در پایگاه داده ها نظیر آماده کردن داده ها ، انتخاب داده ها ، تمیز کردن داده ها و درک درست از فرایند داده کاوی مـــــوجب می شود تا اطلاعاتی که برای ما مفید هستند از داده ها استخراج شوند . داده کاوی از تحلیل های سنتی داده ها و رویکردهای آماری نشات گرفته است به طوری که شامل فنون تحلیلی ای است که از شاخه های دیگری تشکیل شده است ، مانند :

تحلیل های عددی

- الگوهای سازگار و سطوحی از هوش مصنوعی مانند یادگیری ماشین
- شبکه های عصبی و الگوریتم های ژنتیک

با وجود این بسیاری از داد ه کاوی ها بر روش های سنتی و رویکردهای تحلیل داده های مبتنی بر فرضیه تکیه دارد . اساسا دو رویکرد برای داده کاوی وجود دارد که از لحاظ ایجاد و طراحی مدل و یافتن الگوها با هم فرق دارند اولین رویکرد که مربوط به ساخت مدل است (جدا از مشکلاتی که ذاتا در مجموعه داده های

بزرگ وجود دارد) مشابه روش های کاوشگرانه آماری مرسوم است. دراین حالت هدف این است تا خلاصه های کلی ازمجموعه ای ازداده هـــا برای شناخت وتــوضیح خصوصیت های اصلی شکل توزیع به دست آوریم . مثال هایی از این قبیل مدل ها شامل تحلیل خوش های بخشی از مجموعه داده ها مدل رگرسیونی برای پیشگویی و قاعده رده بندی با ساختار درختی است .

نوع دوم رویکرد داده کاوی ، رویکرد تشخیص الگو است . این رویکرد سعی بر آن دارد . تا انحراف هایی هرچند کوچک (از حد مطلوب) را تشخیص دهد (که در هر صورت حائز اهمیت هستند) ، تا الگوها و روند های غیر معمول نایان شود . مثال هایی نظیر الگو های نامعول (برای تشخیص کلاهبرداری) در استفاده از کارت های اعتباری و موضوع هایی که الگوهایی با ویژگی های نا مشابه با سایر الگو ها دارند از این نوع کاربد است . این دسته از راهبردها ست که موجب می شود تا داده کاوی به عنوان علم جستجوی اطلاعات با ارزش از بین توده عظیمی از داده ها به حساب آید . به طور کلی در پایگاه های داده ای کسب و کار (تجاری) ضعف درک الگو ها به خاطر پیچیدگی زیاد آن هاست . این پیچیدگی ها در اثر ناپیوسته بودن ، نامفهوم بودن و کامل نبودن به وجود می آیند. هر چند اکثر الگوریتم های داده کاوی می توانند اثر این گونه خصوصیت های نامربوط برا در تشخیص الگوی اصلی تمییز دهند ، ولی قدرت پیش گویی الگوریتم های داده کاوی با افزایش این انحراف ها کاهش می یابد .

2-3تعاریف داده کاوی

نگاهی به ترجمه لغوی داده کاوی به ما در درک بهتر این واژه کمک می کند . واژه لاتین Mine به معنای استخراج از منابع نهفته و با ارزش زمین اطلاق می شود . ادغام این کلمه با Data به معنی داده بر جستجویی عمیق از داده های قابل دسترس با حجم زیاد برای یافتن اطلاعات مفید که قبلا نهفته بودند ، تاکید دارد

داده کای دارای تعریف های مختلفی است این تعریف ها به مقدار زیادی به پیش زمینه ها و نقطه نظرهای افراد بستگی دارد . هر نویسنده ، محقق و کابر با توجه به پیش زمینه ها و نقطه نظر های افراد بستگی دارد . هر نویسنده ، محقق و کاربر با توجه به دیدگـــاه و نوع نگرش خود تعریف های مختلفی از داده کاوی ارائه کرده اند به عنوان مثال می توان به چند تعریف داده کاوی که در ادامه آمده است اشاره کرد:

- داده کاوی استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده می باشد
- داده کاوی علم استخراج اطلاعات مفید از پایگاه های داده یا مجموعه داده ای می باشد
- داده کاوی استخراج نیمه اتوماتیک الگوها، تغییرات، وابستگی ها، نابهنجاری ها و دیگر ساختارهای معنی دار آماری از پایگاه های بزرگ داده می باشد .
- داده کاوی عبارت است از فرایند استخراج اطلاعات معتبر ، از پیش ناشناخته ، قابل فهم و قابل اعتماد از پایگاه داده های بزرگ و استفاده از آن در تصمیم گیری در فعالیت های تجاری مهم.
- اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود
- داده کاوی یعنی جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها

2-4تفاوت داده کاوی و آنالیز های آماری

داده کاوی با آنالیز های متداول آماری متفاوت است؛در زیرمی توان برخی از اصلی ترین تفاوت های داده کاوی و آنالیز آماری را مشاهده نمود :

آنالیز آماری:

- آمار شناسان همیشه با یک فرضیه شروع به کار می کنند.
- آنها از داده های عددی استفاده می کنند.

- آمارشناسان باید رابطه هایی را ایجاد کنند که به فرضیه آنها مربوط است.
- آنها می توانند داده های نابجا و نادرست را در طول آنالیز مشخص کنند.
- آنها می توانند نتایج کار خود را تفسیر و برای مدیران بیان کنند.

داده کاوی:

- به فرضیه احتیاجی ندارد.
- ابزارهای داده کاوی از انواع مختلف داده ، نه تنها عددی می توانند استفاده کنند.
- الگوریتمهای داده کاوی به طور اتوماتیک روابط را ایجاد می کنند.
- داده کاوی به داده های صحیح و درست نیاز دارد.
- نتایج داده کاوی نسبتا پیچیده می باشد و نیاز به متخصصانی جهت بیان آنها به مدیران دارد.

جهت درک بهتر تفاوت داده کاوی و آنالیزهای آماری به مثال زیر که در مورد شناخت کلاهبرداری های شرکت بیمه می باشد، توجه کنید.

روش آنالیز آماری:

یک مفسر ممکن است متوجه الگوی رفتاری شود که سبب کلاهبرداری بیمه گردد. بر اساس این فرضیه، مفسر به طرح یک سری سوال می پردازد تا این موضوع را بررسی کند. اگر نتایج حاصله مناسب نبود، مفسر فرضیه را اصلاح می کند و یا با انتخاب فرضیه دیگری مجددا شروع می کند. این روش نه تنها وقت گیر است بلکه به قدرت تجزیه و تحلیل مفسر نیز بستگی دارد. مهمتر از همه اینکه این روش هیچ وقت الگوهای کلاهبرداری دیگری را که مفسر به آنها مظنون نشده و در فرضیه جا نداده ، پیدا نمی کند.

روش داده کاوی:

یک مفسر سیستم های داده کاوی را ساخته و پس از طی مراحلی از جمله جمع آوری داده ها، یکپارچه سازی داده ها به انجام عملیات داده کاوی می پردازد. داده کاوی تمام الگوهای غیرعادی را که از حالت عادی و نرمال انحراف دارند و ممکن است منجر به کلاهبرداری شوند را پیدا می کند.

نتایج داده کاوی حالت های مختلفی را که مفسر باید در مراحل بعدی تحقیق کند، نشان می دهند. در نهایت مدل های به دست آمده می توانند مشتریانی را که امکان کلاهبرداری دارند، پیش بینی نمایند.

2-5کاربردهای داده کاوی

از کاربردهای کلاسیک داده کاوی است که می توان به موارد زیر اشاره کرد :

1-خرده فروشی :

- تعیین الگوهای خرید مشتریان
- تجزیه و تحلیل سبد خرید بازار
- پیشگویی میزان خرید مشتریان از طریق پست(فروش الکترونیکی)

2-بیمه :

- تجزیه و تحلیل دعاوی
- پیشگویی میزان خرید بیمه نامه های جدید توسط مشتریان

3-پزشکی :

- تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی
- تعیین میزان موفقیت روشهای درمانی در برخورد با بیماریهای سخت

- تشخیص بیماریها براساس انواع اطلاعات (تصاویر پزشکی، مشخصات بیمار احتمالی)
- تشخیص ناهنجاریهائی که توسط انسان به سختی قابل تشخیص خواهند بود

4-بانکداری :

- پیش بینی الگوهای کلاهبرداری از طریق کارتهای اعتباری
- تشخیص مشتریان ثابت
- تعیین میزان استفاده از کارتهای اعتباری بر اساس گروههای اجتماعی

5-حوزه کاربردی فضا و سفرهای فضائی

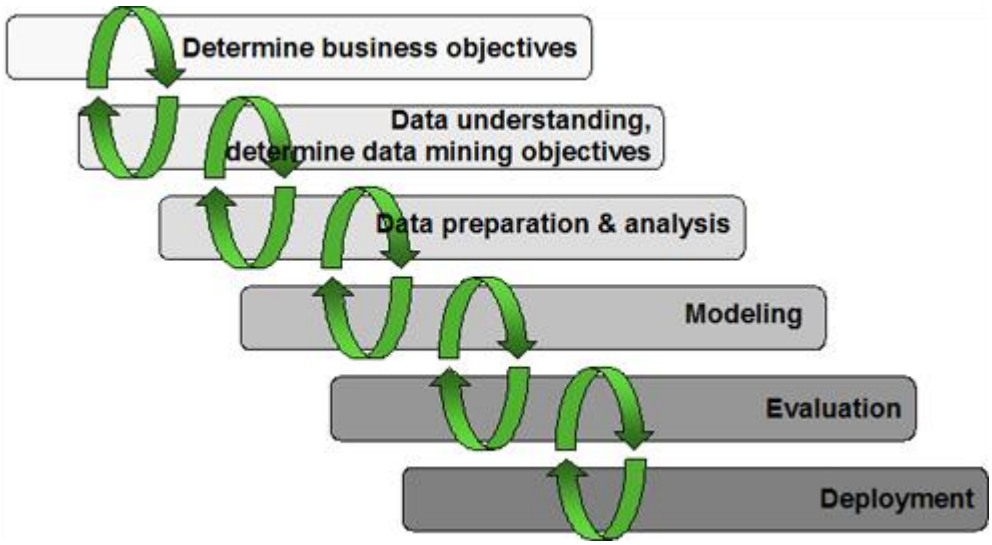
- حجم بسیار زیادی از اطلاعات
- نویز بسیار بالا
- ارزش بسیار زیاد دانش قابل استخراج
- پردازش اطلاعات جمع آوری شده از فضا
- پردازش اطلاعات مربوط به سفینه های فضائی
- ارائه دانش مفید برای اتخاذ تصمیم نهائی جهت پرتاب یا عدم پرتاب یک سفینه به فضا

2-6چند مثال در موردمفهوم داده کاوی

- یکی از نمونه های بارز داده کاوی را می توان در فروشگاه های زنجیره ای مشاهده نمود، که در آن سعی می شود ارتباط محصولات مختلف هنگام خرید مشتریان مشخص گردد. فروشگاه های زنجیره ای مشتاقند بدانند که چه محصولاتی با یکدیگر به فروش می روند.

- برای مثال طی یک عملـیات داده کاوی گستـرده در یـک فروشـگاه زنجیره ای در آمریکای شمالی که بر روی حجـم عظیمـی از داده های فروش صورت گرفت، مشخص گردید که مشتریانی که تلویزیون خریداری می کنند، غالبا گلدان کریستالی نیز می خرند.
- نمونه مشابه عملیات داده کاوی را می توان در یک شرکت بزرگ تولید و عرضه پوشاک در اروپا مشاهده نمود، به شکلی که نتایج داده کاوی مشخص می کرد که افرادی که کراوات های ابریشمی خریداری می کنند، در همان روز یا روزهای آینده گیره کراوات مشکی رنگ نیز خریداری می کنند.
- به روشنی این مطلب قابل درک است که این نوع استفاده از داده کاوی می تواند فروشگاه ها را در برگزاری هوشمندانه فستیوال های فروش و نحوه ارائه اجناس به مشتریان یاری رساند.
- نمونه دیگر استفاده از داده کاوی در زمینه فروش را می توان در یک شرکت بزرگ دوبلاژ و تکثیر و عرضه فیلم های سینمایی در آمریکای شمالی مشاهده نمود که در آن عملیات داده کاوی، روابط مشتریان و هنرپیشه های سینمایی و نیز گروه های مختلف مشتریان بر اساس سبک فیلم ها (ترسناک، رمانتیک، حادثه ای و ...) مشخص گردید. بنابراین آن شرکت به صورت کاملا هوشمندانه می توانست مشتریان بالقوه فیلم های سینمایی را بر اساس علاقه مشتریان به هنرپیشه های مختلف و سبک های سینمایی شناسایی کند.
- از دیگر زمینه های به کارگیری داده کاوی، استفاده بیمارستانها و کارخانه های داروسازی جهت کشف الگوها و مدلهای ناشناخته تاثیر دارو ها بر بیماری های مختلف و نیز بیماران گروه های سنی مختلف را می توان نام برد.
- استفاده از داده کاوی در زمینه های مالی و بانکداری به شناخت مشتریان پر خطر و سودجو بر اساس معیار هایی از جمله سن ، درآمد، وضعیت سکونت، تحصیلات، شغل و غیره می انجامد.

2-7مراحل داده کاوی



شکل2-2 مراحل داده کاوی

2-7-1مرحله اول:Business Understanding

این مرحله مهمترین مرحله فرایند می باشد. در ابتدا بـاید صورت مسئله دانسته شود تا پروژه داده کاوی صورت پذیرد.همچنین باید تاثیرگذارهای بر روی پروژه مشخص شوند که چه کسانی می باشند. سپـس باید دانش بـاشد تا چگونگی عمل نیز مشخص شود.

2-7-2مرحله دوم:Data Understanding

این مرحله مربوط به مفهوم داده ها می باشد. شامل مراحل زیر می باشد:

- جمع آوری داده های اولیه واصلی

- شرح وتوصیف داده ها
- کاوش داده ها
- تحقیق در مورد کیفیت داده ها

1-2-7-2جمع آوری داده ها:

مسئله اصلی در این قسمت این است که :

((ما چه داده هایی را احتیاج داریم؟))

- این داده ها کجا هستند؟
- بزرگی داده های مورد نیاز چقدر باشد؟
- چه مدت طول می کشد تا به داده ها دسترسی پیدا کنیم؟
- آیا روش خاص ومنحصربفردی برای جمع آوری داده ها است؟
- آیا داده های بدست آمده مفید،مفهومی،موثر و بهره ور هستند؟

یکی از سوال هایی که جهت جمع آوری داده مطرح شد ،این بود که داده ها کجا هستند؟

منابع مورد نیاز داده ها شامل:

فایلهای Flat

Database ها

Database های نامتجانس وناهمگون

Database های ناشناس ونامشخص

Database موروثی و

Datawarehouse انبار داده ها است.

انبار داده:Data Warehouse (DWH)

سیستمی است که عمل تلفیق درآن انجام می گیرد.قابل تغییر نیست. به مدیران در گرفتن تصمیم گیری بهتر کمک می کنند. در این سیستم چند خاصیت وجود دارد؟

1- به مسائل به خصوصی درجنبه استراتژیک می پردازد.(مشتریان،محصولات)

2- پس از ورود اعداد به سیستم می توان اعداد را خانه تکانی کرد.(یکسان سازی کدها، نام ها و..)

3-پویا است وبایداطلاعات جدید وارد آن شود.

معماری DWH :

اعداد در سیستم operative وجود دارندواعدادممکن است درچندتا از این DWH ها باشند. باید اول اعداد را تعریف کرده ، ببینیم درکجا قراردارند و بعداعدادمورد نیاز را به DWH میانی می آوریم و بعد درمرحله Staging اینکار انجام می شود و وقتی اعدادآماده شد ، Loud شده و به DWH می رود. سپس بعد از خانه تکانی ، با اعداد تمیز با سیستم ها وابزارهای Olap یا mining یا Reporting عمل می کنیم.

2-2-7-2 بحث شرح وتوصیف داده ها:

برخی از اندازه گیری های شخصیت داده ها شامل:

تعداد مشاهدات :

observation یا مشاهدات در جاهای متفاوت با عناوین مختلفی نام برده شده که از آن جمله می توان به این موارد اشاره کرد: pattern, point, Case, data, object, entity, event, instance, record, sample,…

تعداد صفات : Attribute

هر مشـاهده به وسیله یک یا چند صفت شـرح داده می شود. پس تعـداد صفات حتما باید کمتر از تعداد مشاهدات باشد. صفات یک مشاهده برای تعریف نوع وخاصیت مشاهده لازم وضروری است.

نام های دیگر Attribute به این شرح است: Feature, Field, Variable, …

انواع صفات: انواع صفات بوسیله انواع مقیاس های اندازه گیری اعداد تعریف می شوند.

انواع صفات از نظر مقیاس اندازه گیری:

Ratio	داده های نسبتی
Nominal	داده های اسمی
Ordinal	داده های ترتیبی
Interval	داده های فاصله ای

مقادیر اسمی:

مانند نژاد .آیا این شخص زرد پوست است یا نه؟ فقط در همین حد می باشد و نمی توان روی آن عملیاتی انجام داد.

مقادیر ترتیبی :

برای تمیز دادن هر مشـاهده از دیگرمشاهدات است.

A=B or A=B

وهمچنین ترتیب ورتبه مشاهدات را نیز مشخص می کند. (بیشتر است یا کمتر، بهتر است یا بدتر و ...)

A>B or A<B

مقادیر فاصله ای:

علاوه بر حالات قبل عمل تفاضل را نیز می توان بر روی داده ها انجام داد. در این حالت صفر، صفر مطلق نیست.

بعنوان مثال در مورد درجه حرارت هوا ، می توان گفت که این مقدار درجه هوا گرمترشده . ولی درجه حرارت صفر به این معنا نیست که هوا گرما وسرما ندارد.

مقادیر نسبتی :

تمام خصوصیات مقیاس فاصله ای را دارد.بعلاوه آنکه صفر معنای کامل ومطلق دارد. مثلا اگر گفتیم درآمد صفر است ،یعنی واقعا هیچ پولی وجود ندارد.

انواع دیگر دسته بندی صفات:

discrete اعداد گسسته

continuous اعداد پیوسته

اعدادگسسته : مقادیر محدود (مانند تعداد بچه) یا نامحدود قابل شمارش (مانند شماره اعداد یا فراوانی) هستند،

اغلب با اعداد طبیعی نشان داده می شوند ، حالت خاص آن اعداد دوتایی binary می باشد.

اعدادپیوسته : اعدادحقیقی هستند. تمام مقادیر بین دو مقدار را هم می پذیرند(مانند وزن)

پارامترهای آماری ای که برای خلاصه کردن داده ها مورد نیاز است شامل موارد زیر است:

- فراوانی
- میانگین میانه
- مد
- ماکزیمم داده ها
- مینیمم داده ها
- دامنه یابرد داده ها
- واریانس

- انحراف معیار
- میانگین انحرافات

3-7-2 مرحله سوم:Data Preparation

این مرحله مربوط به آماده سازی داده ها می باشد و شامل مراحل زیر می باشد:

- انتخاب داده ها
- تمیز کردن داده ها
- تبدیل داده ها
- تلفیق داده ها

بصورتی که کدگذاری و نام گذاری داده ها حالت استاندارد ویکسان داشته باشد.

1-3-7-2 Data selectingانتخاب داده

در دو بخش انجام می گیرد: یکی زمانی است که تعداد صفات را کم می کنیم و دیگری زمانی که با کم کردن مشاهدات داده ها را انتخاب می کنیم که ما در اینجا به بخش دوم می پردازیم:

کم کردن تعداد مشاهدات به سه روش می باشد:

- نمونه گیری Sampling
- نمونه گیری هوشمند Intelligent sampling
- یادگیری برای صرفنظر Learn to forget

درآمار،نمونه گیری تصادفی است که داده ها به صورت تصادفی انتخاب می گردند. ولی این تصادفی انتخاب کردن ممکن است که زیـان زیادی برای ما داشته باشد و ما خیلی از داده هـای با ارزشمان را از دست بدهیم. به

عـنوان مثال اگر اخراج یکسری از افراد یک شرکت به صورت نمونه گیری تصادفی باشد، ممکن است یکسری ازکارمندهای حرفه ای وخبره را ازدست بدهیم. عمـل نمونه گیری به این دلیل است کـه حجم بالای اطلاعات وداده ها قابل پردازش نیست. حجم نمونه باید به اندازه بهینه باشد. حجم نمونه به عنوان پارامتراضافی مدل است و میتوان آنرا آنقدرتغییرداد تابه حالت مطلوب برسیم.

نمونه گیری هوشمند:

در این حالت طوری نمونه گیری انجام می شود که طبق قوانین ومقرراتی باشد و ما اطلاعات اصلی را از دست ندهیم.

انواع روشهای نمونه گیری: با جایگذاری / بدون جایگذاری

درنمونه گیری تصادفی اساس براین است که هر کدام از مشاهدات احتمال معادل ومعلوم داشته باشند که بتوانند در نمونه گیری ما انتخاب شوند.

وقتی تعداد جامعه خیلی زیاد باشد می توان از نمـونه گیری بدون جایگذاری استفاده کرد. روش انتخاب نمونه :

$$K=\frac{\text{تعدادافراددرجامعه}}{\text{تعدادافراددرنمونه}}$$

مشاهده اول بصورت تصـادفی بعنوان نقطه شـروع انتخاب می گردد و بعد برای مشاهدات بعدی به هـر مشاهده به اندازه k اضافه می کنیم و به این ترتیب کل نمونه انتخاب می گردد.

نمونه گیری طبقه بندی:

زمانی این نمونه گیری انجام میشودکه یک جامعه نامتجانس داشته باشیم. به این صورت است که ابتداجامعه راطبقه بندی می کنیم و سپس دو حالت داریم: یکی آنکه از کل طبقه ها نمونه گیری تصادفی انجام می دهیم و دیگر آنکه از هر طبقه به تعداد مساوی نمونه می گیریم.

4-7-2مرحله چهارم :Modelling

این مـرحله مربوط به مـدلسازی پس از جمـع آوری داده ها و پیش بینی می باشد .

پیش بینی : به طور کلی مثل وضعیت رده بندی است .

X هایی داریم که بوسیله صفاتی نمایش داده می شوند. می خواهیم سیستمی درست کنیم که برای ما پیش بینی کند. متغیر هدف داریم که می خواهیم آنرا پیش بینی کنیم. متغیر هدف یک متغیر پیوسته است.

بنابراین یکسری عدد به روشهای مختلفی جمع آوری کـرده ایم و جهت مقـایسه وارزیابـی مدل ها در اختیار داریم.

جهت این کار باید اعداد را به دوگروه تقسیم کنیم. مسئله اول چگونگی تقسیم داده ها است.

یک روش تقسیم داده، با توجه به حجم زیاد داده وتجربه خودمان است.

معمولا می توان 70٪ داده هارا برای تولید مـدل و30٪ آنرا برای تست مدل تقسیم کرد که این حالت برای زمانی خوب است که اعداد زیاد باشد. ولی اگر تعداد داده ها کم باشد وما بیائیم 30٪ داده ها را برای تست بگـذاریم ، در واقـع 30٪ داده ها را از دست داده ایم.

زمانی که تعداد اعداد کم باشد، روش Cross Validation ، (اعتبار سنجی متقابل) به کار می رود

5-7-2مرحله پنجم :Evaluation

در این مرحله مدل را ارزیابی می کنیم. ببینیم آیا به هدف رسیده ایم یا نه؟ درآن قسمتهایی که به هدف نرسیده ایم،بعضی جاها را تکرارکنیم یا بعضی وقتها ممکن است مجبور به تغییرهدف شویم ویا گاهی مجبور به تغییر اعـداد اولیه شویم.

6-7-2مرحله ششم :Deployment

این مرحله،مربوط به چگونگی استفاده ازمدل است. زمانی به این مرحله می رویم که به هدف رسیده باشیم. این مرحله به شرح ذیل است:

- گسترش برنامه
- نگهداری وقوت برنامه
- تولیدگزارش نهایی
- تجدیدنظر ونشریه کردن پروژه

8-2مفاهیم اساسی در داده کاوی

1-8-2 **Bagging**: این مفهوم برای ترکیب رده بندی های پیش بینی شده از چند مدل به کار می رود.فرض کنیدکه قصدداریدمدلی برای رده بندی پیش بینی بسازیدو مجموعه داده های مورد نظرتان کوچک است.شمامی توانید نمونه هایی(با جایگزینی) را از مجموعه داده ها انتخاب و برای نمونه های حاصل ازدرخت رده بندی (مثلا RT&C وCHAID)استفاده نمایید.به طورکلی برای نمونه های مختلف به درخت های متفاوتی خواهید رسید.سپس برای پیش بینی با کمک درخت های متفاوت به دست آمده از نمونه ها ،یک رای گیری ساده انجام دهید.رده بندی نهایی ، رده بندی ای‌خواهد بود که درخت های مختلف آنرا پیش بینی کرده اند .

Boosting 2-8-2: این مفهوم برای تولید مدل‌های چندگانه (برای پیش بینی یا رده بندی)به کار می‌رود. Boosting نیزاز روش RT&C یا CHAID استفاده وترتیبی از classifier ها را تولید خواهد کرد .

Meta-Learning 3-8-2 : این مفهوم برای ترکیب پیش بینی‌های حاصل از چند مدل به کار می‌رود.و هنگامی که انواع مدل‌های موجود در پروژه خیلی متفاوت هستند، کاربرد دارد. فرض کنید که پروژه داده کاوی شما شامل Tree classifierها نظیر RT&Cو CHAID، تحلیل خطی و شبکه های عصبی است.هر یک از کامپیوترها،رده بندی هایی رابرای نمونه هاپیش بینی کرده اند.تجربه نشان می‌دهدکه ترکیب پیش بینی های چند روش دقیق تراز پیش بینی های هریک از روشهاست.پیش بینی های حاصل از چند classifier را می توان به عنوان ورودی meta-linear مورد استفاده قرار داد. meta-linear پیش بینی هارا ترکیب می کند تا بهترین رده بندی پیش بینی شده حاصل شود.

9-2 عناصر داده کاوی

توصیف و کمک به پیش بینی دو کارکرد اصلی داده کاوی هستند. تحلیل داده مربوط به مشخصه های انتخابی متغیرها؛ از گذشته و حال، و درک الگو مثالی از تحلیل توصیفی است. برآورد ارزش آینده یک متغیر و طرح ریزی کردن روند مثالی از توانایی پیشگویانه داده کاوی است.

برای عملی شدن هریک از دو کارکرد فوق الذکر داده کاوی، چند گام ابتدایی اما مهم باید اجرا شوند که از این قرارند:

1. انتخاب داده ها
2. پاک سازی داد ها
3. غنی سازی داده ها

4. کد گذاری داده ها

با دارا بودن هدف کلی در مطالعه، انتخاب مجموعه داده های اصلی برای تحلیل، اولین ضرورت است. رکوردهای لازم میتواند از انبار داده ها و یا بانک اطلاعاتی عملیاتی استخراج شود. این رکوردهای داده جمع آوری شده؛ اغلب از آنچه آلودگی داده ها نامگذاری شده است رنج می برند و بنابراین لازم است پاکسازی شوند تا از یکدستی فرمت (شکلی) آنها اطمینان حاصل شود، موارد تکراری حذف شده و کنترل سازگاری دامنه بعمل آید. ممکن است داده های گردآوری شده از جنبه های خاصی ناقص یا ناکافی باشند. در این صورت داده های مشخصی باید گردآوری شوند تا بانک اطلاعات اصلی را تکمیل کنند. منابع مناسب برای این منظور باید شناسایی شوند. این فرایند مرحله غنی سازی داده ها را تکمیل میکند. یک سیستم کدگذاری مناسب معمولا" جهت انتقال داده ها به فرم ساختاربندی شده جدید؛ متناسب برای عملیات داده کاوی تعبیه میشود .

10-2 تکنیک های داده کاوی[2]

1-10-2 دسته بندی

متداولترین تکنیک است و یکسری نمونه های از پیش تعیین شده راشامل می شود که برای توسعه مدل به کار می رود که بتواند انواعی ازموارد ثبت شده را دسته بندی نماید .این شیوه غالباً از درخت تصمیم گیری یا الگوریتم های دسته بندی شبکه استفاده می کند، فرایند شامل یادگیری و رده بندی است. در یادگیری اطلاعات آموزشی با الگوریتم دسته بندی تحلیل می شود و اطلاعات برای برآورد دقیق قواعد به کار می رود، اگر دقت آن در حد مناسبی باشد می توان از آن برای موارد جدید استفاده نمود. الگوریتم دسته بندی کننده موارد آموزشی از نمونه های از پیش دسته بندی شده برای تعیین مجموعه ای از پارامترهای مورد نیاز برای تفکیک صحیح استفاده می کند، سپس الگوریتم این پارامترها رابه مدلی به نام دسته بندی تبدیل می کند.

2-10-2خوشه بندی

دادهها ممکن است حاوی ساختارهای پیچیده ای باشند که حتی بهترین تکنیکهای داده کاوی هم قادر به استخراج الگوهای معنی دار ازآنها نباشند .خوشه بندی راهی را برای یافتن ساختاردادههای پیچیده فراهم می آورد و سیگنالهای رقابتی ناهماهنگ را به اجزایشان تفکیک میکند .خوشه بندی به عمل تقسیم جمعیت ناهمگن به تعدادی اززیرمجموعه ها یا خوشه های همگن گفته میشود **.نقطه تمایز خوشه بندی از دسته بندی:** در دسته بندی براساس یک مدل هر کدام از داده ها به دسته ای از پیش تعیین شده اختصاص مییابد .این دسته ها از طریق پژوهش های پیشین تعیین گردیده اند.لیکن در روش خوشه بندی هیچ دسته از پیش تعیین شده ای وجود ندارد وداده ها صرفا بر اساس تشابه،گروه بندی می شوند وعناوین هر گروه نیز توسط کاربرتعیین می گردد.

3-10-2 رگرسیون گیری

تکنیک رگرسیون گیری را میتوان برای پیش بینی پذیرفت.از تحلیل رگرسیون می توان برای مدل سازی روابط یک یا چند متغیرمستقل ووابسته استفاده نمود.

در استخراج اطلاعات متغیرهای مستقل ویژگی هایی هستند که قبلاً شناخته شده و متغیرهای وابسته مربوط به چیزی هستند که می خواهیم پیش بینی کنیم .متاسفانه، خیلی از مسائل واقعی به راحتی پیش بینی نمی شوند بنابراین، تفکیک های پیچیده تر ممکن است برای پیش بینی مقادیر آینده ضروری باشد، انواعی از مدل ها غالباً برای رگرسیون گیری و دسته بندی به کار می روند برای نمونهCART(دسته بندی و درخت رگرسیون گیری)الگوریتم درخت تصمیم گیری را می توان برای ایجاد درخت دسته بندی ودرخت رگرسیون گیری استفاده نمود.شبکه های عصبی می توانند مدل های رگرسیون گیری و دسته بندی را ایجاد کنند.

4-10-2 تجمع وهمبستگی

تجمع و همبستگی معمولاً برای یافتن یک آیتم تکراری به کار می رود و یافته ها را در میان مجموعه اطلاعات و سیستم تنظیم می کند .این مورد به تاجران کمک می کند تا تصمیمات مشخصی بگیرند مانند طراحی کاتالوگ، بازاریابی و تحلیل رفتار مشتری این الگوریتم لازم است بتواند قوانینی را ایجاد نماید که مقدار اعتباری کمتر از 1 را داشته باشد .به هر حال تعداد قوانین انجمنی احتمالی برای مجموعه داده های مشخص خیلی بزرگ است و نسبت بالای قوانین با مقدار کم همراه است.

5-10-2درخت تصمیم گیری

درخت تصمیم درختی است که در آن نمونه ها را به نحوی دسته بندی می کند که از ریشه به سمت پائین رشد میکنند و در نهایت به گره های برگ میرسد .هر گره داخلی یاغیر برگبایک ویژگی مشخص می شود،این ویژگی سوالی را در رابطه با مثال ورودی مطرح می کند.

در هر گره داخلی به تعداد جواب های ممکن با این سوال شاخه وجود دارد که هر یک با مقدارآن جواب مشخص می شود. برگ های این درخت با یک کلاس و یا یک دسته از جواب ها مشخص میشوند .علت نامگذاری آن با درخت تصمیم این است که این درخت فرآیند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان می دهد.

6-10-2 الگوریتم ژنتیک

الگوریتم ژنتیک جست وجوی اکتشافی است که فرآیند تکامل طبیعی را پیروی می کند،این اکتشاف برای ایجاد راه حل های مفید در بهینه سازی وجست وجوی مسائل به طور مداوم استفاده شده است.الگوریتم های ژنتیک متعلق به کلاس بزگتری از الگوریتم های تکاملی هستند،که برای ایجاد راه حل های بهینه سازی مسائل از روش های الهام گرفته از تکامل طبیعی،مانند وراثت،جهش،گزینش وتقاطع استفاده می کنند،وبه این صورت است که طبیعت،افرادقوی تر را برای زندگی برمی گزیند.

7-10-2 شبکه های عصبی مصنوعی

شبکه های عصبی مصنوعی که معمولا به عنوان شبکه های عصبی نام برده می شوند یک الگوی ریاضی مبنی برسیستم زیستی است.سیستمهای عصبی یک الگوریتم برای بهینه سازی ویادگیری آزادانه بر اساس مفاهیم الهام گرفته ازتحقیق در ماهیت مغز می باشند.مغز با استفاده از قابلیتی شناخته شده به عنوان نورون اجزا ساختاری خود را سازماندهی می کند،در نتیجه محاسبات معینی را بسیارسریع تر از کامپیوتر دیجیتال انجام می دهد.

7-7-2 گام نهایی فرایند داده کاوی، گزارش دادن است

گزارش شامل تحلیل نتایج و کاربردهای پروژه، درصورت بکارگیری آنها، است . و متن مناسب، جداول و گرافیکها را در خود جای می دهد. بیشتر اوقات گزارش دهی یک فرایند تعاملی است که تصمیم گیرنده با داده ها در پایانه کامپیوتری بازی میکند و فرم چاپی برخی نتایج واسطه محتمل را برای عملیات فوری بدست می آورد.

داده کاوی در تولید چهار نوع دانش ذیل مفید است:

- دانش سطحی- کاربردهای SQL

- دانش چند وجهی -کاربردهای OALP

- دانش نهان (تشخیص الگو و کاربردهای الگوریتم یادگیری ماشینی)

- دانش عمیق (کاربردهای الگوریتم بهینه سازی داخلی)

11-2 تکنولوژی های مرتبط با داده کاوی[3]

1-11-2 انبارداده

معمولا داده هایی که در داده کاوی مورد استفاده قرار می گیرند از یک انبار داده استخراج می گردند و در یک پایگاه داده یا مرکز داده ای ویژه برای داده کاوی قرار می گیرند.

اگر داده های انتخابی جزیی از انبار داده ها باشند بسیار مفید است چون بسیاری از اعمالی که برای ساختن انباره داده ها انجام می گیرد با اعمال مقدماتی داده کاوی مشترک است و در نتیجه نیاز به انجام مجدد این اعمال وجود ندارد ، از جمله این اعمال پاکسازی داده ها می باشد.

پایگاه داده مربوط به داده کاوی می تواند جزیی از سیستم انبار داده ها باشد و یا می تواند یک پایگاه داده جدا باشد.

ساختار یک انبارداده مشخصات زیر را نشان میدهد:

1-وابستگی به زمان:

رکوردها بر اساس یک برچسب زمانی نگهداری میشوند. وابستگی زمانی حاصل در ایجاد صفحات زمانی مفید است که درک ترتیب زمانی وقایع را تسهیل میکند.

2-غیر فرار بودن:

رکوردهای داده در انبار داده ها هرگز بطور مستقیم روزآمد نمیشوند. برای هر تغییری در ابتدا داده های عملیاتی روزآمد میشوند و سپس بگونه ای مقتضی به انبار داده منتقل میشوند. این مساله ثبات داده ها را برای

استفاده های وسیعتر تضمین میکند.

3-تمرکز موضوعی:

داده ها از بانکهای اطلاعاتی عملیاتی بصورت گزینشی به انبار داده منتقل میشوند. این استراتژی به ایجاد یک انبار داده بر اساس یک مطلب یا موضوع خاص کمک میکند و بنابراین کاوش انبار داده ها برای پرس و جوهای موضوعی با سرعت بیشتری انجام میشود.

4-یکپارچگی:

داده ها بگونه ای کامل سازماندهی شده اند تا با حذف موارد تکراری و چند عنوانه یکپارچگی رکوردها حفظ شود ؛ به ایجاد ارجاع های متقابل کارآمد بین رکوردها کمک نموده و ارجاع دهی را تسهیل نماید.

واضح است که انبار داده اساسا" برای پرس و جوهای پشتیبان تصمیم گیری ساخته شده است. بر این اساس سازماندهی وعملیات انبار داده چنان طراحی شده اند تا نیازهای اطلاعاتی روزمره یا معمولی را پاسخگو باشند. بدلیل حجم بسیار بالای چنین پایگاه اطلاعاتی یک سیستم کامپیوتری پیشرفته برای عملیات انبارسازی داده ها لازم است. همچنین یک بانک اطلاعات مجزا شامل ابرداده که مشخصه هایی نظیر نوع، فرمت، مکان و پدیدآورندگان داده های ذخیره شده در یک انبار داده ها را توصیف میکند نیز برای کمک به کاربران و مدیران داده ها ساخته میشود. مشخص شد که انبار داده بدلیل اندازه و تنوعش، اگر مبتکرانه پردازش شود میتواند به تولید اطلاعاتی منجر شود که در وهله اول آشکار نیستند. با انتخاب متناسب داده ها، بکار گرفتن فنون مختلف غربال کردن و تفسیر زمینه ای، داده ذخیره شده میتوانست منجر به کشف الگوها یا رابطه هایی شود که بینش نویی به تصمیم گیرنده دهد. ذکر این نکته لازم است که داده کاوی در اصل لزوما" نیاز به سازماندهی یک انبار داده ندارد.

OLAP 2-11-2

بسیاری فکر می کنند که داده کاوی و OLAP دو چیز مشابه هستند در این بخش سعی می کنیم این مسئله را بررسی کنیم و همانطور که خواهیم دید این دو ابزار های کاملا متفاوت می باشند که می توانند همدیگر را تکمیل کنند.

OLAP جزیی از ابزارهای تصمیم گیری می باشد.سیستم های سنتی گزارش گیری و پایگاه داده ای آنچه را که در پایگاه داده بود توضیح می دادند حال آنکه در OLAP هدف بررسی دلیل صحت یک فرضیه است.

بدین معنی که کاربر فرضیه ای در مورد داده ها و روابط بین آنها ارائه می کند و سپس به وسیله ابزار OLAP با انجام چند Query صحت آن فرضیه را بررسی می کند.

اما این روش برای هنگامی که داده ها بسیار حجیم بوده و تعداد پارامترها زیاد باشد نمیتواند مفید باشد چون حدس روابط بین داده ها کار سخت و بررسی صحت آن بسیار زمانبر خواهد بود.

تفاوت داده کاوی با OLAP در این است که داده کاوی برخلاف OLAP برای بررسی صحت یک الگوی فرضی استفاده نمی شود بلکه خود سعی می کند این الگوها را کشف کند.

درنتیجه داده کاوی و OLAP می توانند همدیگر را تکمیل کنند و تحلیل گر می تواند به وسیله ابزار OLAP یک سری اطلاعات کسب کند که در مرحله داده کاوی می تواند مفید باشد و همچنین الگوها و روابط کشف شده در مرحله داده کاوی می تواند درست نباشد که با اعمال تغییرات در آنها می توان به وسیله OLAP بیشتر بررسی شوند.

12-2 محدودیت ها

کاربرد داده کاوی با چند عامل محدود شده است. اولین مورد به سخت افزار و نرم افزار لازم و موقعیت بانک اطلاعاتی مربوط میشود . برای مثال در هند، داده های غیر مجتمع که برای کاربردهای داده کاوی لازم است ممکن است به فرم دیجیتالی در دسترس نباشد. در دسترس بودن نیروی انسانی ماهر در داده کاوی نیز مسأله مهم دیگری است. محرمانه بودن رکوردهای مراجعان ممکن است در نتیجه پردازش داده های مبتنی بر داده کاوی آسیب پذیر شود. کتابداران و مؤسسات آموزشی باید این مسأله را در نظر داشته باشند؛ چرا که در غیر

اینصورت ممکن است گرفتار شکایات قانونی گردند.

محدودیت دیگراز ضعف ذاتی نهفته در ابزارهای نظری ناشی میگردد. ابزارهایی مانند یادگیری ماشینی و الگوریتمهای ژنتیکی بکار گرفته شده در فعالیتهای داده کاوی به مفاهیم وفنون منطق و آمار بستگی دارد. در این حد نتایج به روش مکانیکی تولید شده و بنابراین به یک بررسی دقیق نیاز دارند. اعتبار الگوهای بدست آمده به این طریق؛ باید آزمایش شود. چرا که در بسیاری موارد روابط علل و معلول مشتق شده؛ از برخی استدلالات غلط ذیل رنج میبرند.

از آنجایی که مطالعه الگوها و استخراج روابط میان رکوردها مستلزم کاربرد منطق قیاسی و استقرایی است فرد باید مراقب اشتباهاتی که عموما" رخ میدهد باشد. برای مثال بحثهای قیاسی یا استقرایی، تا زمانیکه وضعیت درست بودن فرضیه آزمایش نشود چیزی درباره درست یا غلط بودن نتایجشان نمی گویند. طبیعتا، نتایج تولید شده ماشینی ممکن است از چنین نقایصی رنج ببرند.

فصل سوم

کاربرد داده کاوی در پزشکی

1-3 داده کاوی درعرصه ی سلامت[4]

تحول در صنعت سلامت به واسطه این هدف واحد که «چگونه سازمان های سلامت هزینه ها را کاهش وکیفیت را افزایش دهند وهمچنان رقابتی باقی بمانند؟» به پیش میرود واین مقوله همواره یک چالش بزرگ محسوب می شود.بهبود کیفیت در صنعت سلامت را می توان به واسطه ی نیروهای محرکی که برآن تاثیرگذار است بهتر تعریف نمود واز جمله این نیروهای محرک داده های سلامت است؛به عبارت دیگر در هرنوع برنامه ی بهبود کیفیت متمرکزبر بیمار،داده ها قلب آن برنامه به حساب می آید.داده ها درعصر امروزی یعنی عصر اطلاعات،عمده ترین دارایی برای سازمان های سلامت بوده وموفقیت سازمان های سلامت درگروی جمع آوری،ذخیره وتحلیل آن هاست.با این وجود،جمع آوری وذخیره ی میزان زیادی از داده ها می تواند یک نوع اتلاف محسوب شود؛مگر این که داده ها به شکل سودمند استفاده شده وتبدیل به یک منبع مالی برای سازمان گردد.برای تبدیل این ارزش بالقوه به اطلاعات استراتژیک،بسیاری از سازمان ها به داده کاوی روی آورده اند؛چرا که به واسطه ی داده کاوی امکان کشف روابط،روندها والگوهای مخفی بین داده ها و دستیابی به دانش نوین در زمینه ی چالش های آشکارونهان سازمان میسر خواهد شد.

مفهوم کشف دانش از داده ها بیش از یک دهه است که در محیط های مالی-تجاری در حال استفاده می باشد ودر علوم مدیریت ارتباطات،مهندسی،وب کاوی،تحلیل جرایم وپزشکی جای خود را باز کرده است.اگرچه کشف دانش باهدف شناسایی اختلاس مالی وارد عرصه ی سلامت شد،اما به تدریج در حوزه ی بالینی نیز مورد استفاده قرارگرفت. این مهم ناشی از تغییرسریع هوشیاری نسبت به اطلاعات در حوزه ی سلامت است.

صنعت سلامت به طور مستمردرحال تولید میزان زیادی از داده ها می باشدوافرادی که با این نوع داده ها مواجه هستند،دریافته اند که بین جمع آوری تا تفسیرآنها شکاف وسیعی وجود دارد؛حوزه ی به نسبت جوان ودرحال رشدداده کاوی درسلامت ازجمله شیوه هایی است که میتواند این صنعت را از تحلیل عمیق این داده ها بهره مند سازد وبه توسعه ی تحقیقات پزشکی وتصمیم گیری های علمی درزمینه ی تشخیص ودرمان منتج شود.

داده کاوی به کندی اما به طور فزاینده ای برای رفع مشکلات متعدد در کشف دانش ودربخش سلامت به کار گرفته شده است.4مورد از مهم ترین دلایل رشدکند این علم درسلامت؛حساسیت علم پزشکی وگره خوردن آن باجان انسان ها(تفاوت جزیی درالگوهای داده کاوی می تواند به تغییر تعادل بین مرگ وزندگی بیانجامد)،سردرگمی درتعریف داده کاوی (گاه ایجاد یک طرح ساده از پایگاه داده های پزشکی به غلط به عنوان الگوی حاصل از داده کاوی مطرح می شود)،حریم شخصی ومحرمانگی داده های سلامت ودر نهایت مهم ترین

چالش این است که اگر فرض براین باشد که نتایج داده کاوی به طور کامل قابل اعتماد است؛تغییر عادت ارایه دهندگان مراقبت از پزشکی سنتی به پزشکی مبتنی بر شواهد دشوار است.

با این وجود امروزه بخش سلامت بیش ترین نیاز را به داده کاوی پیدا کرده است وحرکت ازپزشکی سنتی به سمت پزشکی مبتنی برشواهد از جمله مواردی است که می تواند موکد این امر باشد.در ادامه رایج ترین استراتژی های داده کاوی،تکنیک های داده کاوی ونمونه هایی از کاربردهای آن در سلامت بیان می گردد

2-3 استراتژی های داده کاوی

به طور کلی هدف داده کاوی،یادگیری وآموختن از داده ها است وبر این اساس ودو دسته کلی از استراتژی های داده کاوی شامل یادگیری نظارت شده ویادگیری فاقد نظارت وجود دارد.شیوه های یادگیری نظارت شده زمانی به کار می رود که ارزش متغیرهای ورودی برای ما شناخته شده باشد.یافتن مدل های پیش بینی خطا در مطالبات بیمه ی یک موسسه ی سلامت،نمونه ای از استراتژی یادگیری نظارت شده است ؛در این استراتژی مدل ها و ویژگی ها برای ما شناخته شده وبا هدف پیش بینی داده ها وکشف اطلاعات به کار می رود اما درشیوه ی یادگیری فاقد نظارت،ویژگی ها ومدل های خطاهای مطالبات شناخته شده نیست،اما الگوها و خوشه های حاصل از داده کاوی منجر به کشفیات جدید می شود.

3-3 نمونه هایی از کاربرد های داده کاوی درسلامت

- **داده کاوی در تشخیص های غیر تهاجمی:**برخی از اقدامات تشخیصی وآزمایشگاهی برای بیماران،تهاجمی وهزینه بر و در عین حال رنج آور هستند.به عنوان مثال بافت برداری از گردن رحم به منظور تشخیص سرطان گردن رحم از جمله این موارد است.

Thangavelوهمکاران از طریق الگوریتم های خوشه بندی به تحلیل بیماران مبتلا به سرطان گردن رحم پرداختند ونتایج پیش گیری کننده تری را نسبت به عقاید پزشکی کنونی پیدا کردند.

- **داده کاوی در تعیین نوع درمان:** به کارگیری داده کاوی بر روی داده های پزشکی دستاوردهای حیاتی واثرگذاری را در انتخاب نوع درمان مناسب ونجات جان انسان ها به ارمغان آورده است.به عنوان مثال

دربیمارستان شهید هاشمی نژاد تهران برای تعیین نوع درمان سنگ حالب از راهکار داده کاوی استفاده می شود. در این بیمارستان یک الگوریتم درختی وجود دارد که پزشک بر اساس آن درمانی را که میزان موفقیت بالاتری برای بیمار دارد انتخاب می کند و بررسی ها نشان داده است که میزان عملکرد این الگوریتم 77درصد بوده و بسیار بهتر از عملکرد الگوریتم بیمارستانی و مدل ذهنی پزشکان است.

- **داده کاوی در شناسایی عوارض جانبی داروها:**برخی از داروها که در ابتدا به عنوان داروهای غیر مضر برای انسان به تایید رسیده اند،بعد از استفاده ی عموم در طولانی مدت اثرات زیان بار خود را نشان می دهند.

اداره ی دارو وغذای آمریکا با استفاده از داده کاوی به کف دانش درباره ی عوارض جانبی داروها در پایگاه داده ی خود پرداخته است. الگوریتم مورد استفاده در این روش MGPSنام دارد که توانسته است با موفقیت 67درصد عوارض جانبی داروها را 5سال زودتر از شیوه ی سنتی شناسایی کند.

- **داده کاوی در پرونده ی الکترونیک سلامت** :در حال حاضر مطالعات متعددی موکد این است که تکنیک های داده کاوی ابزار موثری را برای شناسایی الگوهای مهم سلامت از درون پرونده های پزشکی فراهم می کند.

پرونده های سلامت کامپیوتری به واسطه در برداشتن مجموعه ای از داده ها درباره تشخیص،درمان،اقدامات آزمایشگاهی ودارویی به طور بالقوه منبع غنی از دانش هستند. اگرچه کشف دانش از انبوه داده ها در آن ها برای انسان غیرممکن نیست،اما امری دشوار است وداده کاوی بهترین شیوه برای حل این چالش می باشد.

- **داده کاوی در رتبه بندی بیمارستان ها:**رتبه بندی بیمارستان ها وبرنامه های بهداشتی می تواند بر مبنای اطلاعات گزارش شده توسط ارایه دهندگان مراقبت باشد، بنابراین گزارش دهی استاندارد برای مقایسه ی معنی دار بیمارستان ها ورتبه بندی آ نها بسیار مهم است .از جمله شیو ه های استاندارد کردن این گزارشات استفاده از تکنیک های داده کاوی می باشد،به عنوان مثال اگر کدهای ICD(کدهای عوامل خطر) اختصاص یافته به پرونده های بیمار با تکنیک های داده کاوی مانند خوشه بندی و ارتباطات همراه شود، می تواند به ایجاد گزارشاتی بیانجامد که با تعیین نرخ واقعی

میزان ناخوشی ها،مرگ و میر و سایر شاخص های کیفیت در رتبه بندی بیمارستان ها مورد استفاده واقع شود.

- **داده کاوی در بهره برداری از خدمات سلامت:** به واسطه ی داده کاوی در این برنامه، بیمارستان ها دردستیابی به متغیرهای کلیدی در پیش بینی بهره برداری ازخدمات، بهبود پیامدهای کیفیت، پیش بینی رفتارهای آتی بیماران و بهبود برنامه های درمانی آ نها، شناسایی بیماران پرخطر و مدیریت مراقبت آن ها توانمند می شوند.

با استفاده از داده کاوی و مد لسازی داده ها می توان بیماران با شرایط پرخطر را شناسایی کرد .در واقع داده کاوی با ارایه ی اطلاعات به ارایه دهندگان مراقبت، آن ها را در شناسایی بیماران پرخطر به گونه ای که بتوان کیفیت مراقبت آن ها را بهبود داد و از مشکلات آتی آ نها جلوگیری کرد،کمک می کند و باطراحی مداخله ی مناسب منجر به کاهش پذیرش های بیمارستانی می شود.به عنوان مثال تکنیک های مدل سازی پیش بینی کننده ی داده کاوی در رابطه با مدیریت بیماری دیابت منجر به ارتقای کیفیت وکاهش هزینه ی بیماران مبتلا به دیابت می شود.

3-4 مقایسه الگوریتم‌های هوشمند در شناسایی بیماری دیابت[5]

بیماری دیابت یکی از شایع‌ترین بیماری‌های حاضر دنیا شناخته شده است که علی‌رغم گستردگی شیوع این بیماری هنوز روشی به منظور ریشه‌کن کردن و از بین بردن آن در دنیا شناخته نشده است هرچند که روش‌های مختلفی جهت تشخیص و کنترل آن در حال حاضر مورد استفاده قرار می‌گیرد. از جمله عوارضی که به دنبال مبتلا شدن افراد به این بیماری گریبان‌گیر آن‌ها خواهد شد می‌توان به گرفتگی عروق قلبی و در نوع پیشرفته آن به نابینایی، قطع اعضای بدن، اختلالات فکری و غیره اشاره نمود.

بیماری دیابت را از نظر تقسیم‌بندی می‌توان به دو نوع وابسته به انسولین(IDDM) که در این نوع لوزالمعده شخص مبتلا به دیابت قادر به ترشح انسولین نمی‌باشد و یا نوع غیر وابسته به انسولین(NIDDM) که در آن لوزالمعده شخص مبتلا به دیابت قادر به تولید و ترشح انسولین می‌باشد اما میزان جذب آن در بدن بسیار اندک است.

مشکل عمده‌ای که در رابطه با بیماری دیابت وجود دارد عدم تشخیص به موقع و یا به‌طور کلی ضعف در تشخیص این بیماری است که این ضعف نیز به‌دلیل عدم انتخاب الگوی مناسب توسط پزشک و یا عدم استفاده

مناسب از الگوهای استاندارد است. بنابراین پیاده‌سازی روشی که بتواند هر فرد را در تشخیص صحیح ابتلا یا عدم ابتلا به این بیماری یاری رساند می‌تواند گام مهمی در جهت پیشگیری و کنترل این بیماری به‌خصوص در مراحل ابتدایی آن باشد.

1-4-3 دسته‌بندی کننده Bagging

Bagging(متراکم شدن خودکار) توسط لئوبریمن در سال 1994 پیشنهاد شد که برای بهبود دادن رده‌بندی توسط ترکیب کردن رده‌بندی‌های مجموعه‌های آموزشی به‌طور تصادفی تولید شده، می‌باشد.
این روش یک متا الگوریتم می‌باشد که برای بهبود دادن یادگیری ماشین رده‌بندی و مدل‌های پسرفتی بر حسب پایداری و دقت رده‌بندی می‌باشد. این روش همچنین واریانس را کاهش داده و به دوری از اورفیتینگ کمک می‌کند. اگر چه این روش در درخت تصمیم به‌کار می‌رود اما می‌تواند در هر نوع مدل استفاده شود. Bagging یک حالت مخصوص از روند مدل میانگین می‌باشد.
یک مجموعه آموزشی استاندارد D به اندازه n را فرض کنید، Bagging توسط نمونه‌گیری به‌طور یکنواخت و باجایگزینی مثال‌ها از D، m مجموعه آموزشی جدید D_i با اندازه $n \geq n'$ تولید می‌شود. نمونه‌گیری با جایگزینی این امکان را می‌دهد که بعضی از مثال‌ها امکان تکرار در هر D_i را داشته باشند. اگر n=n' باشد لذا برای n بزرگ، مجموعه D_i انتظار داشتن ٪63.2 از مثال‌های بی‌همتای D را دارد و بقیه مثال‌ها تکراری می‌باشند. این نوع نمونه‌گیری به عنوان نمونه‌گیری هورراه‌انداز شناخته می‌شود. M مدل برای استفاده کردن m نمونه‌های خودکار بالا گنجانیده شده و این مدل‌ها توسط متوسط‌گیری خروجی یا رای‌گیری(برای رده‌بندی) ترکیب می‌شوند.
از آنجایی‌که این روش چندین پیشگویی کننده را میانگین می‌گیرد، لذا برای بهبود مثال‌های خطی مفید نمی‌باشد.
Bagging مفهومی برای ترکیب رده‌بندی‌های پیش‌بینی شده از چند مدل به‌کار می‌رود.

2-4-3 دسته‌بندی کننده Naive Bayse

هدف ما در این بخش ایجاد قانونی‌ست که قادرمان سازد اعضای بعدی را در یک دسته قرار دهیم و اینکار را تنها با داشتن بردارهایی از متغییرهای توصیف کننده اجسام بعدی می‌توانیم انجام دهیم. مشکلاتی از این نوع که دسته‌بندی نظارت شده نام دارند و روش‌های بسیاری برای ایجاد چنین قوانینی بوجود آمده‌اند. یک روش بسیار مهم روش بیز ساده است که بیز سطحی، بیز ساده، و بیز مستقل نیز نامیده می‌شود. این روش به دلایل متعددی اهمیت دارد. ساخت آن بسیار ساده است و نیازی به برنامه‌های تخمین پارامتر تکرار شونده پیچیده ندارد. یعنی از آن می‌توان برای مجموعه داده‌های وسیع استفاده کرد و در نهایت این روش معمولاً فوق‌العاده عمل می‌کند. ممکن است بهترین دسته‌بندی کننده ممکن در یک کاربرد خاص نباشد اما اغلب می‌توان به قوی بودن و عملکرد عالی آن اطمینان کرد.

الف- قوانین اصلی در Naive Bayse

برای راحتی تفسیر در اینجا دو دسته را در نظر می‌گیریم که به‌صورت i=0 and 1 رتبه‌بندی می‌شود. هدف ما استفاده از یک مجموعه اعضای اولیه با عضویت‌های دسته شناخته شده(مجموعه آموزشی) برای ایجاد یک امتیاز است به‌طوری‌که امتیازات بالاتر با اعضای دسته 1 مرتبطند و امتیازات کوچک‌تر با اعضای دسته 0 در ارتباطند. بنابراین دسته‌بندی با مقایسه این امتیازات با یک آستانه به‌دست می‌آید. اگر ما P(i|x) را این احتمال تعریف کنیم که یک عضو با بردار مکان $x = (x_1,\ldots,x_p)$ به دسته i تعلق داشته باشد، هر عملکرد یکنواخت P(i|x) یک امتیاز مناسب ایجاد خواهد کرد. بطور ویژه نسبت P(1|x) P(0|x) مناسب خواهد بود. احتمال ابتدایی به مامی‌گوید که P(i|x) را به نسبت f(x|i) / p(i) تجزیه کنیم که در آن f(x|i) توزیع شرطی x برای اجسام دسته i است و p(i) احتمال تعلق یک عضو به دسته i است در صورتیکه ما چیز بیشتری درباره آن ندانیم (احتمال اولیه دسته دسته i). یعنی نسبت به این صورت در می‌آید:

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)} \qquad (1)$$

برای استفاده از این نسبت در دسته بندی‌ها ما نیاز به تخمین f(x|i) و P(i) داریم.

اگر مجموعه آموزشی یک نمونه تصادفی از جمعیت کل باشد P(i) را می‌توان مستقیماً از جمعیت اجسام دسته i محاسبه کرد. برای محاسبه f(x|i) روش بیز ساده اینطور در نظر می‌گیرد که اعضای x مستقل‌اند، $\prod_{j}^{p} = 1^{f(x_j|i)}$ سپس هر یک از توزیعات یکنواخت $f(x_j \mid i), j = 1,\ldots, p;$ را جداگانه اندازه‌گیری می‌کند. بنابراین مشکل چند شکلی P بعدی به مشکلات اندازه‌گیری یک شکلی کاهش می‌یابد. تخمین یک شکلی آشنا و ساده بوده و به اندازه‌های مجموعه یادگیری کوچکتری برای دستیابی به اندازه‌های درست نیازمند است. این

یکی از جذابی‌های خاص و نیز منحصربه فرد روش بیز ساده است: تخمین ساده، و بسیار سریع بوده و به برنامه-های اندازه‌گیری تکرار شونده پیچیده نیاز ندارد.

اگر توزیع حاشیه‌ای $f(x_j \mid i)$ مجزا باشد و هر x_j تعداد اندکی از ارزش‌ها را شامل شود، اندازه $f(x_j \mid i)$ یک اندازه گیرنده سابقه نمای چند جمله‌ایست که تنها نسبت اعضای دسته i را که در یک سلول قرار می‌گیرند محاسبه می‌کند. اگر $f(x_j \mid i)$ ادامه داشته باشند، یک روش رایج تقسیم آن‌ها به فواصل کوچک‌تر و استفاده مجدد از اندازه گیرنده چند جمله‌ای است اما انواع پیشرفته‌تر آن که بر پایه اندازه‌های مداوم هستند(مانند اندازه‌های هسته) نیز مورد استفاده قرار می‌گیرند نسبت داده شده به این صورت در می‌آید:

$$\frac{P(1 \mid x)}{P(0 \mid x)} = \frac{\prod_{j=1}^{p} f(x \mid 1)P(1)}{\prod_{j=1}^{p} f(x \mid 0)P(0)} = \frac{p(1)}{p(0)} \prod_{j=1}^{p} \frac{f(x_j \mid 1)}{f(x_j \mid 0)} \quad (2)$$

اکنون با یادآوری این‌که هدف ما تنها معرفی یک امتیاز بود که بطور یکنواخت به $P(i|x)$ مرتبط باشد، می‌توانیم از آن لگاریتم بگیریم.

$$\ln \frac{P(1 \mid x)}{P(0 \mid x)} = \ln \frac{p(1)}{p(0)} + \sum_{j=1}^{p} \ln \frac{f(x_j \mid 1)}{f(x_j \mid 0)} \quad (3)$$

اگر ما $w = \ln(\frac{f(x_j \mid 1)}{f(x_j \mid 0)})$ و یک $k = \ln(\frac{p(1)}{p(0)})$ مداوم را تعریف کنیم می‌بینیم که نسبت معادله بالا شکل ساده زیر را می‌گیرد:

$$\ln \frac{P(1 \mid x)}{P(0 \mid x)} = k + \sum_{j=1}^{p} w_j, \quad (4)$$

بطوری‌که دسته‌بندی ساختار ساده‌ای دارد.

فرض استقلال x_j هر دسته مشخص در مدل بیز ساده ممکن است زیادی سخت‌گیرانه به نظر برسد. با این وجود در واقع فاکتورهای مختلفی ممکن است وارد بازی شوند که به این معناست که فرض آنقدر که به نظر می‌رسد مضر نیست.اول این‌که، یک مرحله انتخاب متغییر اولیه معمولاً اتفاق می‌افتد که در آن متغییرهای بسیار مرتبط با یکدیگر در زمینه‌هایی که ممکن بود به شباهت جدایی بین دسته‌ها کمک کنند رفع شدند. دوم این‌که، در نظر گرفتن ارتباطات به عنوان صفر یک مرحله نظم‌دهی مشخص فراهم می‌کند که باعث کاهش واریانس مدل و دسته‌بندی‌های درست‌تر می‌شود. سوم این‌که، در برخی موارد وقتی متغیرها به هم مرتبط هستند سطح تصمیم‌گیری بهینه با سطح ایجاد شده تحت فرض مستقل تصادف می‌کند بطوری‌که ایجاد فرض به هیچ عنوان زیان‌بار نیست. چهارم این‌که، مسلماً سطح تصمیم ایجاد شده توسط مدل بیز ساده می‌تواند یک شکل غیر خطی پیچیده داشته باشد.

ب- نکات نهایی درباره بیز ساده

این مدل از قدیمی‌ترین الگوریتم‌های دسته‌بندی رسمی است و هنوز حتی در ساده‌ترین شکل بسیار موثر است. از این مدل در دسته‌بندی متون و جداسازی Spamها به‌طور گسترده استفاده می‌شود. بسیاری از اصطلاحات توسط جوامع آماری، استخراج اطلاعات، یادگیری ماشینی، و الگوشناسی در تلاش برای انعطاف بیشتر آن ارائه شده‌اند اما باید دانست که چنین اصطلاحاتی پیچیدگی‌هایی را ایجاب می‌کنند که باعث جدایی از سادگی اولیه می‌شود.

3-4-3دسته‌بندی کننده SVM

این الگوریتم از روش‌های یادگیری باناظر است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند.

این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی-تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته‌بندی کننده‌ی SVM دسته‌بندی خطی داده‌هاست و در تقسیم خطس داده‌ها سعی بر آن است خطی انتخاب شود که حاشیه اطمینان بیشتری داشته باشد. حل معادله‌ی پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد.قبل از تقسیم خطی برای این‌که ماشین بتواند داده‌های باپیچیدگی بالا را دسته‌بندی کند داده‌ها را باید به وسیله توابع phi به فضای با ابعاد خیلی بالاتر برد.

این الگوریتم دارای مبانی نظری قوی و بی‌نقصی می‌باشد، فقط به یک دوجین نمونه احتیاج دارد و به تعداد ابعاد مسأله حساس نمی‌باشد. همچنین متدهایی کارآمد برای یادگیری SVM به سرعت در حال رشد هستند. در یک فرایند یادگیری که شامل دو کلاس می‌باشد، هدف SVM پیدا کردن بهترین تابع برای طبقه‌بندی می‌باشد به نحوی که بتوان اعضای دو کلاس را در مجموعه داده‌ها از هم تشخیص داد. معیار بهترین طبقه‌بندی به‌صورت هندسی مشخص می‌شود، برای مجموعه داده‌هایی که به‌صورت خطی قابل تجزیه هستند. به‌طور حسی آن مرزی که به‌صورت بخشیاز فضا تعریف می‌شود یا همان تفکیک بین دو کلاس بوسیله hyperplane تعریف می‌شود. همین تعریف هندسی به ما اجازه می‌دهد تا کشف کنیم که چگونه مرزها را بیشینه کنیم ولو اینکه تعداد بیشماری hyperplane داشته باشیم و فقط تعداد کمی، شایستگی راه حل برای SVM دارند دلیل اینکه SVM روی بزرگ‌ترین مرز برای hyperplane پافشاری می‌کند این‌ست که قضیه قابلیت عمومیت بخشیدن

به الگوریتم را بهتر تامین می‌کند. این نه تنها به کارایی طبقه‌بندی و دقت آن روی داده‌های آزمایشی کمک می‌کند، فضا را نیز برای طبقه‌بندی بهتر داده‌های آتی مهیا می‌کند.

یکی از مشکلات SVM پیچیدگی محاسباتی آن است. با این‌حال این مشکل به‌طور قابل قبولی حل شده است. یک راه‌حل این است که یک مسأله بهینه‌سازی بزرگ را به یک سری از مسائل کوچک‌تر تقسیم کرد که هر مسأله شامل یک جفت با دقت انتخاب شده از متغیرها که مساله بطور موثر بتواند از آن‌ها بهره ببرد. این پروسه تا زمانی که همه این قسمت‌های تجزیه شده حل شوند ادامه خواهد داشت.

4-4-3 دسته‌بندی کننده Random Forest

از جمله دسته‌بندهایی که متد Bagging را به کار می‌گیرد Random Forest است که حاوی چندین درخت تصمیم می‌باشد و خروجی آن از خروجی‌های درخت‌های انفرادی به دست می‌آید. این الگوریتم متد Bagging را با انتخاب تصادفی ویژگی‌ها ترکیب می‌کند تا مجموعه‌ای از درخت‌های تصمیم با تنوع تحت کنترل ایجاد گردد. دقت بسیار بالای دسته‌بند از مزایای آن است، ضمن این‌که می‌تواند با تعداد زیاد ورودی به‌خوبی عمل کند.

5-4-3 دسته‌بندی کننده C4.5

الف- توصیف الگوریتم

C4.5 یک الگوریتم نیست بلکه مجموعه‌ای از الگوریتم‌هاست. همه روش‌های استنتاج درخت از گره ریشه آغاز می‌شوند که همه اطلاعات داده شده را ارائه می‌دهد و بطور بازگشتی اطلاعات را به مجموعه‌های کوچک‌تری تقسیم‌بندی می‌کند. و این‌کار را به وسیله آزمایش هر نسبت در هر گروه انجام می‌دهد. درختان فرعی، دسته‌بندی‌های مجموعه اطلاعات اصلی را نشان می‌دهند که آزمایشات ارزش‌گذاری نسبت‌های مشخص شده را تکمیل می‌کنند. این فرایند معمولاً تا خالص‌سازی مجموعه‌ها ادامه می‌یابد، یعنی همه نمونه‌ها در یک دسته قرار می‌گیرند و در این زمان رشد درخت متوقف می‌گردد.

ب- قوانین دسته‌بندی کننده‌ها

C4.5 یک لیست از قوانین در قالب یک فرم معرفی می‌کند. قوانین برای هر کلاس با هم گروه‌بندی می‌شوند یک نمونه با پیدا کردن اولین قانون، به موقعیت امن می‌رود و اگر قانونی برای نمونه پیدا نشد به کلاس پیش فرض می‌رود. فایده مجموعه قوانین C4.5 در مقدار زمان CPU و حافظه مورد نیاز است.

فصل چهارم

درخت تصمیم

وپیاده سازی نرم افزاروکا

4-1 مقدمه[6]

ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش بینی کننده می باشد که حقایق مشاهده شده در مورد یک پدیده را به استنتاج هایی در مورد مقدار هدف آن پدیده نقش می کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده ها، یادگیری درخت تصمیم نامیده می شود که یکی از رایج ترین روش های داده کاوی است.

هر گرهء داخلی متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گرهء برگ، با داشتن مقادیر متغیرها که با مسیری از ریشهء درخت تا آن گرهء برگ بازنمایی می شود، مقدار پیش بینی شدهء متغیر هدف را نشان می دهد..یک درخت تصمیم ساختاری را نشان می دهد که برگ ها نشان دهندهء دسته بندی و شاخه ها ترکیبات فصلی صفاتی که منتج به این دسته بندی ها را بازنمایی می کنند. یادگیری یک درخت می تواند با تفکیک کردن یک مجموعهء منبع به زیرمجموعه هایی براساس یک تست مقدار صفت انجام شود. این فرآیند به شکل بازگشتی در هر زیرمجموعهء حاصل از تفکیک تکرار می شود. عمل بازگشت زمانی کامل می شود که تفکیک بیشتر سودمند نباشد یا بتوان یک دسته بندی را به همهء نمونه های موجود در زیرمجموعهء بدست آمده اعمال کرد.

درختان تصمیم قادر به تولید توصیفات قابل درک برای انسان، از روابط موجود در یک مجموعهء داده ای هستند و می توانند برای وظایف دسته بندی و پیش بینی بکار روند. این تکنیک به شکل گسترده ای در زمینه های مختلف همچون تشخیص بیماری دسته بندی گیاهان و استراتژی های بازاریابی مشتری بکار رفته است.

این ساختار تصمیم گیری می تواند به شکل تکنیک های ریاضی و محاسباتی که به توصیف، دسته بندی و عام سازی یک مجموعه از داده ها کمک می کنند نیز معرفی شوند. داده ها در رکوردهایی به شکل $(x, y) = (x_1, x_2, x_3\ldots, x_k, y)$ داده می شوند. با استفاده از متغیرهای $x_1,x_2,..,x_k$ سعی در درک، دسته بندی یا عام سازی متغیر وابستهء Y داریم. انواع صفات در درخت تصمیم به دو نوع صفات دسته ای و صفات حقیقی بوده که صفات دسته ای، صفاتی هستند که دو یا چند مقدار گسسته می پذیرند (یا صفات سمبلیک) درحالی که صفات حقیقی مقادیر خود را از مجموعهء اعداد حقیقی می گیرند.

4-2 اهداف اصلی درخت‌های تصمیم‌گیری دسته‌بندی کننده

- داده‌های ورودی را تا حد ممکن درست دسته‌بندی کنند.
- دانش یادگیری شده از داده‌های آموزشی را به گونه‌ای عام سازی کنند که داده‌های دیده نشده را با بالاترین دقت ممکن دسته‌بندی کنند.
- در صورت اضافه شدن داده‌های آموزشی جدید، بتوان به راحتی درخت تصمیم‌گیری را گسترش داد(دارای خاصیت افزایشی باشند).
- ساختار درخت حاصل به ساده‌ترین شکل ممکن باشد.

4-3 گام‌های لازم برای طراحی یک درخت تصمیم‌گیری:

- انتخاب مناسبی برای ساختار درخت.
- انتخاب ویژگی‌هایی مورد نظر برای تصمیم‌گیری در هر یک از گره‌های میانی.
- انتخاب قانون تصمیم‌گیری یا استراتژی مورد استفاده در هر یک از گره‌های میانی.

4-4جذابیت درختان تصمیم

نواحی تصمیم پیچیدهء سراسری (خصوصاً در فضاهای با ابعاد زیاد) می توانند با اجتماع نواحی تصمیم محلی ساده تر در سطوح مختلف درخت تقریب زده شوند.

برخلاف دسته بندی کننده های تک مرحله ای رایج که هر نمونهء داده ای روی تمام دسته ها امتحان می شود، در یک دسته بندی کنندهء درخت، یک نمونه فقط روی زیرمجموعه های خاصی از دسته ها امتحان شده و محاسبات غیرلازم حذف می شود.

در دسته بندی کننده های تک مرحله ای، فقط از زیرمجموعه ای از صفات، برای تفکیک بین دسته ها استفاده می شود که معمولاً با یک معیار بهینهء سراسری انتخاب می شوند. در دسته بندی کنندهء درخت، انعطاف پذیری انتخاب زیرمجموعه های مختلفی از صفات در گره های داخلی مختلف درخت وجود دارد؛ به شکلی که زیرمجموعهء انتخاب شده به شکل بهینه بین دسته های این گره را تفکیک می کند. این انعطاف پذیری ممکن است بهبودی در کارایی را نسبت به دسته بندی کننده های تک مرحله ای ایجاد کند.

در تحلیل چندگونگی با تعداد صفات و دسته های زیاد، معمولاً نیاز به تخمین توزیع های ابعاد-زیاد یا پارامترهای خاصی از توزیع های دسته همانند احتمالات اولیه از یک مجموعهء داده های آموزشی کوچک می باشد. در این حالت مشکل ابعاد-بالا وجود دارد که امکان دارد در درخت دسته بندی کننده، با بکاربردن تعداد کمتری از صفات در هر گرهء داخلی بدون افت شدید کارایی، این مسئله حل شود.

4 -5بازنمایی درخت تصمیم

درختان تصمیم، نمونه ها را با مرتب کردن آنها در درخت از گرهء ریشه به سمت گره های برگ دسته بندی می کنند. هر گرهء داخلی در درخت، صفتی از نمونه را آزمایش می کند و هر شاخه ای که از آن گره خارج می شود متناظر یک مقدار ممکن برای آن صفت می باشد. همچنین به هر گرهء برگ، یک دسته بندی منتسب می شود. هر نمونه، با شروع از گره ریشهء درخت و آزمایش صفت مشخص شده توسط این گره و حرکت در شاخهء متناظر با مقدار صفت داده شده در نمونه، دسته بندی می شود. این فرآیند برای هر زیردرختی که گرهء جدید ریشهء آن می باشد تکرار می شود.

در حالت کلی، درختان تصمیم یک ترکیب فصلی از ترکیبات عطفی قیود روی مقادیر صفات نمونه ها را بازنمایی می کنند. هر مسیر از ریشهء درخت به یک، برگ متناظر با یک ترکیب عطفی صفات تست موجود در آن مسیر بوده و خود درخت نیز متناظر با ترکیب فصلی همهء این ترکیبات عطفی می باشد.

4- 6 مسائل مناسب برای یادگیری درخت تصمیم

انواع مختلفی از روش های یادگیری درخت تصمیم با قابلیت ها و نیازمندیهای گوناگونی توسعه یافته اند اما در حالت کلی، یادگیری درخت تصمیم در کل برای مسائلی با ویژگی های زیر مناسب است:

مسائلی که در آنها نمونه ها به شکل جفت های صفت-مقدار بازنمایی می شوند- در این گونه مسائل، نمونه ها با محموعهء ثابتی از صفات و مقادیر آنها بیان می شوند. ساده ترین وضعیت برای یادگیری درخت تصمیم، زمانی

است که هر صفت تعداد کوچکی از مقادیر ممکن گسسته بگیرد. هرچند با توسعه به الگوریتم پایه، امکان استفاده از صفات با مقادیر پیوسته نیز امکان پذیر است.

مثال: صفت: دما مقدار: {گرم، معتدل، خنک}

مسائلی که در آنها تابع هدف مقادیر خروجی گسسته دارند- می توان روش های درخت تصمیم به گونه ای گسترش داد که بتوانند توابعی با خروجی بیش از دو مقدار را یادگیری کنند. نوع دیگری از توسعه، امکان دستیابی به توابع هدف یادگیری با خروجی های مقدار پیوسته را امکان پذیر می سازد.

مثال: خروجی یک تابع هدف فرضی: {بلی، خیر}

ممکن است به توصیفات فصلی نیاز باشد- درختان تصمیم به شکل ذاتی عبارات فصلی را بازنمایی می کنند.

ممکن است داده های آموزشی حاوی خطا باشند- روش های یادگیری درخت تصمیم نسبت به وجود خطا در داده های آموزشی مقاوم هستند. (خطا در دسته بندی مثالهای آموزشی و خطا در مقادیر صفاتی که مثالهای آموزشی را تشریح می کنند.)

ممکن است داده های آموزشی حاوی صفات فاقد مقدار باشند- روش های درخت تصمیم می توانند زمانی که برخی مثالهای آموزشی مقادیر ناشناخته دارند نیز استفاده شوند.

نمونه هایی از مسائلی که دارای این ویژگی ها بوده و یادگیری درخت تصمیم به آنها بکارگرفته شده اند، مانند یادگیری دسته بندی بیماران پزشکی بوسیلهء بیماریهای آنها، دسته بندی عمل کردن اشتباه تجهیزات بوسیلهء دلایل آنها، و دسته بندی اعطای وام توسط احتمال عدم پرداخت آنها.

4-7مسائل در یادگیری درخت تصمیم

مسائل عملی در یادگیری درختان تصمیم

- تا چه عمقی درخت باید رشد داده شود
- بکاربردن صفات پیوسته
- انتخات یک معیار انتخاب صفت مناسب
- بکاربردن داده های آموزشی با صفات فاقد مقدار
- استفاده از صفات با هزینه های مختلف
- بهبود کارایی محاسباتی
- خطا در مثال های آموزشی
- یک درخت تصمیم اورفیت شده
- مثال های آموزشی ناکافی
- ممانعت از اورفیتینگ داده ها

4-8 اورفیتینگ داده ها

الگوریتم تصمیم هر شاخهء درخت را تا عمقی رشد می دهد که درخت بتواند مثال های آموزشی را کاملاً دسته بندی کند. هرچند این استراتژی مناسب می باشد، زمانی که در داده ها نویز وجود داشته باشد و یا تعداد مثال های آموزشی برای تولید نمونه ای از تابع هدف صحیح بسیار کم است می تواند باعث مشکلاتی شود. در هر یک از این حالات، الگوریتم تصمیم می تواند درختانی را تولید کند که مثال های آموزشی را اورفیت می کنند. وقتی درخت با نویز در داده ها سازگار شود اورفیتینگ رخ داده است و در این هنگام ممکن است روی داده های مجموعهء تست بدتر عمل کند.

اورفیتینگ یک مسئلهء عملی مهم برای یادگیری درختان تصمیم و بسیاری متدهای یادگیری دیگر می باشد. برای مثال با مطالعات آزمایشگاهی روی پنج وظیفهء یادگیری با داده های حاوی نویز غیرقطعی، نشان داده شد که اورفیتینگ دقت درخت تصمیم یادگیری شده را 10 تا 25 درصد کاهش می دهد.

- **روشهای موجود برای ممانعت از اورفیتینگ**

معمولاً از قبل نمی دانیم که چه نمونه هایی نامربوط هستند و این ممکن است که به نوع و زمینهء داده ها برگردد اما می توان از روشهای آماری ساده برای هشدار در زمان وقوع اورفیتینگ استفاده کرد مانند (تست Chi-Squared). روش های بکار رفته برای ممانعت از اورفیتینگ، عموماً روشهای هرس نامیده می شوند.

9-4 انواع روش های هرس کردن

1) **روش هرس از قبل**: روش هایی که در آنها، رشد درخت قبل از اینکه به نقطه ای برسد که کاملاً مثال های آموزشی را دسته بندی کند متوقف می شوند. (مشکل این روش ها در تعیین زمان توقف رشد درخت است.). نمونه هایی از معیارهای متوقف کردن رشد درخت:

- وقتی نفع اطلاعات کمتر از مقدار ثابت E باشد .
- از تست chi-squared برای آزمایش معناداری تقسیم بیشتر مثال های آموزشی استفاده می کنیم.
- معیار MDL[1]را بکار گرفته و در صورتی که رشد بیشتر باعث کاهش این معیار شود، اجازهء رشد درخت را می دهیم.

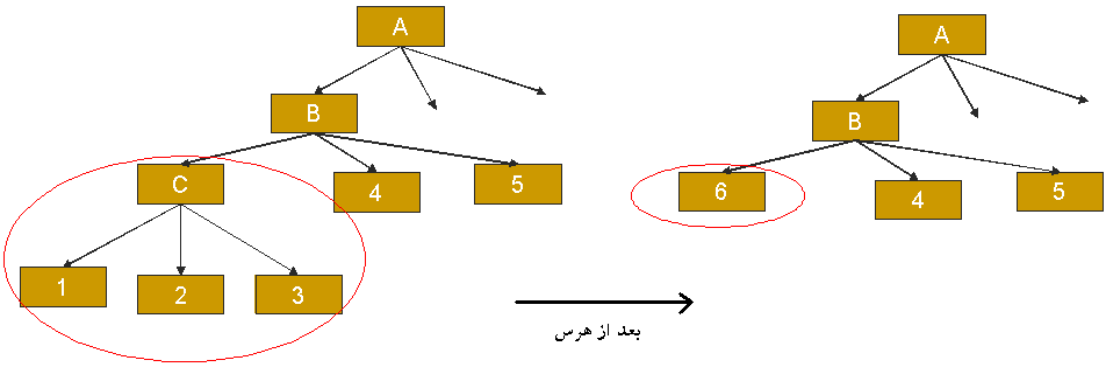
2) **روش هرس بعدی:** روشهایی که اجازه می دهند درخت کاملاً ساخته شود و داده ها را اورفیت کند، سپس آن را هرس می کنند. (این روش ها در عمل موفق تر هستند.)

- دو تکنیک برای هرس بعدی استفاده می شود:

الف) جایگزینی زیردرخت: تمام زیر درخت با یک گرهء برگ جایگزین می شود. این تکنیک درخت را

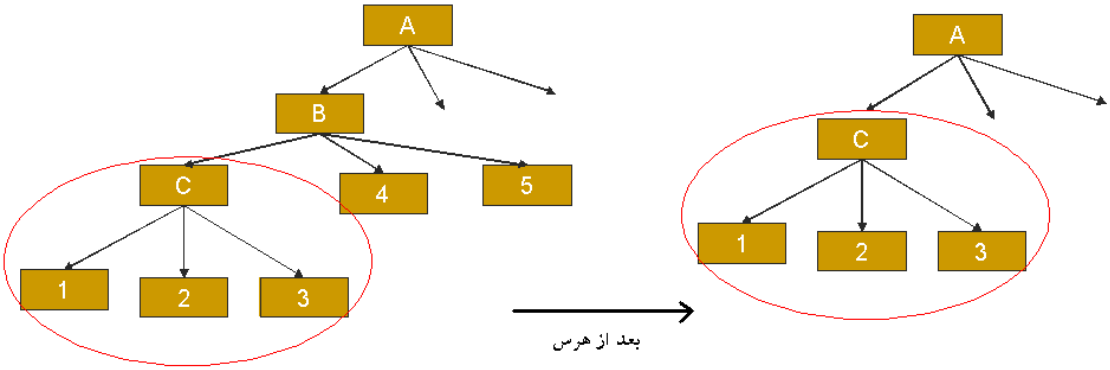
[1] MDL (minimum description length) = minimize size(tree) + size(misclassifications(tree))

کمی عام تر می کند اما دقت را افزایش می دهد.



شکل 4-1 جایگزینی زیر درخت

ب) بالا کشیدن زیردرخت: تمام زیردرخت به جای گره ای دیگر جایگزین می شود.



شکل 4 -2 بالا کشیدن زیر درخت

4-10 عام سازی درخت

در ساخت درخت دو انتخاب وجود دارد: 1) طراحی درختی که تمام نمونه های آموزشی را به شکل صحیح دسته بندی کند (درخت کامل) و انتخاب کوچکترین درخت یا 2) ساخت درختی که کامل نبوده اما کمترین نرخ خطای ممکن در دسته بندی نمونه های تست را دارد (در عمل این روش مطلوب تر است).

در هر دو حالت کوچک نگاه داشتن اندازهء درخت تاحد امکان مطلوب است. دلایل آن این است که درختان کوچکتر زمان تست و نیاز حافظهء کمتری داشته و تمایل به عام سازی بهتر نمونه های نادیده دارند (چون به بی قاعدگی های آماری و طبیعت داده های آموزشی کمتر حساس هستند.)

11-4مزایای درختان تصمیم نسبت به روش های دیگر داده کاوی

- قوانین تولید شده و به کارگرفته شده قابل استخراج و قابل فهم می باشند.
- درخت تصمیم، توانایی کار با داده های پیوسته و گسسته را دارد. (روشهای دیگر فقط توان کار با یک نوع را دارند. مثلاً شبکه های عصبی فقط توان کار با داده های پیوسته و قوانین رابطه با داده های گسسته)
- درخت تصمیم از نواحی تصمیم گیری ساده استفاده می کند.
- مقایسه های غیرضروری در این ساختار حذف می شود.
- از ویژگی های متفاوت برای نمونه های مختلف استفاده می شود.
- احتیاجی به تخمین تابع توزیع نیست.
- آماده سازی داده ها برای یک درخت تصمیم، ساده یا غیرضروری است. (روش های دیگر اغلب نیاز به نرمال سازی داده یا حذف مقادیر خالی یا ایجاد متغیرهای پوچ دارند)
- درخت تصمیم یک مدل جعبهء سفید است. توصیف شرایط در درختان تصمیم به آسانی با منطق بولی امکان پذیر است در حالی که شبکه های عصبی بدلیل پیچیدگی در توصیف نتایج آنها یک جعبهء سیاه می باشند.
- تایید یک مدل در درخت های تصمیم با استفاده از تست های آماری امکان پذیر است. (قابلیت اطمینان مدل را می توان نشان داد)
- ساختارهای درخت تصمیم برای تحلیل داده های بزرگ در زمان کوتاه قدرتمند می باشند.
- روابط غیرمنتظره یا نامعلوم را می یابند.
- درخت های تصمیم قادر به شناسایی تفاوت های زیرگروه ها می باشند.
- درخت های تصمیم قادر به سازگار کردن داده های فاقد مقدار می باشند.

12-4معایب درختان تصمیم

1. در مواردی که هدف از یادگیری، تخمین تابعی با مقادیر پیوسته است مناسب نیستند.
2. در موارد با تعداد دسته های زیاد و نمونهء آموزشی کم، احتمال خطا بالاست.

3. تولید درخت تصمیم گیری، هزینه محاسباتی بالا دارد.
4. هرس کردن درخت هزینه بالایی دارد.
5. در مسائلی که دسته های ورودی با نواحی مکعبی به خوبی جدا نشوند و دسته ها همپوشانی داشته باشند، خوب عمل نمی کنند.
6. در صورت همپوشانی گره ها تعداد گره های پایانی زیاد می شود.
7. در صورتی که درخت بزرگ باشد امکان است خطاها از سطحی به سطحی دیگر جمع شوند (انباشته شدن خطای لایه ها بر روی یکدیگر).
8. طراحی درخت تصمیم گیری بهینه، دشوار است. کارایی یک درخت دسته بندی کننده به چگونگی طراحی خوب آن بستگی دارد.
9. احتمال تولید روابط نادرست وجود دارد.
10. بازنمایی درخت تصمیم دشوار است.

11. وقتی تعداد دسته ها زیاد است، می تواند باعث شود که تعداد گره های پایانی بیشتر از تعداد دسته های واقعی بوده و بنابراین زمان جستجو و فضای حافظه را افزایش می دهد.

13-4 انواع درختان تصمیم: [7]

1-13-4 درختان رگراسیون

وظیفهء یادگیری در درختان رگراسیون، شامل پیش بینی اعداد حقیقی بجای مقادیر دسته ای گسسته است. که این عمل را با داشتن مقادیر حقیقی در گره های برگ خود نشان می دهند. بدین صورت که میانگین مقادیر هدف نمونه های آموزشی را در این گرهء برگ بدست می آورند. این نوع از درختان، تفسیر آسان داشته و می توانند توابع ثابت تکه ای را تقریب بزنند.

نسخهء پیچیده تر درختان رگراسیون، درختان مدل هستند که عمل رگراسیون را با داشتن مدل خطی در گره های داخلی یا پایانی نشان می دهند (در هر گره، توابع رگراسیون خطی دارند). بعداز اینکه درخت رگراسیون کامل ساخته شد، عمل رگراسیون خطی، به نمونه هایی که به این گره رسیده اند اعمال می شود و فقط از یک زیرمجموعه از صفات (صفاتی که در زیردرخت دیده خواهند شد) برای این کار استفاده می شوند. بدلیل استفاده از زیرمجموعه ای از صفات در هر گره، سربار عمل رگراسیون خطی زیاد نخواهد شد.

2-13-4 الگوریتمid3

-این الگوریتم درختان تصمیم از بالا به پایین می سازد و با طرح این سوال که چه صفتی باید در ریشهء درخت آزمایش شود آغاز می کند. برای پاسخ به این سوال، با استفاده از یکی از انواع آزمایش های آماری برای تعیین مناسب ترین صفت برای دسته بندی مثال های آموزشی، تصمیم براساس هر صفت نمونه را ارزیابی می کند. سپس بهترین صفت را انتخاب کرده و به عنوان تست در گرهء ریشهء درخت استفاده می کند. برای هر مقدار ممکن صفت تست شده در ریشه، یک گرهء متناظر ایجاد شده و مثال های آموزشی براساس مقادیر صفت تست، بین این گره ها افراز می شوند. تمام فرایند ذکر شده، با استفاده از مثال های آموزشی نسبت داده شده به هر گره، برای انتخاب بهترین صفت برای آزمایشی در آن گرهء درخت تکرار می شود. این روش جستجویی حریصانه را برای یک درخت تصمیم قابل قبول ارائه می دهد که در این الگوریتم، هیچ گاه برای در نظر گرفتن دوبارهء انتخاب های قبلی، به عقب برگشت نمی شود. این الگوریتم در یادگیری نمونه هایی با صفات فاقد مقدار مشکل داشته و غیرافزایشی و ارزان می باشد

3-13-4 الگوریتم Id4-hat

تغییریافتهء الگوریتم ID4 است؛ به شکلی که اگر درخت موجود نتواند نمونهء جدید را به شکل صحیح دسته بندی کند این الگوریتم درخت را دوباره می سازد. اگر درخت نتواند دوباره ساخته شود بنابراین بهترین درخت نخواهد بود. در این حالت نتیجه با درخت نهایی تولید شده توسط ID3 متفاوت خواهد بود. هر دوی این الگوریتم ها وقتی بی نظمی صفر است یا تعداد خروجی های تصمیم را نگه می دارند و وقتی تمام ... بجز یکی صفر است آن را متوقف می کنند.

4-13-4 الگوریتمid5

یک الگوریتم افزایشی بهبود یافته است که توسط Utgoff توسعه یافته و همانند الگوریتم ID4 شروع می شود. وقتی یک نمونهء جدید اضافه می شود، اگر توسط درخت موجود به شکل صحیح دسته بندی نشود بهترین

صفت بعدی را با استفاده از نفع اطلاعات برای دسته بندی این مثال اضافه می کند (در غیر این صورت درخت موجود را حفظ می کند.). در هر مرحله، اگر برای تمام نمونه هایی که تا این مرحله دیده شده اند، صفت پایین تر به نسبت صفت بالاتر بی نظمی شرطی کوچکتری داشته باشد درخت را با انجام تقسیم، معکوس کردن، ادغام و ساده سازی دوباره می سازد.

5-13-4 الگوریتم id5-hat

در الگوریتم ID5 هرگاه که یک نمونهء آموزشی اضافه می شود، بی نظمی های شرطی دوباره کنترل می شوند (و در صورت لزوم ساختار درخت تغییر می یابد.). این الگوریتم مشابه الگوریتم ID5 می باشد جز اینکه بی نظمی های شرطی فقط زمانی دوباره درنظر گرفته می شوند که درخت قادر به دسته بندی صحیح یک نمونهء جدید نباشد

6-13-4الگوریتم c4.5

نسل بعدی الگوریتم ID3 است و از نوعی از قانون هرس بعدی استفاده می کند. همچنین قادر است صفات گسسته، صفات فاقد مقدار و داده های نویزی را استفاده کند. این الگوریتم بهترین صفت را با استفاده از معیار بی نظمی انتخاب می کند و به دلیل استفاده از عامل GainRatio قادر به بکارگیری صفات با مقادیر بسیار زیاد می باشد. حتی اگر هیچ خطایی در داده های آموزشی وجود نداشته باشد هرس انجام می شود که باعث می شود درخت عام تر شده و کمتر به مجموعهء آموزشی وابسته شود.

هرس در این الگوریتم نسبتاً پیچیده و برپایهء توزیع دوجمله ای و به شکل بازگشتی به برگ های درخت است. وقتی هرس یک شاخه متوقف می شود به سمت بالا ادامه نمی یابد. برای ممانعت از داشتن برگ هایی با یک نمونهء آزمایشی، جداسازی بیشتر روی دسته هایی که در حال حاضر به دو عنصر کاهش یافته اند انجام نمی شود. هرس فقط زمانی انجام می شود که تعداد پیش بینی شدهء خطاها افزایش نیابد. این الگوریتم، با در نظر گرفتن بی نظمی های هریک از آنها برای هر موردی که برای آنها داده داده شده است یک صفت را انتخاب می کند. بعد از انتخب بهترین صفت، موارد صفات فاقد مقدار با مقادیری از صفت در بخشی از مواردی که داده فراهم است تخصیص می یابند و الگوریتم ادامه می یابد.

7-13-4 الگوریتم cart

- نخستین بار توسط olshen, friedman, stone,bremiman در سال 1984 برای درختان رگرسیون و کلاسه بندی طراحی شد.

- روش عملکرد این الگوریتم surrogate splitting نام دارد. این الگوریتم شامل یک متد بازگشتی است. الگوریتم cart در هر مرحله رکورد های آموزشی را به دو زیر مجموعه تقسیم می کند.به طوریکه رکوردهای هر زیر مجموعه نسبت به زیر مجموعه های قبلی همگن تر باشد.این تقسیم شدنها به دفعات انجام می شودتا شرایط تو قف برقرار شود. در cart بهترین شکست با تعیین مقدار پارامتر impurity تعیین می شود. اگر بهترین شکست برای یک شاخه impurity را از حد تعریف شده کمتر کند آن انشعاب ساخته نمی شودمفهوم impurity در اینجا به میزان شباهت مقدار فیلد هدف قرار بگیرد ان گره pure نامیده می شود قابل توجه است که در الگوریتم cart یک فیلد پیشگو ممکن است به دفعات در سطوح مختلف درخت تصمیم گیری بکار گرفته شود. همچنین این الگوریتم فیلدهایهدف و پیشگوی از نوع categorical و continues را پشتیبانی می کند.

8-13-4 الگوریتم c5.0

- c 5.0 یکی از الگوریتم های درختان تصمیم گیری می باشدکه تحقیقاتمان به منظور کشف دانش و قوانین با کیفیت تر مورد استفاده قرار گرفت. الگوریتم c5.0 یک نوع درخت تصمیم گیری تک متغییره و بهبود یافته الگوریتم c4.5 است که تو سط محقق استرلیایی j.ross quinlan در سال 1993 طراحی شد. این الگوریتم مشابه با cart ابتدا درختی تقریبا پر ایجاد می کند ولی استراتژی هرس ان کاملا متفاوت است . این الگوریتم کلاسه بندی را با تقسیم کردن داده ها به زیر مجموعه هایی که شامل رکورد های همگن تر از والد خود هستند انجام می دهد . در c5.0 تقسیم کردن نمونه ها براساس فیلدی که بیشترین بهره اطلاعات را دارد صورت میگیرد.

4- 14 نرم افزارهای داده کاوی[8]

تا به امروز نرم افزار های تجاری و آموزشی فراوانی برای داده کاوی در حوزه های مختلف داده ها به دنیای علم و فناوری عرضه شده‌اند. هریک از آنها با توجه به نوع اصلی داده هایی که مورد کاوش قرار می‌دهند، روی الگوریتمهای خاصی متمرکز شده‌اند. مقایسه دقیق و علمی این ابزارها باید از جنبه های متفاوت و متعددی مانند تنوع انواع و فرمت داده های ورودی، حجم ممکن برای پردازش داده ها، الگوریتمها پیاده سازی شده، روشهای ارزیابی نتایج، روشهای مصور سازی ، روشهای پیش پردازش داده ها، واسطهای کاربر پسند ، پلت فرم های سازگار برای اجرا، قیمت و در دسترس بودن نرم افزار صورت گیرد.

4-14-1 نرم افزار WEKA

میزکار Weka ، مجموعه‌ای از الگوریتم‌های روز یادگیری ماشینی و ابزارهای پیش پردازش داده‌ها می‌باشد. این نرم‌افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش نمود. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرآیند داده کاوی های تجربی فراهم می‌کند. این پشتیبانی‌ها، آماده سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در بر می‌گیرند .همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش پردازش داده‌هاست. این جعبه ابزار متنوع و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی را که برای مسایل مدنظر مناسب‌تر هستند، تشخیص دهد .

نرم‌افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و اسم آن از عبارت "Waikato Environment for knowledge Analysis"استخراج گشته است .همچنین Weka ، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی‌کند و در نیوزلند، یافت می‌شود. این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر GNU انتشار یافته است Weka .تقریباً روی هر پلت فرمی اجرا می‌شود و نیز تحت سیستم عامل‌های لینوکس، ویندوز، و مکینتاش، و حتی روی یک منشی دیجیتالی شخصی ، آزمایش شده است .

این نرم‌افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش و ارزیابی نتایج طرح های یادگیری روی همه مجموعه های داده موجود، قابل اعمال است .

نرم افزار Weka ، پیاده سازی الگوریتم‌های مختلف یادگیری را فراهم می‌کند و به آسانی می‌توان آنها را به مجموعه های داده خود اعمال کرد .

همچنین، این نرم‌افزار شامل مجموعه متنوعی از ابزارهای تبدیل مجموعه های داده‌ها، همانند الگوریتم‌های گسسته سازی می‌باشد. در این محیط می‌توان یک مجموعه داده را پیش پردازش کرد، آن را به یک طرح یادگیری وارد نمود، و دسته‌بندی حاصله و کارآیی‌اش را مورد تحلیل قرار داد. (همه این کارها، بدون نیاز به نوشتن هیچ قطعه برنامه‌ای میسر است.)

این محیط، شامل روش‌هایی برای همه مسایل استاندارد داده کاوی مانند رگرسیون، رده‌بندی، خوشه‌بندی، کاوش قواعد انجمنی و انتخاب ویژگی می‌باشد. با در نظر گرفتن اینکه، داده‌ها بخش مکمل کار هستند، بسیاری از ابزارهای پیش پردازش داده‌ها و مصورسازی آنها فراهم گشته است. همه الگوریتم‌ها، ورودی‌های خود را به صورت یک جدول رابطه‌ای به فرمت ARFF دریافت می‌کنند. این فرمت داده‌ها، می‌تواند از یک فایل خوانده شده یا به وسیله یک درخواست از پایگاه داده‌ای تولید گردد .

یکی از راه‌های به کارگیری Weka ، اعمال یک روش یادگیری به یک مجموعه داده و تحلیل خروجی آن برای شناخت چیزهای بیشتری راجع به آن اطلاعات می‌باشد. راه دیگر استفاده از مدل یادگیری شده برای تولید پیش‌بینی‌هایی در مورد نمونه‌های جدید است. سومین راه، اعمال یادگیرنده‌های مختلف و مقایسه کارآیی آنها به منظور انتخاب یکی از آنها برای تخمین می‌باشد. روش‌های یادگیری Classifier نامیده می‌شوند و در واسط تعاملی Weka ، می‌توان هر یک از آنها را از منو انتخاب نمود. بسیاری از classifier ها پارامترهای قابل تنظیم دارند که می‌توان از طریق صفحه ویژگی‌ها یا object editor به آنها دسترسی داشت. یک واحد ارزیابی مشترک، برای اندازه‌گیری کارآیی همه classifier به کار می‌رود .

پیاده سازی‌های چارچوب‌های یادگیری واقعی، منابع بسیار ارزشمندی هستند که Weka فراهم می‌کند.

ابزارهایی که برای پیش پردازش داده‌ها استفاده می‌شوند filter .نامیده می‌شوند. همانند classifier ها، می-توان filter ها را از منوی مربوطه انتخاب کرده و آنها را با نیازمندی‌های خود، سازگار نمود.

علاوه بر موارد فوق، Wekaشامل پیاده سازی الگوریتم‌هایی برای یادگیری قواعد انجمنی، خوشه‌بندی داده-ها در جایی که هیچ دسته‌ای تعریف نشده است، و انتخاب ویژگی‌های مرتبط در داده‌ها می‌شود .

4-14-1-1قابلیت های weka

مستندسازی در لحظه، که به صورت خودکار از کد اصلی تولید می‌شود و دقیقاً ساختار آن را بیان می‌کند، قابلیت مهمی است که حین استفاده از Weka وجوددارد .

نحوه استفاده از این مستندات و چگونگی تعیین پایه‌های ساختمانی اصلی Weka ، مشخص کردن بخش-هایی که از روش‌های یادگیری با سرپرست استفاده می‌کند، ابزاری برای پیش پردازش داده‌ها بکار می‌رود. مستندات در لحظه همیشه به هنگام شده می‌باشد. اگر ادامه دادن به مراحل بعدی و دسترسی به کتابخانه از برنامه جاوا شخصی یا نوشتن و آزمایش کردن برنامه‌های یادگیری شخصی مورد نیاز باشد، این ویژگی بسیار حیاتی خواهد بود .

در اغلب برنامه‌های کاربردی داده کاوی، جزء یادگیری ماشینی، بخش کوچکی از سیستم نرم‌افزاری نسبتاً بزرگی را شامل می‌شود. در صورتی که نوشتن برنامه کاربردی داده کاوی مد نظر باشد، می‌توان با برنامه‌نویسی اندکی به برنامه‌های Weka از داخل کد شخصی دسترسی داشت. اگر پیدا کردن مهارت در الگوریتم‌های یادگیری ماشینی مدنظر باشد، اجرای الگوریتم‌های شخصی بدون درگیر جزییات دست و پا گیر شدن مثل خواندن اطلاعات از یک فایل، اجرای الگوریتم‌های فیلترینگ یا تهیه کد برای ارزیابی نتایج یکی از خواسته‌ها می‌باشد Weka .دارای همه این مزیت‌ها است.

4-14-2 نرم افزار JMP[2]

jmp،یک بسته نرم افزاری بسیار قدرتمند است که به کاربران اجازه می دهد مدل های بسیار عالی با یک قیمت نسبتا ارزان ایجاد کنند.این نرم افزار برای شرکت هایی مهم است که می خواهند قدرت تجزیه وتحلیل کسب وکار خود را بدست آورند.اما نمی خواهند بودجه بیش از حد برروی آن صرف کنند.

4-14-2-1قابلیت های JMP

- تجزیه وتحلیل داده های اکتشافی
- تجزیه وتحلیل خوشه ای
- تجزیه وتحلیل مولفه های اصلی
- مدل های رگرسیون:منطق وکمترین نظم
- شبکه های عصبی ودرخت تصمیم

تجزیه وتحلیل داده های اکتشافی در JMP به یک شیوه گرافیکی انجام می شود. میزان اطلاعات بسیار غنی وگرافیک سطح بالامی باشد.همچنین درآن لیستی از انتخاب آنالیزهاوجود دارد که ممکن است یک طراح آرزوی استفاده از آن را دادشته باشد.با استفاده از بخش تجزیه و تحلیل داده اکتشافی ما قادر خواهیم بود که یک درک از توزیع ارزش داده ها داشته باشیم.

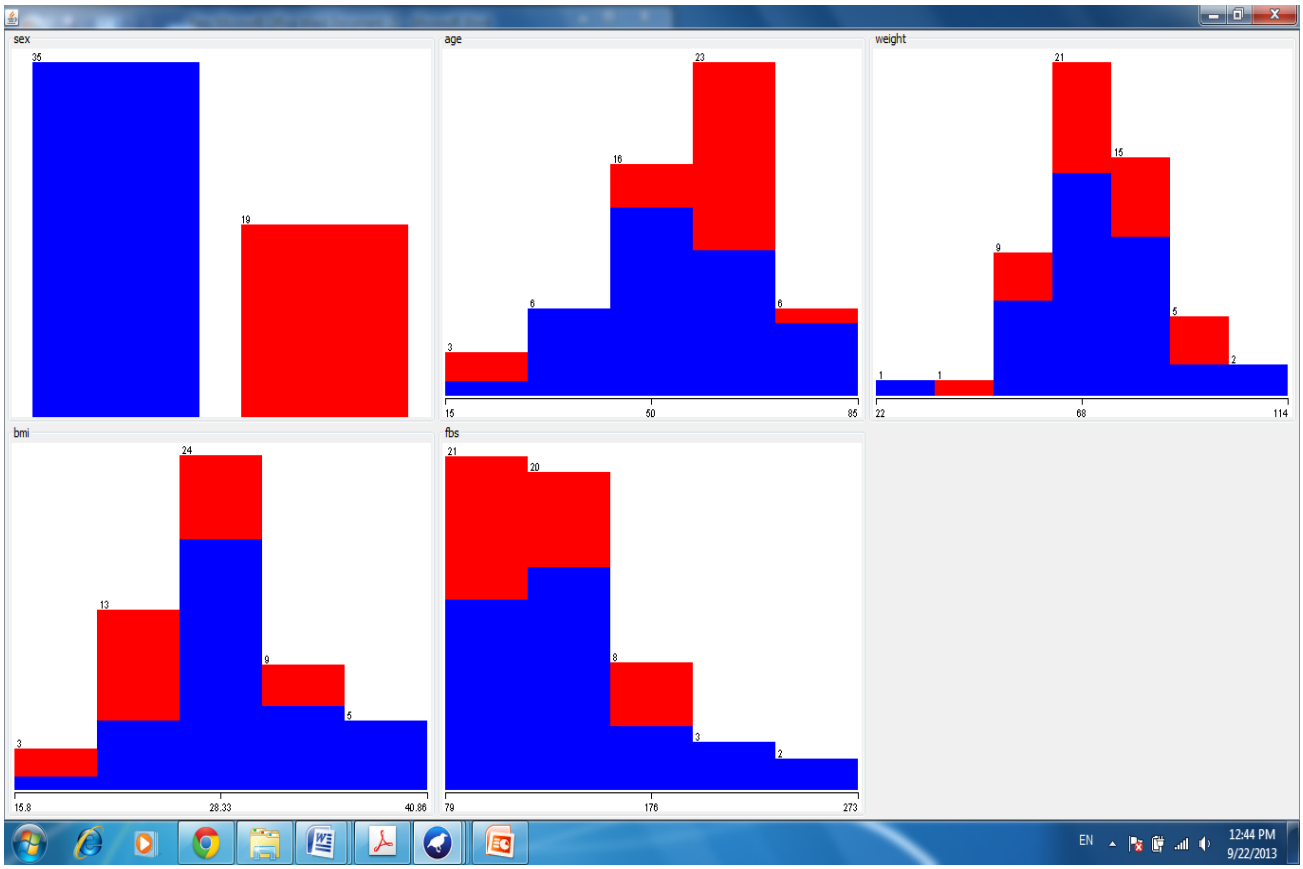
4-15پیاده سازی نرم افزار وکا

در این قسمت با استفاده از نرم افزار وکا واطلاعات 54 بیمار دیابتی که شامل 5خوصیت sex,age,weight,bmiوfbsمی باشد شیوع این بیماری در دو جنس مونث ومذکر مورد تحلیل ومقایسه قرار گرفته است.

4-15-1پیاده سازی توسط الگوریتم Naïve Bayse:

بعد از باز کردن نرم افزار وباز کردن پنجرهExplorerفایل اطلاعات را به نرم افزار میدهیم و در قسمت Arttibuteگزینه allرا به جهت انتخاب تمام فیلدها انتخاب می کنیم

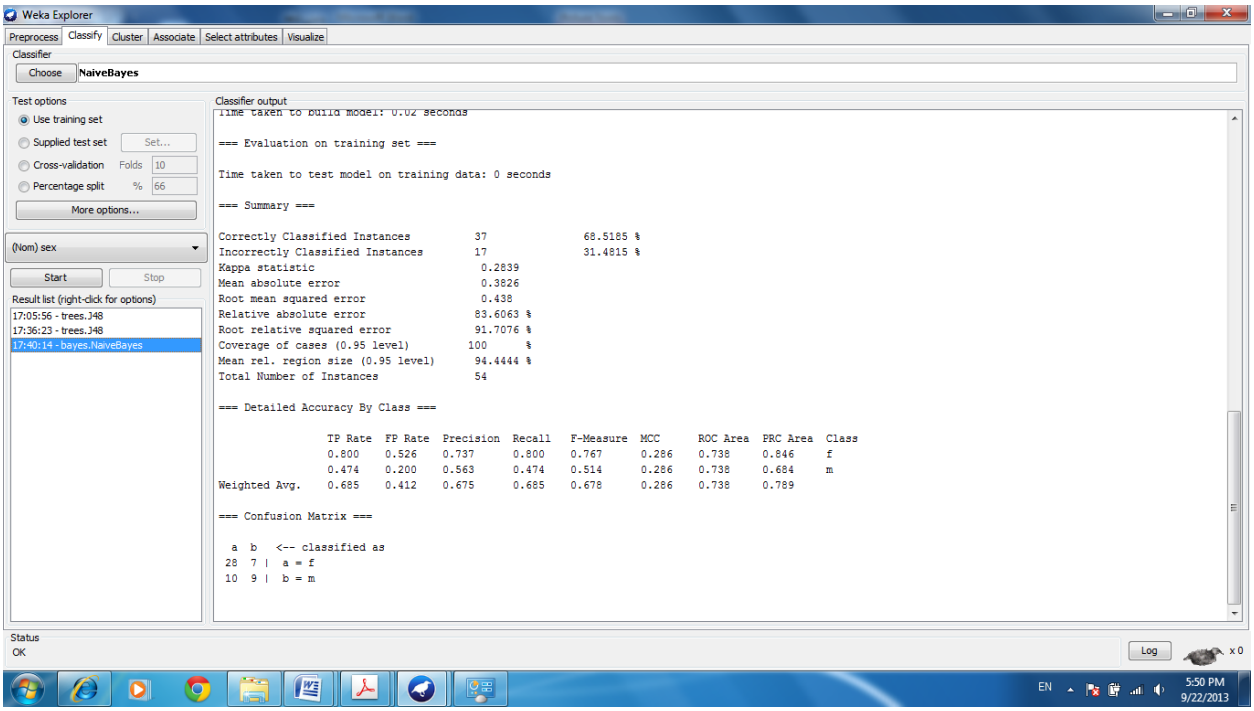
با انتخاب گزینه Visualize All نمودار ستونی برای فراوانی مقادیرمختلف نمایش داده می شود(شکل4-3)



شکل 4-3 نمودار ستونی برای فراوانی مقادیرمختلف ستون ها در بازه هایی با طول یکسان

سپس باانتخاب تب Classifyبه صحفه جدید وارد می شویم در این صفحه نوع الگوریتم را از قسمت Classifierکه Naïve Bayseمی باشدرا انتخاب و روی گزینه Startکلیک می کنیم.

در نهایت خواهیم داشت (شکل4-4)

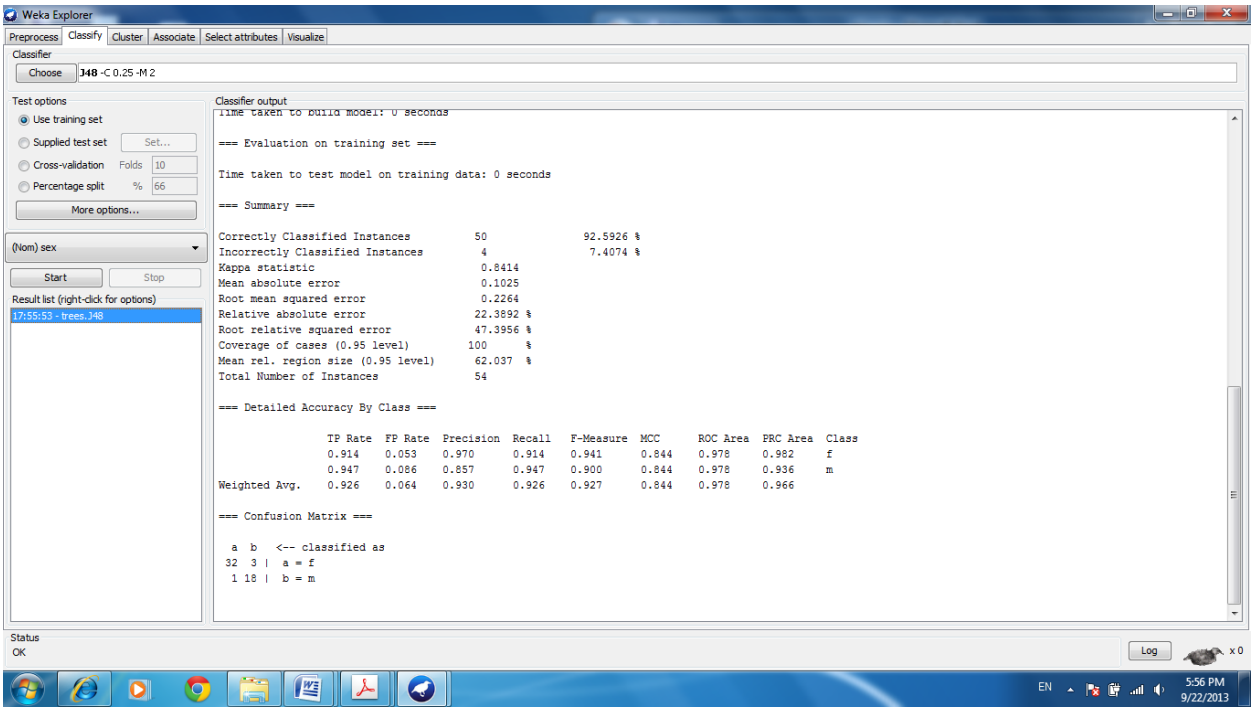


شکل 4-4 اجرای الگوریتمNaive Bayse

همانطور که در قسمت Precisionدر شکل مشاهده می کنید زنان با 0.737درصدومردان با 0.563 درصد محاسبه شده اند.

4-15-2پیاده سازی توسط الگوریتم Decision Trees:

در این قسمت ازمیان الگوریتم ها ی Treesالگوریتم J48 را انتخاب می کنیم که نتایج زیر را به دنبال دارد.(شکل4-5)



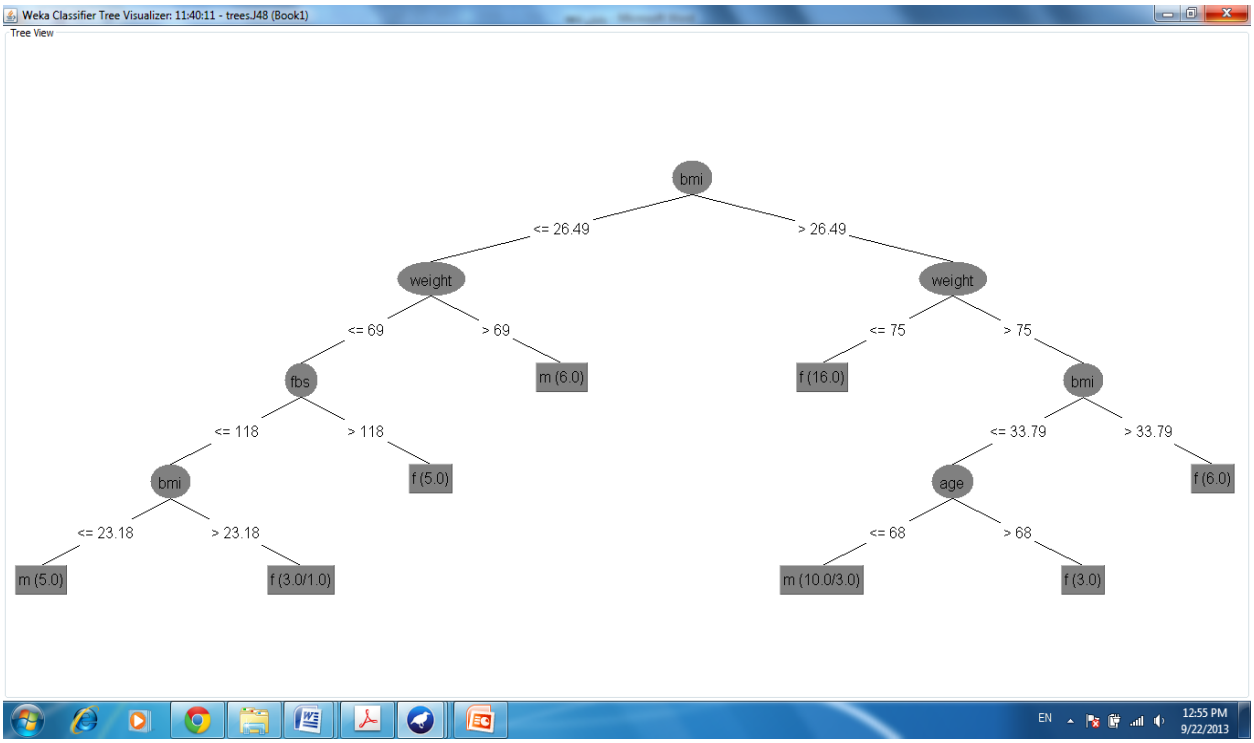
شکل 4-5 اجرای الگوریتمDecision Trees

طبق شکل زنان با 0.970 درصد ومردان با 0.857 درصد محاسبه شده است.

از جمله اعداد مهم که در خروجی نمایش داده شده Correctly Classified Instances است به میزان 92.5926که نشان می دهد بیش از 92 درصد از نمونه هایمان به درستی دسته بندی شده اند. در مقابل این معیار نیز Incorectly Classified Instances قرار دارد. اما در الگوریتم Naïve Bayesحدود 68 درصد از داده ها به درستی دسته بندی شده است در نتیجه الگوریتم J48 نتیجه ی بهتری را در اختیارمان قرار می دهد.

در قسمت Confusion Matrixمیزان مثبت کاذب ومنفی کاذب را مشخص می کند. در اینجا مثبت کاذب 3ومنفی کاذب 1 می باشد. مثبت کاذب زمانی رخ می دهد که آزمایشات آماری فرض صحیح را رد می کند. منفی کاذب نیز هنگامی رخ می دهد که آزمایش آماری فرضی نادرست را بپذیرد.

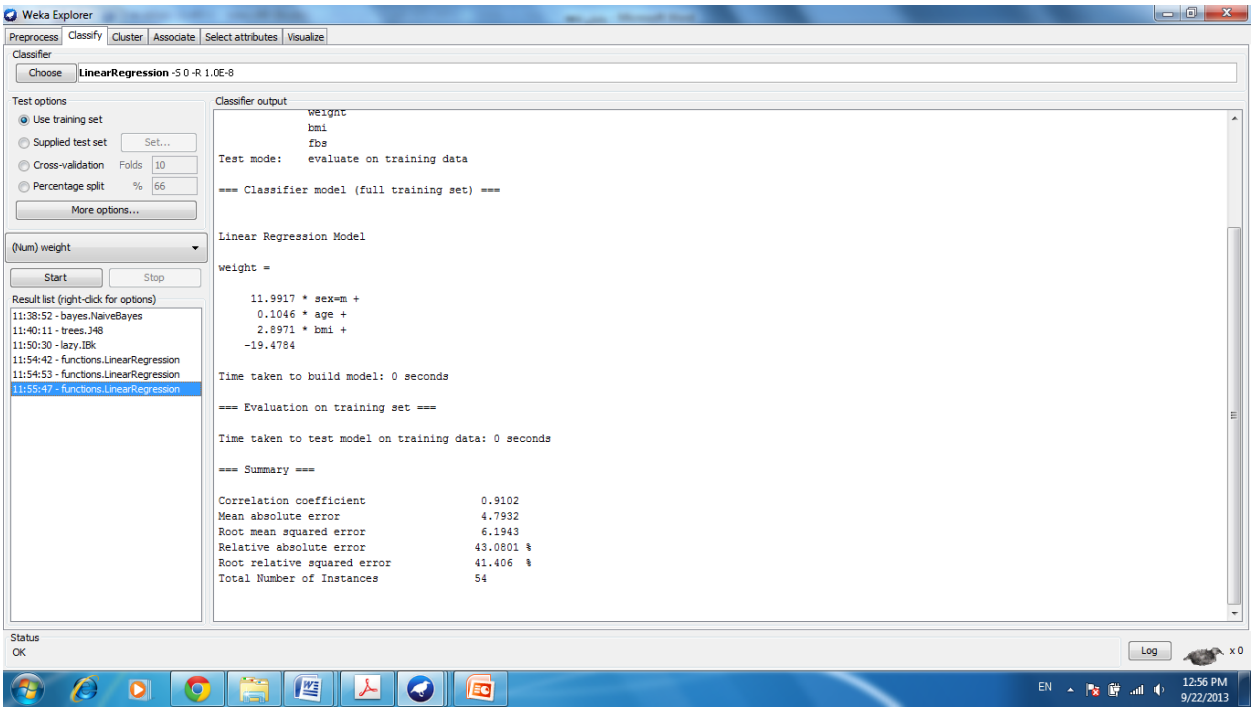
برای مشاهده درخت تصمیم در پنجره ی Result Listراست کلیک می کنیم وگزینه Visualize treeرا انتخاب می کنیم. (شکل4-6)



شکل 4-6 درخت تصمیم

4-15-3 ایجاد مدل رگرسیون

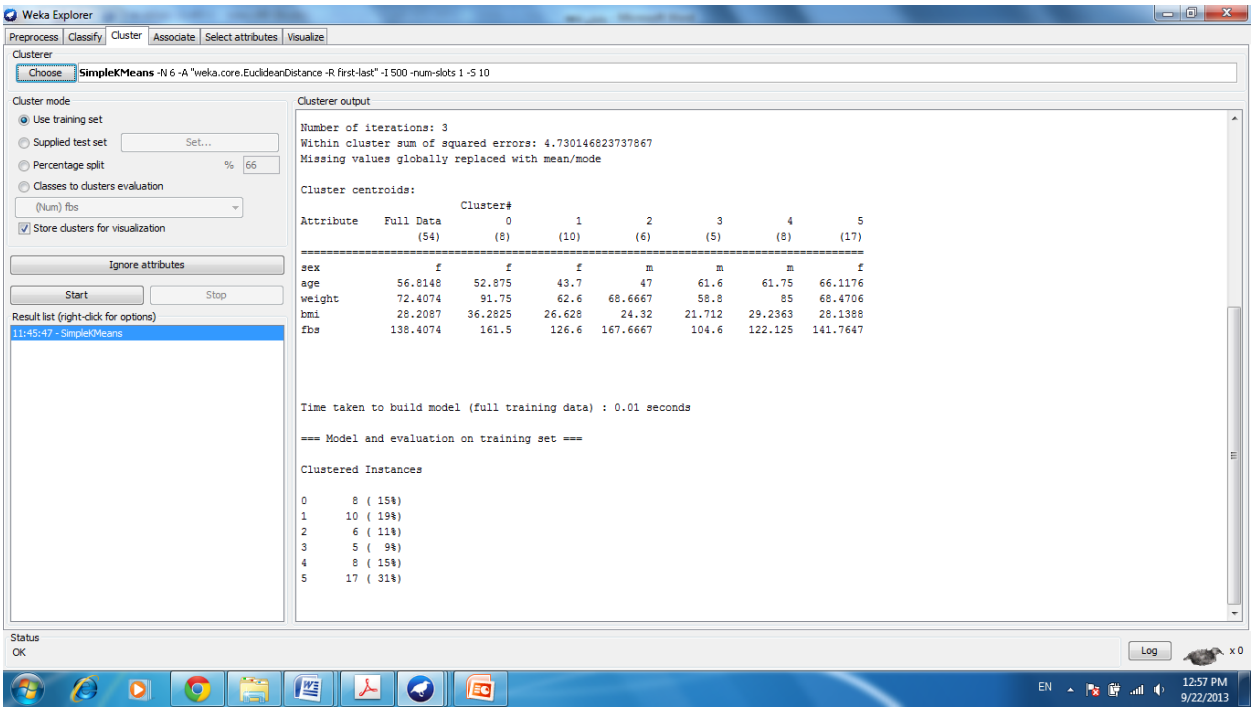
با انتخاب تب Classifyوانتخاب Linear Regression نتیجه ی زیر حاصل می شود.(شکل4-7)



شکل 4-7 اجرای مدل رگرسیون

4-15- 4 ایجاد مدل با خوشه بندی

با انتخاب تب Clusterوارد بخش خوشه بندی می شویم .در این قسمت از میان روش های خوشه بندی SimpleKMeansرا انتخاب می کنیم وتعداد خوشه ها را برابر 6 قرار می دهیم تا بتوانیم تغییر ماهیت خوشه ها را به خوبی مشاهده کنیم.(شکل4-8)



شکل 4-8 اجرای مدل خوشه بندی

توضیح روش کلی K_means:

این روش ابتدا kخوشه را مشخص کرده ومراکز آنها را تعیین می کند.حال الگوریتم به اجرای متناوب این سه فاز می پردازد:

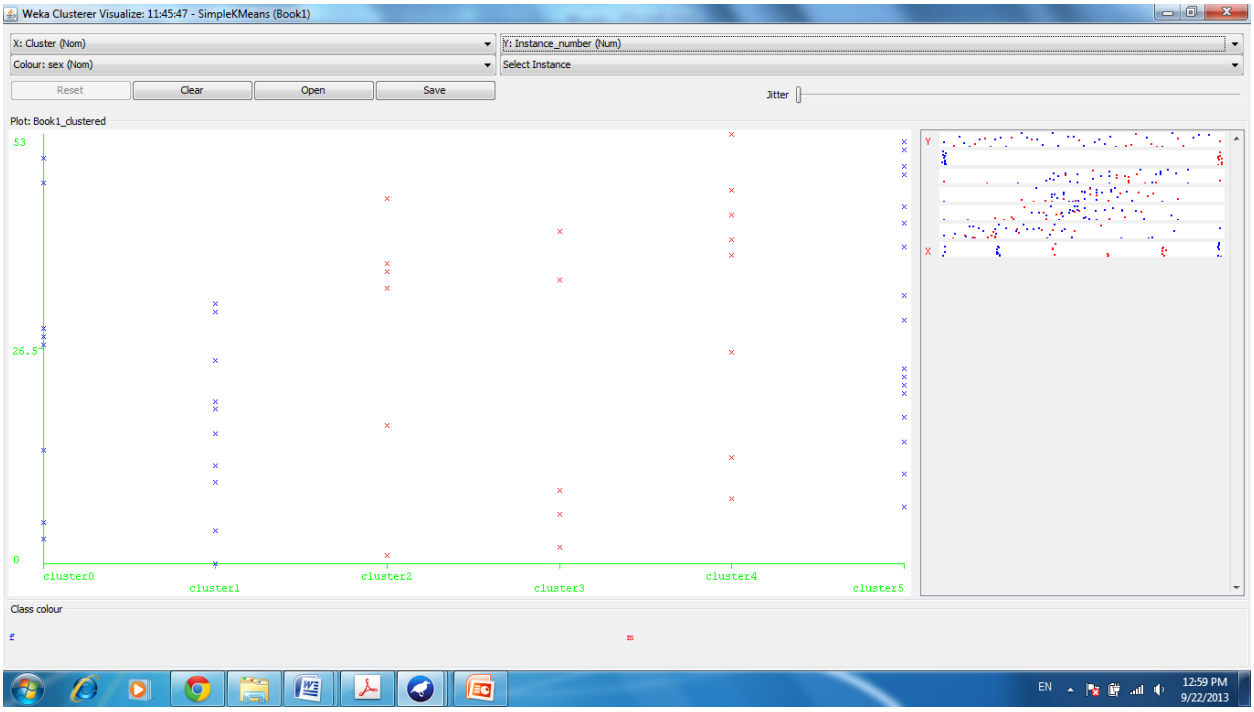
1-مختصات مرکز را مشخص کن.

2-فاصله هر نقطه تا مرکز را مشخص کن.

3-هرداده را در خوشه ای که فاصله مرکز آن تا داده کمترین است قرار بده .این کار تا جایی ادامه پیدا می کند که دیگر نقاط داده میان خوشه ها حرکت نکنند.نتیجه نهایی خوشه بندی این است که در آن فاصله هر نقطه نسبت به مرکز خوشه ای که درآن قرار دارد کمتر از فاصله اش نسبت به مراکز خوشه های دیگر است.

مطابق (شکل) نه تنها مراکز هر خوشه مشخص شده بلکه میزان درصد داده هایی که به هر خوشه نسبت داده شده اند،نیز ذکر شده اند.به عنوان مثال خوشه 5را در نظر بگیرید.نتایج نشان می دهد نزدیک به31درصد از کل مجموعه در این خوشه قرار گرفته اند.در ضمن از اطلاعات داده شده درباره مرکز این خوشه می توان برداشت کرد که عموم افراد حاضر در این خوشه،زنانی حدودا 66ساله وبا وزن 68 کیلو باbmi28وباقند خون 141.

برای نمایش بصری داده ها روی خوشه بندی انجام شده راست کلیک کرده Visualize cluster assignmentsرا انتخاب می کنیم.به عنوان مثال می توان محور xرا برای شماره خوشه ومحور yروی تعداد داده هایی که در آن خوشه قرار گرفته اند،رنگ را نیز با جنسیت تنظیم کنیم.(شکل4-9)



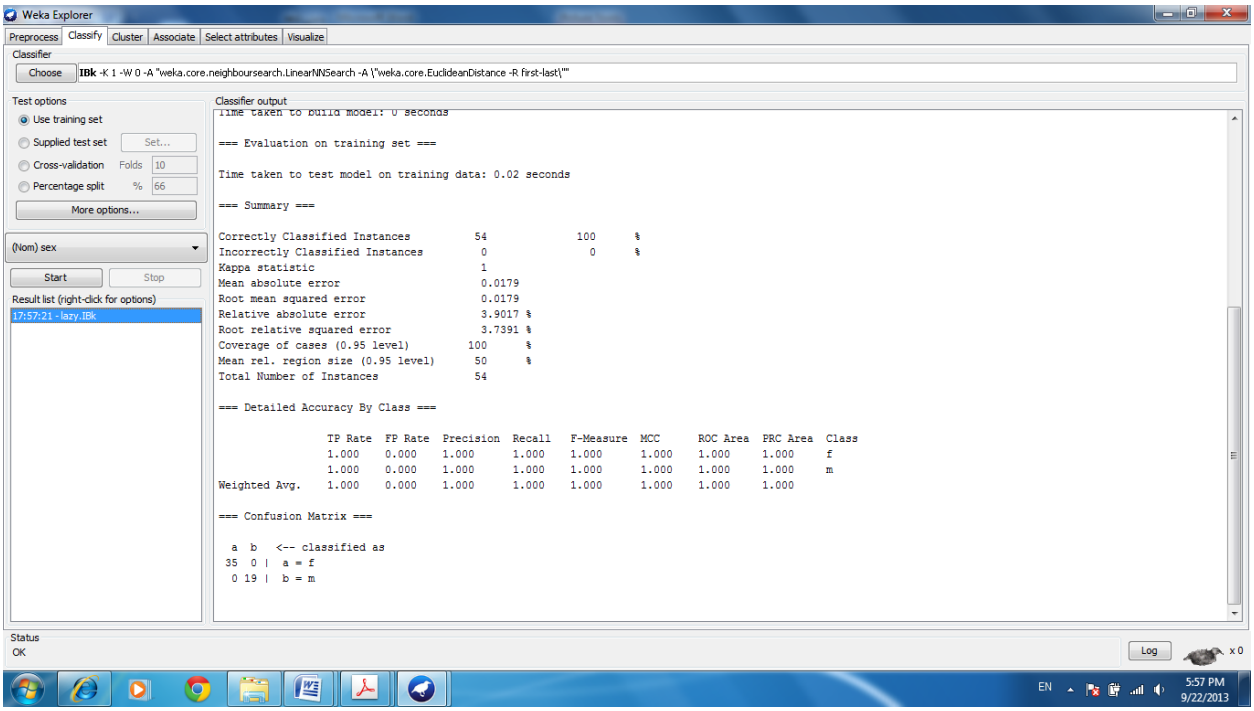
شکل 4-9 نمایش داده ها در خوشه بندی

همانطور که مشاهده می شود تعداد زنان در خوشه های ابتدا وانتهایی بیشتر ودر خوشه های میانی تعداد مردان بیشتر است.

4-15-5پیاده سازی با الگوریتم نزدیک ترین همسایه:

در این روش سعی می شود تا ویژگی های نقاط داده از روی ویژگی های نزدیک ترین همسایگانشان تعیین شود واز نوعی رای گیری استفاده شود وتنها به نزدیک ترین همسایه صرف اعتماد نمی شود.از آنجا که مادر این روش در حال انجام نوعی پیش بینی هستیم، کارمان تاحدی شبیه همان کاری است که در رگرسیون وطبقه بندی انجام می دادیم.از طرف دیگر مفهوم اولیه نزدیک ترین همسایه در واقع همان ایده خوشه بندی است.

در قسمت Classifyاززیر منوی Lazyگزینه Ibkرا انتخاب می کنیم وگزینه startرا کلیک می کنیم.(شکل4-10)

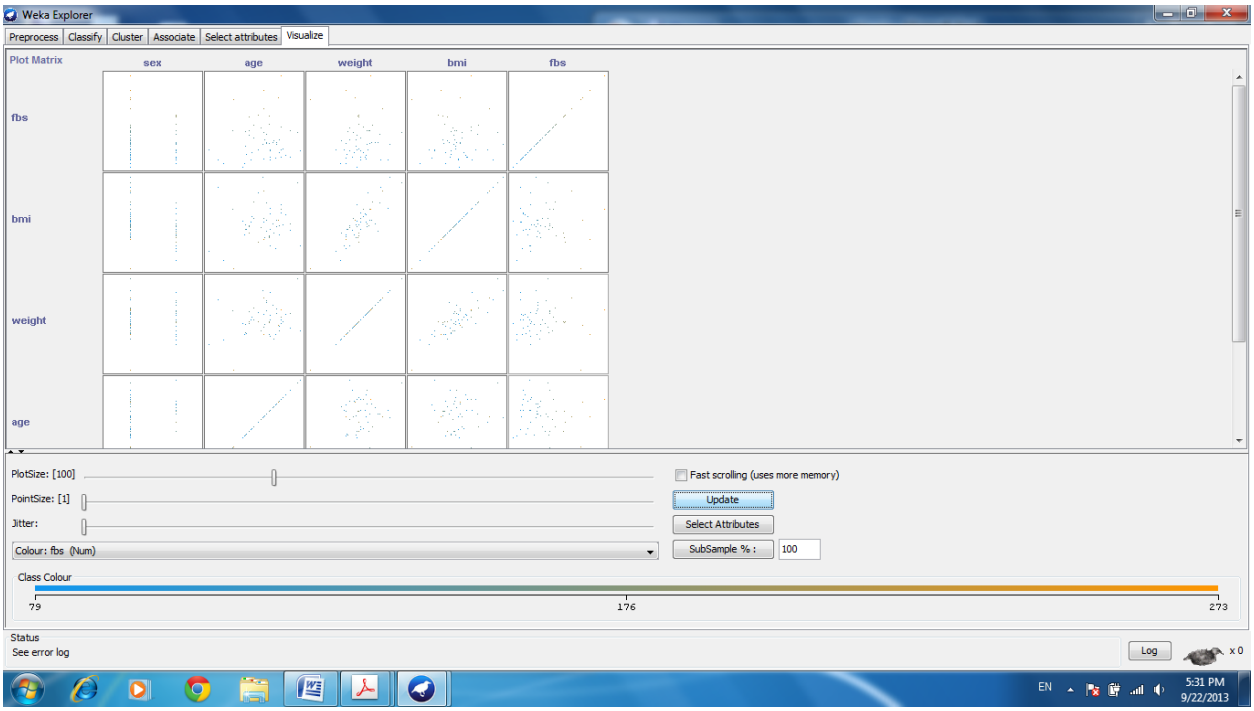


شکل4-10 اجرای الگوریتم نزدیک ترین همسایه

همانطور که مشاهده می کنید فرمت نتیجه نهایی نزدیک ترین همسایه بسیار شبیه به فرمت نهایی روش طبقه بندی است.در این روش 100درصد از دادها به درستی دسته بندی شده اند ومثبت کاذب ومنفی کاذب نیز 0است.اما در روش طبقه بندی به 92 درصد دقت دست یافتیم.نکته ای که در اینجا باید به آن توجه داشته باشید این است که این ترتیب همیشه به این گونه حفظ نخواهد شد.

6-15-4 برگه Visualize

در این برگه نحوه توزیع داده های خصیصه های متفاوت براساس متغیر کلاس نشان داده شده است.(شکل4-11)



شکل 4-11 نحوه توزیع داده ها بر اساس متغیر کلاس

فصل پنجم

بحث ونتیجه گیری

5-1 بحث

یکی از مهمترین موضوعات چالش برانگیز در مراقبت سلامت، تغییر شکل داده های بالینی خام به اطلاعات معنی داربه دنبال تولید مستمر انبوهی از داده ها است چرا که بسیاری از سازمان های مراقبت سلامت با وجود غنای داده با فقر دانش روبرو شده اند.به عبارت دیگر رویارویی با مجموعه های عظیم داده ها وتوسعه ی پایگاه های داده ها نسبت به دهه های گذشته نیازهای جدیدی را مانند خلاصه سازی خودکار داده ها،استخراج اطلاعات ذخیره شده وکشف الگوها از داده های خام به وجود آورده است که داده کاوی نمونه ای از آن می باشد.

استخراج اطلاعات ودانش از داده ها مفهومی دیرینه در مطالعات علمی وپزشکی می باشد و آنچه که جدید است هم گرایی واشتراک چندین رشته وفن آوری های متناظر آنها است که فرصت منحصر به فردی را برای داده کاوی ایجاد کرده است.

داده کاوی با ایجاد پزشکی مبتنی بر شواهد نقش حیاتی در سلامت داردرومنجر به کشف دانش جدید،سودمندوماندگار در پایگاه های داده ای سازمان های سلامت می شود؛چرا که برای دستیابی به پزشکی مبتنی بر شواهد بایدازشناسایی شکاف و خلاء دانش در فرایندهای مراقبت سلامت کنونی شروع کرد و سپس به دنبال بهترین ادله بود، در قدم بعدی باید به بررسی صحیح بودن و معتبر بودن اقدامات شناسایی شده در بهترین ادله پرداخت و در قدم آخر این ادله را برروی بیماران اجرا کرد. داده کاوی دستیابی به اولین گام در این زمینه را هموار می سازد.

در محیط رقابتی امروز،سازمان های سلامتی که به واسطه ی فن آوری هایی هم چون داده کاوی بتوانند داده ها را در راستای بهبود کیفیت سلامت به کاربرند سریع تر به قله ی موفقیت خواهند رسید،لذا لازم است سازمان های سلامت ما نیز با استفاده از تخصص صاحبان این فن از این عرصه ی رقابت باز نمانند.

بسیاری از مراکز تحقیقاتی کشورمان دارای حجم زیادی از داده ها هستند که یا هرگز تحلیل نمی شوند ویا اگر هم تحلیل و به دانش منتج شوند به واسطه ی استفاده از شیوه های سنتی،امری مقطعی وزمان بر است؛حال آنکه با روی آوردن به داده کاوی واجرای آن می توانند داده ها را به ابزاری نیرومند و رقابتی تبدیل نموده و گام های جدیدی را در پیش گیری،تشخیص،درمان وارایه ی خدمات با کیفیت به مشتریان سلامت بردارند.

5-2 نتیجه گیری

در این مقاله مقوله داده کاوی وتکنیک های آن مورد بررسی قرار گرفت.همچنین از میان روش های دسته بندی وپیش بینی ، درخت تصمیم به عنوان یکی از ابزار های قوی و متداول در داده کاوی که درک،پیاده سازی وکاربرد آسان داشته واز نظر محاسباتی ارزان می باشد مورد بحث قرار گرفت .

با توجه به اهمیت وحساسیت داده کاوی در پزشکی وهمچنین نیاز مبرم این صنعت به حرکت از پزشکی سنتی به سمت پزشکی مبتنی برشواهد لذا در این مقاله کاربرد داده کاوی در عرصه سلامت مورد بررسی قرار گرفت .

نتایج به دست آمده روی مجموعه بیماران دیابتی نشان می دهد که در میان روش های طبقه بندی ،درخت تصمیم نتیجه ی بهتری را به دست می آورد والگوریتم نزدیک ترین همسایه نیز دقت بالاتری نسبت به روش طبقه بندی دارد اما با توجه به مطالعه سایر مقالات در زمینه تشخیص دیابت نتایج حاکی از آن است که هرگز نمی توان الگوریتمی را به عنوان الگوریتم بهینه معرفی کرد.در نتیجه برای هر کاربرد با توجه به مجموعه داده مورد استفاده می توان الگوریتمی را به عنوان الگوریتم بهینه معرفی نمود.

5-3پیشنهادات

- مقایسه اثربخشی مراقبت ودرمان قبل از اجرای داده کاوی وبعداز اجرای آن در مراکز سلامت
- مقایسه نتایج نرم افزارهای مختلف داده کاوی
- پیشنهاد درمانی برای بیماران با استفاده از داده کاوی

منابع

[1]اشکاوند راد،ایمان،داده کاوی،

[2]نوروزی،فردین- طائفی همراه،عارف،مروری بر داده کاوی وبررسی شبکه های عصبی

[3]کیخا، مصطفی- عباسی،علی،مقدمه ای بر داده کاوی

[4]مقدسی،حمید-حسینی،اعظم السادات-اسدی،فرخنده-جهانبخش،مریم،(1390)داده کاوی وکاربرد آن در سلامت

[5]صادق زاده ،مهدی - عشیر،امین-فروتن،فراز-سیفی پور،مرضیه،(1390)مقایسه الگوریتم های هوشمند در شناسایی بیماری دیابت

[6]عزیزی،(1385) یادگیری درخت های تصمیم

[7] http://www.iran-eng.com/

[8] http://www.eca.ir/forum2/index.php?topic=27035.0