

استفاده از داده کاوی در پیش بینی و دسته بندی وضعیت ترافیک

فرید سیفی^۱، محمد رضا کنگاوری^۲

۱- دانشجوی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران

۲- استادیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران

چکیده

داده کاوی یکی از شاخه های علم کامپیوتر است که در سالهای اخیر بطور وسیعی توسط سازمانهای بزرگ به منظور تحلیل و پردازش داده ها مورد توجه قرار گرفته است. تکنیکهای داده کاوی عموما برای استخراج داش و الگو از مجموعه داده های بزرگ و به منظور یادگیری و پیش بینی شرایط جدید به کار می روند. این تکنیکها در زمینه هایی که داده های زیادی تولید می کنند و پردازش این داده ها به صورت دستی ممکن نمی باشد مورد استفاده قرار می گیرند. با پیشرفت تکنولوژی امروز داده های بسیاری در زمینه ترافیک تولید می شوند که امکان پردازش دستی آنها وجود ندارد. یکی از مسائلی که در این زمینه طرح می شود پیش بینی وضعیت ترافیک است. در این مقاله از روش دسته بندی که یکی از تکنیکهای داده کاوی است، به منظور یادگیری رفتار ترافیک و پیش بینی وضعیت های جدید و تصمیم گیری در مورد آن استفاده کرده ایم و نشان داده ایم که این تکنیکها را چگونه می توان در این زمینه استفاده کرد.

کلید واژه: داده کاوی، دسته بندی، پیش بینی، مهندسی ترافیک، حمل و نقل.

۱- مقدمه

با پیشرفت و توسعه دانش کامپیوتر در سالهای اخیر و بکارگیری تکنولوژیهای سخت افزاری پیشرفت در نمونه برداری داده ها، حجم داده هایی که در حل مسائل مختلف جمع آوری می شوند و باید مورد بررسی قرار گیرند به شدت افزایش یافته است. با وجود اینکه افزایش حجم نمونه داده های گرفته شده از محیط می تواند به ایجاد راه حلی قطعی تر منجر شود ولی در اکثر موارد به دلیل دستی و کند بودن روشهای تحلیل و پردازش داده ها استفاده از این حجم بالای داده ها غیر ممکن است. متناسب با پیشرفت ابزار های نمونه گیری داده ها که حجم زیادی از داده های خام را در اختیار می گذارند، الگوریتمها و روشهای نرم افزاری جدیدی ایجاد شده اند که بررسی و تحلیل این داده ها را ممکن ساخته اند. داده کاوی طیف وسیعی از این روشها و الگوریتمهای مختلف را ارائه می کند که

^۱ دانشجو، ۰۹۱۵۵۰۲۳۹۰۴، farid@comp.iust.ac.ir

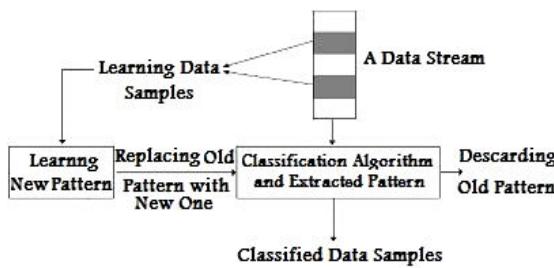
^۲ استادیار، ۰۹۱۲۳۵۰۰۹۸۳، kangavari@iust.ac.ir

می توانند در استخراج مفاهیم از داده ها، دسته بندی و خوش بندی داده ها، پیش بینی وضعیت های جدید و تحلیل و تصمیم گیری بکار روند.

استفاده از تجهیزاتی مثل سنسورها، شبکه های کابلی و بی سیم، فیبرهای نوری، دوربینها و کامپیوتروهای پر سرعت در سیستمهای ترافیکی امروز، بستر مناسبی برای استفاده از روشهای داده کاوی را فراهم کرده است. استفاده از این تجهیزات به جمع آوری اتوماتیک و سریع داده های ترافیکی منجر شده اند. به عنوان یکی از مهمترین استفاده های این داده های ترافیکی می توان به پیش بینی وضعیت ترافیک در آینده و تصمیم گیری در مورد آن اشاره کرد. حجم بالای داده های جمع آوری شده و از طرفی رسیدن بی درنگ داده های جدید امکان تحلیل دستی آنها را غیر ممکن می سازد. اینجاست که تکنیکهای داده کاوی اهمیت می یابند و راه حلهایی برای این مساله ارائه می کنند.

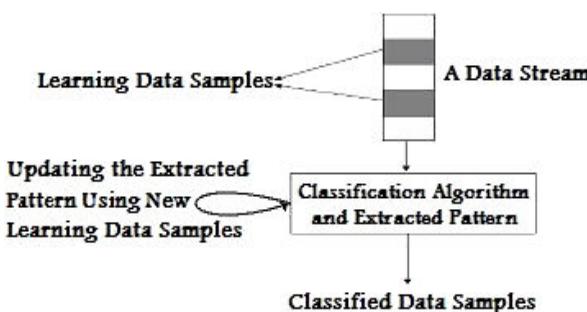
۲- داده کاوی و دسته بندی

به عنوان یک تعریف مختصر می توان گفت داده کاوی تکنیکهایی برای تحلیل و درک داده های رقمی ارائه می کند. اکثر تکنیکهای داده کاوی دارای دو قسمت می باشند. یکی، الگوریتم یادگیری است که توسط آن و با استفاده از مجموعه داده های آموزشی الگوی موجود در داده ها توسط الگوریتم به عنوان دانش استخراج می شود و در قسمت بعدی که قسمت تست نامیده می شود بکار می رود. داده ها ممکن است همگی در دسترس باشند و تعداد آنها ثابت باشد، که در اینصورت به الگوریتمهای مورد استفاده الگوریتمهای داده کاوی بروی داده های ایستا می گوییم. در این صورت الگو و مفهوم داده ها ثابت است و کافیست یکبار مرحله یادگیری الگوی داده ها انجام شود و در ادامه ای استفاده از الگوریتم الگوی یادگرفته شده بکار می رود. از طرفی ممکن است همه داده ها در ابتدا در اختیار الگوریتم قرار نداشته باشند بلکه با گذشت زمان داده های جدیدی در اختیار سیستم قرار گیرند. به الگوریتم هایی که بروی چنین داده هایی کار می کنند الگوریتمهای داده کاوی بروی داده های جریانی اطلاق می شود. در این حالت احتمال تغییر الگوی داده ها در طول زمان وجود دارد و بنابر این انجام یک مرحله یادگیری در ابتدا و استفاده از دانش استخراج شده برای تمام داده های جدیدی که وارد سیستم می شوند کاری صحیح نیست بلکه در این قبیل مسائل اغلب یادگیری را در بازه های زمانی مشخص و با استفاده از داده های آموزشی جدید تجدید می کنند، یعنی دانش استخراج شده قبلی را دور ریخته و دانش جدید را از داده های آموزشی جدید استخراج می کنند این مدل را در شکل ۱ مشاهده می کنید.



شکل ۱: مدل الگوریتم دسته‌بندی داده‌های جریانی که مدل‌های قدیمی را در صورت تغییر الگوی داده‌ها با مدل‌های جدید جایگزین می‌کند.

در راهکار دیگری برای داده‌های جریانی داده‌های آموزشی جدیدی که در دسترس سیستم قرار می‌گیرند به منظور بروز رسانی دانش اولیه بکار می‌روند و بدین ترتیب دانش موجود در سیستم با تغییر الگوی داده‌ها به مرور زمان سازگار می‌شود. این مدل را در شکل ۲ مشاهده می‌کنید.



شکل ۲: الگوریتم دسته‌بندی که مدل‌های قبلی را به مرور زمان بروزرسانی می‌کند.

مساله ترافیک و داده‌هایی که در این مساله وجود دارند دارای ماهیت جریانی هستند و به مرور زمان جمع آوری می‌شوند و الگوی آنها ممکن است به دلیل بروز مسائیل مثل تغییر فصلها، سهمیه بندی بنزین، تعطیلات رسمی و غیره تغییر کند. بنابر این برای انجام داده کاوی بروی داده‌های ترافیکی باید از روش‌های داده کاوی داده‌های جریانی استفاده کنیم.

داده کاوی، زمینه وسیعی برای تحقیق می‌باشد و شاخه‌های بسیاری مثل دسته‌بندی، خوش بندی، کاوش قوانین انجمنی و غیره دارد. در اینجا می‌خواهیم از داده کاوی برای یادگیری تاثیر پارامترهایی خاص بروی وضعیت ترافیک استفاده کنیم و از دانش استخراج شده از این یادگیری برای پیش‌بینی وضعیت ترافیک و تصمیم‌گیری در مورد آن استفاده کنیم.

دسته‌بندی (Classification) یکی از شاخه‌های داده کاوی است که می‌تواند به منظور یادگیری استنتاجی، استخراج دانش، پیش‌بینی و تصمیم‌گیری بکار رود. در مساله دسته‌بندی یک مجموعه داده آموزشی تعریف می‌شود که به منظور ایجاد یک دسته‌بند (Classifier) بکار می‌رود. هر نمونه داده در مجموعه داده‌های آموزشی تعدادی متغیر و یک شماره کلاس دارد که مشخص می‌کند آن نمونه داده به کدام کلاس تعلق دارد. پس از آموزش و تست الگوریتم دسته‌بند از آن به منظور پیش‌بینی کلاس داده‌های جدید استفاده می‌شود [۱]. در مساله ترافیک مورد نظر نیز داده‌های موجود دارای

تعدادی متغیر هستند که شامل ساعت نمونه برداری، دمای هوای روز هفته و متغیرهای مشابه هستند و کلاسی که برای هر نمونه داده تعریف می‌شود میزان ترافیک در آن شرایط است. هدف نیز یادگیری ارتباط بین مقادیر متغیرهای نمونه داده‌ها و کلاسها یا میزان ترافیک است تا برای داده‌های جدید که وضعیت ترافیکشان مشخص نیست پیش‌بینی صورت گیرد. این پیش‌بینی می‌تواند همان تصمیم‌گیری نیز باشد بدین معنی که برای هر وضعیت ترافیکی یک تصمیم‌گیری یا راه حل پیش‌فرض تعريف می‌شود. پس از دسته‌بندی و تعیین کلاس یک نمونه داده جدید، با توجه به اینکه کلاس آن چه چیزی پیش‌بینی شده است تصمیم متناظر با آن کلاس را در مورد آن شرایط اتخاذ می‌کنیم. روش‌های دسته‌بندی بسیاری تاکنون ابداع شده‌اند که بیشتر آنها بر مفهوم درختهای تصمیم تکیه دارند [۲]. در این مقاله می‌خواهیم نحوه استفاده از درختهای تصمیم دسته‌بند را برای یادگیری، پیش‌بینی و تصمیم‌گیری ترافیکی نشان دهیم.

۳- به کارگیری درختهای تصمیم به منظور یادگیری و پیش‌بینی وضعیت ترافیک

همانطور که پیش از این ذکر شد، هدف دسته‌بندی اینست که - با توجه به مقادیر متغیرهای هر پدیده جدید - کلاسی از بین کلاس‌های تعریف شده در محیط را به هر پدیده جدید تخصیص دهد. در حقیقت این عمل پیش‌بینی کلاس پدیده جدید بر اساس مقادیر متغیرهایش است [۳,۴]. روش‌های متعددی برای این پیش‌بینی وجود دارند که دسته‌بندی با استفاده از درختهای تصمیم یکی از آنهاست [۵,۶]. ساختار درخت تصمیم‌گیری یک ساختار درختی شبیه فلوچارت است که در این ساختار هر گره داخلی تستی (آزمونی) را بروی یک ویژگی مشخص می‌کند و هر شاخه خارج شده از این گره، دستاورد این تست را نشان می‌دهد و گره‌های برگ کلاسها یا توزیع کلاسها را ارائه می‌نمایند. بالاترین گره درخت گره ریشه نامیده می‌شود. دسته‌بندی با استفاده از درختهای تصمیم یک عملیات دو مرحله ایست. در اولین مرحله این روش، با استفاده از مجموعه داده‌های آموزشی، قوانین و روابط بین مقادیر متغیرها و کلاسها در قالب یک درخت تصمیم استخراج می‌شوند. درخت تصمیم حاصل در حقیقت دانشی است که استخراج شده و در مرحله بعد مورد استفاده قرار خواهد گرفت. در مرحله بعد مجموعه داده‌های آزمایشی به منظور تخمین میزان دقت درخت تصمیم ساخته شده بکار می‌رond. درخت تصمیم حاصل را می‌توان به منظور دسته‌بندی داده‌های دنیای واقعی بکار برد [۷]. این دسته‌بندی در حقیقت می‌تواند پیش‌بینی وضعیت‌های جدید و حتی تصمیم‌گیری در مورد آنها باشد. روند کلی استفاده از درخت تصمیم ایجاد شده بدین ترتیب است که یک گره ریشه در بالای درخت و برگها در پایینترین قسمت درخت قرار دارند. یک رکورد جدید در گره ریشه وارد می‌شود و در این گره یک تست صورت می‌گیرد تا معلوم شود این رکورد به کدامیک از گره‌های فرزند (شاخه پایینتر) می‌رود. این فرایند آنقدر ادامه پیدا می‌کند تا رکورد جدید به گره برگ برسد. تمام رکوردهایی که به یک برگ از درخت می‌رسند در یک کلاس قرار می‌گیرند. همچنین برای رسیدن از ریشه به یک برگ تنها یک مسیر وجود دارد و آن مسیر در حقیقت بیان قانونی است که برای دسته



بندی رکوردهایی که به آن برگ رسیده اند و در قالب کلاس متناظر با آن برگ دسته بندی شده اند استفاده شده است.

اثر بخشی یک درخت تصمیم گیری پس از ایجاد شدن باید اندازه گیری شود. برای این کار از یک مجموعه رکورد یا داده های آزمایشی استفاده می شود که از داده های اولیه ایجاد کننده درخت متفاوت می باشند. معیاری که در این قسمت محاسبه می شود عبارتست از درصد داده هایی که درست دسته بندی می شوند. همچنین کیفیت هریک از شاخه ها و مسیرهای ایجاد شده از ریشه به برگها باید مدنظر قرار گیرد. در واقع هر مسیر از ریشه به یک برگ یک قانون است که باید دقیق آن مورد نظر قرار گیرد. در اکثر مواقع بریدن و حذف شاخه ها و مسیرهای کم دقیق موجب بهبود قدرت پیش بینی درخت می شود. به این عمل حذف و بریدن شاخه های زاید، هرس کردن درخت (Pruning) گفته می شود.

در ادامه با یک مثال ساده یادگیری و پیش بینی یا تصمیم گیری با استفاده از درختهای تصمیم روی داده های ترافیکی را روشن تر توضیح می دهیم. مجموعه داده مختصراً به صورتی که در جدول ۱ مشاهده می کنیم در نظر می گیریم.

جدول ۱: مجموعه داده های آموزشی ترافیک

ساعت	دما	میزان ترافیک
۷	۲۰	۳
۷:۳۰	-۲	۳
۸:۳۰	۱۵	۲
۹	۱۵	۲
۸:۳۰	۶	۲
۱۰	-۲	۱
۹	-۲	۱
۹:۳۰	-۵	۱

این مجموعه داده که داده های آموزشی الگوریتم هستند دارای دو متغیر می باشند که یکی ساعت نمونه برداری و دیگری دمای محیط در زمان نمونه برداری است. کلاس این مجموعه داده وضعیت ترافیک در زمان برداشت نمونه را نشان می دهد و آن را با مقادیر ۱، ۲ و ۳ نشان می دهد که مقدار ۱ ترافیک روان، مقدار ۲ ترافیک کند و مقدار ۳ ترافیک شدید را نشان می دهد. همانطور که اشاره شد در این مجموعه داده تنها از دو پارامتر استفاده شده است تا پیچیدگی مساله کاهش یابد و قابلیت درک آن افزایش یابد. در مسائل واقعی تعداد این پارامترها ممکن است حتی تا صد مورد افزایش یابد و تعداد کلاسها هم که در این مثال ۳ مورد در نظر گرفته شده می تواند بیشتر باشد. هدف از ایجاد درخت تصمیم در این مساله استخراج رابطه بین دو متغیر ذکر شده با وضعیت ترافیک در قالب درخت و استفاده از درخت مورد نظر برای پیش بینی وضعیت ترافیک حالتی جدید است.

ایجاد درخت تصمیم یا به عبارتی یادگیری و استخراج الگو از دو مرحله تشکیل شده است:

• مرحله رشد و ایجاد درخت

• مرحله هرس کردن درخت

الگوریتم کلی ایجاد درخت برای یک درخت دودویی یعنی درختی که در هر گره آن تنها دو شاخه زده می شود به صورت زیر است:

Apply ss to D to find the splitting criterion

If n split

 Use best split to partition D to D1 and D2

 Build tree(n1,D1,ss)

 Build tree(n2,D2,ss)

End if

شیوه های مختلفی برای انتخاب نقطه شکست و انشعاب وجود دارند که تفاوت آنها مبنا و ریشه تفاوت روش های مختلف ایجاد درخت تصمیم است. از جمله این شیوه ها می توان به موارد زیر اشاره کرد:

$$(gini(T) = 1 - \sum p_j^2) : \text{Gini Index} \bullet$$

$$(entropy(T) = -\sum p_j * \log^2(p_j)) : \text{Entropy} \bullet$$

Cart •

$$(p_j / 2p_j) \bullet \text{ فراوانی نسبی از کلاس j در درخت T می باشد.}$$

$$Min(p_j) \bullet$$

C4.5 •

در روش Gini Index، معیار را برای تمام متغیر ها در گره حساب کرده و متغیری را که از همه کوچکتر باشد برای شاخه زدن انتخاب می کنیم. برای انتخاب نقطه انشعاب مورد نظر ابتدا مجموعه داده های آموزشی (S) را بر اساس متغیر منتخب مرتب می کنیم و سپس فرمول زیر را برای تمام حالتهایی که می توان شاخه را ایجاد کرد یعنی مجموعه S را به دو مجموعه S1 و S2 تقسیم کرد حساب می کنیم. نقطه انشعابی که (S) I(S) را مینیمم می کند را به عنوان نقطه انشعاب در نظر می گیریم.

$$I(S) = |S1|/|S| * I(S1) + |S2|/|S| * I(S2)$$

در اینجا روند ساخت درخت تصمیم گیری با استفاده از روش Gini Index برای نمونه داده های ترافیکی ذکر شده (جدول ۱) بررسی می کنیم. برای این منظور ابتدا مجموعه داده های آموزشی را بر اساس متغیر ساعت مرتب می کنیم (جدول ۲). سپس تک تک مقادیر متغیر مورد نظر را به عنوان

مقادیر نقطه شکست در نظر می گیریم و معیارهای فوق را برای آنها محاسبه می کنیم تا کمترین I(S) را بدست آوریم. مقداری از متغیر که این کمترین مقدار برای آن محاسبه شده را به عنوان بهترین نقطه انشعاب درخت در نظر می گیریم.

جدول ۲: مجموعه داده های آموزشی که بر اساس متغیر ساعت نمونه بردازی مرتب شده اند.

ساعت	دما	میزان ترافیک
۷	۲۰	۳
۷:۳۰	-۲	۳
۸:۳۰	۱۵	۲
۸:۳۰	۶	۲
۹	۱۵	۲
۹	-۲	۱
۹:۳۰	-۵	۱
۱۰	-۲	۱

برای محاسبه I(S) هر نقطه ساختاری به شکل زیر را تشکیل می دهیم:

جدول ۳: ساختار داده در نظر گرفته شده برای محاسبه I(S)

کلاس	۱	۲	۳
کمتر	a	b	c
بیشتر	d	e	f

که در ردیف اول کلاسها قرار دارند. در ردیف دوم مشخص شده است که چه تعداد از داده های متعلق به هر کلاس مقدار متغیر منتخبشان از مقدار مورد نظر که جدول برای آن ایجاد شده کمتر است. ردیف سوم هم نشان دهنده تعداد داده هایی است که مقدار متغیرشان بیشتر بوده است. پس از ایجاد این ساختار داریم:

$$I(S1) = 1 - (a / (a + b + c))^2 - (b / (a + b + c))^2 - (c / (a + b + c))^2$$

$$I(S2) = 1 - (d / (d + e + f))^2 - (e / (d + e + f))^2 + (f / (d + e + f))^2$$

$$I(S) = (a + b + c) / (a + b + c + d + e + f) * I(S1) + (d + e + f) / (a + b + c + d + e + f) * I(S2)$$

با توجه به ساختار و فرمولهای فوق برای داده های ترافیکی ذکر شده داریم (جدول ۴):

جدول ۴: مقداردهی ساختار داده تعریف شده برای حالت ۷ ساعت

کلاس	۱	۲	۳
کمتر	۰	۰	۱
بیشتر	۳	۳	۱

$$I(s1) = 1 - (1/1)^2 = 0$$

$$I(S2) = 1 - (3/7)^2 - (3/7)^2 - (1/7)^2 = 0.61$$

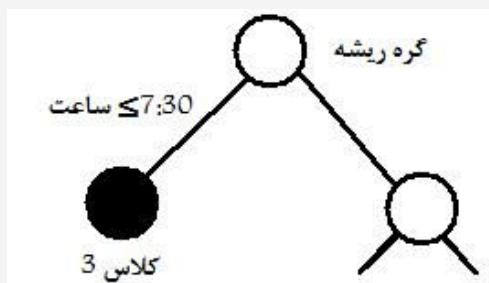
$$I(S) = 1/8 * 0 + 7/8 * 0.61 = 0.53$$

به همین ترتیب برای مقادیر دیگر متغیر ساعت داریم (جدول ۵):

جدول ۵: محاسبه مقادیر $I(S)$ برای متغیر ساعت

ساعت	≤ 7	$\leq 7:30$	$\leq 8:30$	≤ 9	$\leq 9:30$	≤ 10
$I(S)$	۰,۵۳	۰,۳۷	۰,۴۳	۰,۴۵	۰,۵۷	۰,۶۵

با توجه به اینکه ۰,۳۷ کمترین مقدار محاسبه شده برای $I(S)$ است، مقدار $7:30$ به عنوان نقطه انشعاب و $\leq 7:30$ به عنوان شرط انشعاب انتخاب می شود (شکل ۳). از مجموعه داده های آموزشی دو داده در شرط $\leq 7:30$ صدق می کنند که هر دو متعلق به یک کلاس مشترک یعنی کلاس ۳ هستند. بنابراین گره سمت چپ به عنوان یک برگ در نظر گرفته می شود که دیگر نیازی به شاخه زدن ندارد. ولی گره سمت راست که نمونه داده های با شرط $> 7:30$ به آن راه یافته اند به دلیل یکدست نبودن کلاس داده هایش نیاز به شاخه زدن دارند. برای این منظور روند فوق را روی متغیر دمای داده هایی که به این گره راه یافته اند ادامه می دهیم.



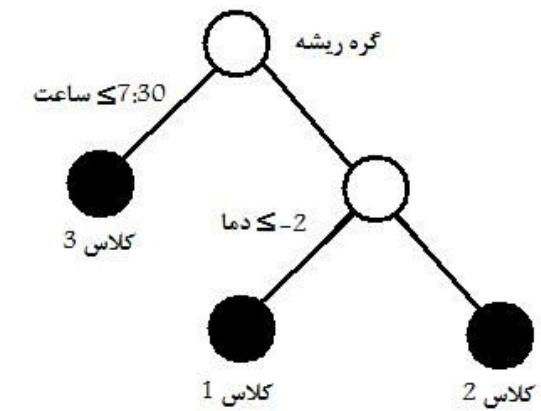
شکل ۳: انشعاب اول در ساختار درخت ایجاد شده برای مجموعه داده های ترافیک

پس از محاسبه مقادیر $I(S)$ برای متغیر دما و داده هایی که در شاخه سمت راست قرار می گیرند داریم (جدول ۶):

جدول ۶: محاسبه مقادیر $I(S)$ برای متغیر دما

دما	≤ -5	≤ -2	≤ 6	≤ 15
$I(S)$	۰,۴	۰	۰,۲۵	۰,۵

با توجه به اینکه کمترین مقدار $I(S)$ برای مقدار -2 بودست آمده است این مقدار متغیر به عنوان نقطه انشعاب انتخاب می شود. در اینصورت از مجموعه داده هایی که از مرحله قبل به این گره راه یافته بودند سه نمونه داده که کلاس آنها 1 است به زیر شاخه سمت چپ و سه نمونه داده که کلاس همه آنها 2 است به زیر شاخه سمت راست انتقال می یابند. بنابراین گره ایجاد شده در سمت چپ یک گره برگ با کلاس 1 و گره ایجاد شده در سمت راست نیز یک گره برگ با کلاس 2 می باشد (شکل ۴).



شکل ۴: درخت کامل ایجاد شده برای مثال ذکر شده

در دنیای واقعی، تعداد پارامترهایی که در مرحله ای ایجاد درخت بکار می روند بسیار بیشتر است. تعداد کلاسها و نیز نمونه داده های آزمایشی نیز بسیار بیشتر است. در چنین مساله ای ممکن است درخت ایجاد شده چندین سطح داشته باشد و تعداد گره های آن نیز بسیار بیشتر است.

پس از ایجاد درخت، با استفاده از نمونه داده های تست که کلاس آن داده ها نیز مشخص است درخت حاصل را تست می کنیم. بدین ترتیب که با استفاده از درخت حاصل کلاس آنها را پیش بینی می کنیم و آن را با کلاس واقعی آنها مقایسه کرده درصد پیش بینی درست درخت را محاسبه می کنیم. به عبارتی دقت درخت حاصل را بدست می آوریم تا از صحت آن اطمینان حاصل کنیم.

برای پیش بینی کلاس یک نمونه داده جدید بدین ترتیب عمل می کنیم که از گره ریشه شروع می کنیم. شرط هر گره را بروی نمونه داده مورد نظر بررسی می کنیم و در مسیری که داده مورد نظر در شرط آن صدق می کند پیش می رویم تا به یک گره برگ برسیم. کلاسی که در مرحله ایجاد درخت به آن گره برگ نسبت داده شده است به عنوان کلاس پیش بینی شده برای این نمونه داده در

نظر گرفته می شود. بدین ترتیب دانش یادگرفته شده در قالب درخت در این مرحله به منظور تعیین یا پیش بینی کلاس داده مورد نظر بکار می رود.

۴- نتیجه گیری و کارهای آتی

همانطور که در ابتدا ذکر شد در این مقاله سعی کرده ایم سازگاری و قابلیت استفاده از روش دسته بندی با استفاده از درختهای تصمیم که یکی از روشهای داده کاوی است را در مساله ترافیک و پیش بینی آن نشان دهیم. برای این منظور و برای جلوگیری از پیچیدگی مساله از یک مثال بسیار ساده استفاده کردیم و نشان دادیم که چگونه درخت تصمیم برای آن ایجاد می شود و چگونه دسته بندی داده های جدید با استفاده از درخت حاصل انجام می شود. با توجه به مطالبی که بیان کردیم می توان نتیجه گرفت که داده کاوی می تواند کاربرد مناسبی در مساله پیش بینی ترافیک داشته باشد.

استفاده از سایر روشهای دسته بندی و بررسی و مقایسه راه حل های این روشهای متنوع را به کارهای آتی موقول می کنیم. همچنین می توان کاربرد روشهای دیگر مثل شبکه های عصبی که برای یادگیری و پیش بینی بکار می روند را روی این مساله بررسی و نتایج حاصل را با روشهای داده کاوی مقایسه کرد.

۵- مراجع

- 1- Qiang Ding, Qin Ding, William Perrizo, 2002, Decision Tree Classification of Spatial Data Streams Using Peano Count Trees. SAC 02: 413-417.
- 2- Ruoming Jin, Gagan Agrawal, 2003, Efficient Decision Tree Construction on Streaming Data. KDD 03: 571-576.
- 3- Nicholas R. Howe, Toni M. Rath, and R. Manmatha, 2005, Boosted Secision Trees for Word Recognition in Handwritten Document Retrieval. SIGIR: 377-383.
- 4- Patrick Knab, Martin Pinzger, and Abraham Bernstein, 2006, Predicting Defect Densities in Source Code Files with Decision Tree Learners. MSR: 119-125.
- 5- J. Gehrke, V Ganti, R. Ramakrishnan, and W.-Y. Loh. , 1999, BOAT - Optimistic Decision Tree Construction. In A. Delis, C. Faloutsos, and S. Ghandeharizadeh, editors, Proc. ACM SIGMOD'99, Philadelphia, USA, pages 169-1 80. ACM.
- 6- J. Gehrke, R. Ramakrishnan, and V. Ganti. , 1998, RainForest – A Framework for Fast Decision Tree Construction of Large Datasets. In A. Gupta, O. Shmueli, and J. Widom, editors, Proc. VLDB '98, New York, USA, pages 4 16427. Morgan Kaufmann.
- 7- F. Seifi, H. Ahmadi, M. Kangavari, 2007, Twins Decision Tree Classification: A Sophisticated Approach to Decision Tree Construction, ICCSA'07, Florida USA.