dataacademy.ir

تحلیل نمونه های متفاوت در داده کاوی

سوسن حسینی

دانشجوی کارشناسی ارشد،نرم افزارکامپیوتر،دانشگاه آزاداسلامی مشهد

Sosan.hosseini@gmail.com

چکیده

داده کاوی وسیله ای برای کشف داده ها از منابع اطلاعاتی محسوب می گردد و به بررسی نمونه های مختلف و متشابه در پایگاه داده ها می پردازد .هر نمونه با چندین ویژگي مشخص شده است به طوري كه انواع مختلفي از مقادير براي هر ويژگي وجود دارد،كه می توان باتحلیلی صحیح برروی نمونه ها به دانش مفیدی وجدیدی دست یافت.

کلید واژه- داده کاوی ،کشف داده، نمونه های متفاوت، پایگاه داده، دانش مفید.

1- مقدمه

داده كاوي (Data Mining)، يكي از مهمترين روش هايي است كه به وسيله آن الگوهاي مفيد نهان در داده ها كشف شده ، اطلاعاتي را در اختيار كاربران و تحليل گران قرار مي دهد تا بر اساس آنها تصميمات مهم و حياتي اتخاذ شوند. اطلاعات و دانش بدست آمده از اين طريق، جهت كاربردهاي گوناگوني از قبيل تحليل بازار، تشخيص كلاهبرداري، حفظ مشتري، كنترل توليد و ... را شامل مي شود.[1] به كل فرايند بكارگيري متدولوژي مبتني بر كامپيوتر از جمله روش هاي جديد براي دريافت دانش واطلاعات ازداده هاداده كاوي گفته مي شود.درحقيقت،داده كاوي جستجوي لازم براي يافتن اطلاعات ارزشمندوغيربديهي ازميان حجم زياد داده ها مي باشد. فعاليتهاي داده كاوي را مي توان در دو گروه زير طبقه بندي كرد:

الف-داده كاوي پيشگويانه ، مدلي از سيستم ارائه مي دهد كه توسط مجموعه داده هاي مشخصي توصيف مي شود.

ب- داده كاوي توصيفي ، اطلاعات جديد و غيربديهي را بر اساس مجموعه داده هاي موجود ارائه مي دهد.

داده كاوي بخشي از يک فرايند بزرگتر موسوم به كشف دانش در پايگاه داده (Knowledge Discovery in Databases) مي باشد اما در اكثر موارد از واژه داده كاوي بجاي KDD استفاده مي شود[2] .ادامه این مقاله اینگونه سازماتدهی شده است دربخش دوم مروری برنمونه ها درداده کاوی ودربخش سوم به تحلیل نمونه های متفاوت ودرانتهابه نتیجه گیری پرداخته خواهدشد.

۲-بررسی نمونه

هر نمونه با چندين ويژگي مشخص مي شود، طوري كه انواع مختلفي از مقادير براي هر ويژگي وجود دارد[3].دو نوع رايج آن categorical و numeric مي باشد ،مقادير numeric شامل متغيرهاي real و يا متغير هاي integer مي باشد، نوع numeric دوويژگي ، مهم در رابطه دارد :

۱ـمنظم.۲ـفاصله .در مقابل categorical ، هيچ يك از اين دو رابطه را ندارند .راه ديگر [4] براي نمونه بندي متغيرها ، مبتني بر مقاديرشان مي باشد،به طوري كه آيا متغيرهاي پيوسته انديا گسسته متغيرهاي پيوسته به متغيرهاي metric (كمي) نيز معروف اند .آنها هم با مقياس فاصله اي و هم با مقياس نسبت اندازه گيري مي

آکادمی داده
dataacademy.ir

dataacademy.ir

شوند . تفاوت اين روش با مقياس در چگونگي تعريف نقطه صفي در مقياس مي باشد،نقطه صفر در مقياس فاصله به طور قرار دادي مشخص شده است و چيزي نيست كه بتوان آن را با اندازه گيري مشخص كرد . براي مثال ، مقياس دما ، كه درجه صفر ، رابطه نسبت براي متغيرهايي كه با استفاده از مقياس فاصله اندازه گيري شده اند درست كار نمي كند . براي مثال ۸۰ درجه فايرنهات بر دو برابر بودن گرماي ۴۰ درجه دلالت نمي كند . در مقابل مقياس نسبتي داراي نقطه صفر مطلق مي باشد و به تبع آن ، رابطه نسبتي براي اين نوع اندازه گيري به خوبي كار مي كند،كميت هاي مانند ارتفاع ، قد و حقوق با استفاده از اين مقياس اندازه گيري مي شود . متغيرهاي پيوسته به صورت مجموعه داده بزرگ با مقاديري چون real ،Integer نمايش داده مي شوند .متغيرهاي گسسته(متغيرهاي كيفي)يكي از دو نوع مقياس هاي كيفي ترتيبي يا اسمي اندازه گيري مي شوند يا مقاديرشان مشخص مي شود . يك مقياس اسمي يك مقياس بدون ترتيب مي باشدكه نشانه ها،كاراكترها و اعداد مختلف حالت هاي (مقادير) مختلف متغيير را نشان مي دهد،يك مثال از متغيرهاي اسمي (غير ترتيبي) ،يك كاربرد با مقاديرصنعتي ، تجاري و خانگي مي باشد . اين مقادير مي تواند به صورتي الفبايي C,B,A و يا اعداد به صورت ۱و۲و۳ كد شوند،اما آنها داراي هيچگونه ترتيبي نمي باشند.مثال ديگر ، Zip-code مي باشد .در هر دو مثال ، اعداد استفاده شده براي يك متغير ، نسبت به هم هيچگونه نظم و رابطه خاصي ندارند . يك مقياس ترتيبي شامل ، نظم و گسستگي مي باشد ، مانند Ranking يك متغيير ترتيبي يك متغير categorical مي باشدكه داراي رابطه نظم مي باشد،اما فاقد رابطه فاصله مي باشد . مثالي از اين نوع متغير ، رتبه دانش آموزان داخل يك كلاس و مدل هاي طلا و نقره و برنز در رقابت هاي ورزشي مي باشد . براي مقياس منظم [5]، خطي بودن نياز ضروري نمي باشد . براي مثال ، فاصله بين رتبه هاي ۴و۵ دانش آموزان ، نياز ندارد كه فاصله ۱۵ و ۱۴ يكسان مي باشد . عموماً متغيرهاي منظم يك متغير عددي را مجموعه كوچكي از مقادير orerlap متناظر با مقادير يك متغير منظم مي كند.يك كلاس ويژه ازمتغيرهاي گسسته ، متغيرهاي دوره اي يا تناوبي مي باشد . متغيرهاي تناوبي داراي ويژگي رابطه فاصله اي باشند اما رابطه نظم در آنها وجود ندارد.براي مثال ، روزهاي هفته،دوشنبه و سه شنبه به عنوان مقاديراين متغير مي باشندكه فاصلة بين آنهايك مي باشد ، اما دوشنبه مي تواند بعد و يا قبل از جمع قرار گيرد .در نهايت ، يك بعد ديگر از طبقه بندي داده ها مبني رفتارشان نسبت به زمان مي باشد . بعضي از داده ها با توجه به زمان تغيير نمي كنند و داده هايي استاتيك مي باشند . در طرف ديگر بعضي از داده ها با توجه به زمان تغيير مي كنند و داده هاي ديناميك يا زماني مي باشند . اكثر روش هاي Data mining براي داده هاي استاتيك مناسب مي باشند و بعضي از پيش پردازش ها و رسيدگي هاي ويژه اي براي داده هاي ديناميك لازم مي باشد .اكثر مشكلات Data mining ناشي از وجود مقدار زيادي از مثال ها با انواع مختلفي از ويژگي ها مي باشد .

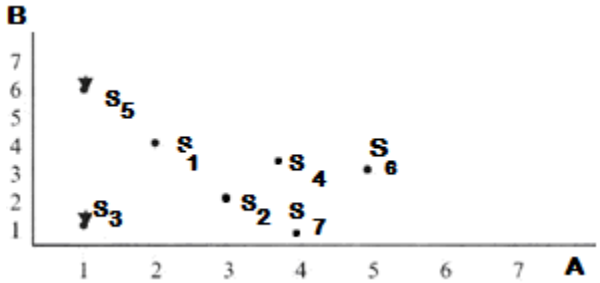
3-تحليل نمونه متفاوت:

بسياری از مواقع ، در مجموعه داده های بزرگ، نمونه هايی وجود دارد كه از رفتار كلی مدل داده ها تبعيت نمی كنند. چنين نمونه هايی كه تا حد زيادی متفاوت انديا،مابقی مجموعه داده ها ناهماهنگ هستند، نمونه های متفاوت ناميده می باشند،كه می توانند توسط خطای ارزيابی ايجاد گردند يا نتيجه تنوع فطری داده ها بوده اند. به عبارت ديگر،اگر در پايگاه داده ها، تعداد بچه های يک فرد ۲۵ باشد اين داده غير معمول است و بايد چک شود. اين اشتباه می تواند يک خطای چاپی يا می تواند صحيح بوده و نشان دهنده تنوع واقع برای خصيصه داده شده باشد.بسياری از الگوريتم های داده كاوی در تلاش اند،تا تأثير نمونه های متفاوت بر نمونه نهايی را كاهش داده يا آنها را در ،مراحل پيش از پردازش ريشه كن سازند. تحليل گر داده كاوی بايد در ريشه كنی اتوماتيک نمونه های متفاوت بسيار مراقب باشد،زيرا اگر داده ها درست باشند می تواند منجر به فقدان اطلاعات مهم مخفی ، گردد. برخی از كاربردهای داده كاوی بر روی شناسايی نمونه متفاوت متمركز و اين امر نتيجه فوری يک تحليل داده می باشد. به عنوان مثال در حين شناسايی تعاملات كارت اعتباری جعلی در يک بانک نمونه های متفاوت

آکادمی داده
dataacademy.ir

dataacademy.ir

، نمونه هایی معمولی هستند که ممکن است فعالیت جعلی را نشان دهند و کل فرآیند داده کاوی شوند نمونه های متفاوت چندان مفید نیستندو بیشتر نتیجه خطاهایی در جمع آوری داده ها و سپس مشخصه ای از یک مجموعه داده ها هستند.شناسایی نمونه متفاوت و رفع بالقوه آن از یک مجموعه داده ها می تواند به عنوان فرآیندی از انتخاب k در خارج از نمونه هایی توصیف شود که با توجه به داده های باقی مانده به طور قابل توجه ای ناهماهنگ هستند. مشکل تشریع نمونه های متفاوت، بی اهمیت نمی باشد، خصوصاً در نمونه های چند بعدی، شیوه های تجسم داده ها که شناسایی نمونه متفاوت برای ابعاد یک تا سه مفید هستند، در داده های چند بعدی ضعیف تر می باشند،نمونه ای از یک تجسم نمونه های دو بعدی و شناسایی بصری نمونه های متفاوت در نمودار(1) داده می شود.



نمودار (1): تجسم مجموعه داده های دو بعدی برای شناسایی نمونه های متفاوت[6].

ساده ترین رویکرد جهت شناسایی نمونه متفاوت برای نمونه ها، بر اساس آمار می باشد. با فرض اینکه تنوع مقادیر داده شود یافتن پارامترهای اساسی آماری از قبیل مقدار میانگین و واریانس ضروری است. بر اساس این مقادیر و عدد مورد انتظار (یا پیش بینی شده)درموردنمونه های متفاوت،اثبات مقدار ورودی به عنوان عملکردی از واریانس، امکان پذیر است.تمام نمونه های خارج از مقدار ورودی،کاندیداهای نمونه های متفاوت می باشند.مشکل اساسی موجود در این شیوه ساده، یک برداشت پیشوا در مورد توزیع داده ها است. در بسیاری از نمونه های دنیای واقعی، توزیع داده ها ممکن است شناخته شده نباشد.اگر مجموعه ی معین ازداده نشان دهنده مشخصه سن با،بیست مقدار متفاوت باشدومقدارورودی توزیع نرمال داده،به صورت(انحراف استاندارد*۲+ میانگین = ورودی) انتخاب شود،تمام داده هایی که خارج از دامنه { ۱۳۱/۲ و ۵۴/۱ - } می باشند، نمونه های متفاوت بالقوه خواهندبود.(دانش اضافی ازمشخصه های خصیصه سن همیشه بیشترازصفراست).ممکن است دامنه به صورت{ ۱۳۱/۲ و ۰ } بیشتر کاهش دهد. بر اساس ضابطه معین در نمونه ما سه مقدار وجود دارد که نمونه های متفاوت می باشند : ۱۵۶ ، ۱۳۹ ، ۶۷ - . شناسایی نمونه متفاوت بر اساس فاصله، شیوه ثانویه می باشدکه برخی ازمحدودیت های تحمیلی از سوی رویکرد آماری را از بین می برد.مهم ترین تفاوت این است که این شیوه برای نمونه های چند بعدی قابل کاربردمی باشددرحالی که توصیف کنندگان تنها یک بعد واحدیاچندین بعدرا، به صورت جداگانه ارزیابی می نماید.دشواری،محاسباتی این شیوه،ارزیابی مقادیرفاصله بین،نمونه ها،دریک مجموعه n بعدی ازداده ها می باشد.بنابراین یک نمونهsi دریک مجموعه داده های s،یک نمونه متفاوت می باشد.اگرحداقل یک بخشp،ازنمونه هایsدرفاصله ای بیشتر از d قرار داشته باشدو ملاک شناسایی نمونه متفاوت برحسب دو پارامتر p,d می باشدکه قبل ازاستفاده ازدانش مربوط به داده هاداده شودیادرحین تکراررویکرد آزمون وخطا،انتخاب گویاترین نمونه متفاوت،تغییرداده شوند.برای نشان دادن رویکرد،می توان مجموعه ای ازنمونه های دو بعدی راتحلیل کرد.نتیجه نمونه های متفاوت،مقادیرورودی $d \geq 3, p \geq 4$ می باشد.مسافت های اقلیدوسی $d = [(X1 - X2)^2 (y1 - y2)^2\}^{y2}$،برای مجموعه s در جدول۱ دیده می شود.

آکادمی داده
dataacademy.ir

dataacademy.ir

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1		2.236	3.162	2.236	2.236.	3.162	2.828
S_2			2.236	1.414	4.472	2.236	1.000
S_3				3.605	5.000	4.472	3.162
S_4					4.242	1.000	1.000
S_5						5.000	5.000
S_6							1.414

جدول ۱:مسافت های اقلیدسی برای p>=4 و [7] d>=3

بر اساس جدول (۱) می توان با مسافت ورودی داده شده $(d=3)$ برای هر نمونه،یک مقدار،را برای پارامتر p،بالای مقدار ورودی است: $(p=4)$. همین نتایج را می توان از طریق بازدید بصری یک مجموعه داده بدست آورد که در جدول۲ ارائه شده است. البته مجموعه داده تعیین شده بسیار کوچک است و یک نمودار گرافیکی دو بعدی امکان پذیر و مفید است. برای تحلیل داده های n بعدی در دنیای واقعی، فرآیند تجسم بسیار دشوارتر است و رویکردهای تحلیل در شناسایی نمونه متفاوت غالباً کاربردی تر و قابل اطمینان تر هستند.

Sample	p
S_1	2
S_2	1
S_3	5
S_4	2
S_5	5
S_6	3

جدول۲: نتایج مسافت های اقلیدسی[8]

تکنیک های بر اساس انحراف سومین طبقه از شیوه های شناسایی نمونه متفاوت هستند. این تکنیک ها را شبیه سازی می نمایند که بشر می تواند در آن ، نمونه های غیر معمول را،از مجموعه نمونه های مشابه دیگر جداسازی نماید. این شیوه هاخصوصیات اساسی مجموعه ستونه را تعریف می نماید و تمام نمونه هایی که از این خصوصیات منحرف می گردند، نمونه های متفاوت می باشند.تکنیک غیر متوالی یک رویکرد امکان پذیر است که بر اساس یک عملکرد عدم شباهت می باشد . یک عملکرد امکان پذیر،عدم شباهت برای مجموعه ای معین از نمونه های n عبارت است از انحراف کلی مجموعه نمونه می باشدو باتعریف کردن، کوچکترین زیر مجموعه نمونه ها است،که رفع کردن آن منجربه بزرگترین کاهش عملکردعدم شباهت،مجموعه باقیمانده می گردد. وظیفه کلی برای پیدا کردن نمونه های متفاوت با،این شیوه می تواند بسیار پیچیده باشد (رشد سریع ترکیبی گزینه های متفاوت مجموعه نمونه های متفاوت بالقوه ـ مجموعه بسیار معروف استثنا می گردد) و می تواند به صورت تئوری به عنوان یک مشکل سخت NP تعریف شود.اگر به یک پاسخ کمتر از حد مطلوب رضایت بدهیم،سختی الگوریتم می تواند تا سطح تک بعدی کاهش یابد.

آکادمی داده
dataacademy.ir

dataacademy.ir

۴-نتیجه گیری

شناخت ویژگی ها،درتحلیل نمونه متفاوت بسیارمفیدمی باشدبراساس ضابطه در، نمونه های متفاوت آزمایش شده(۱۳۹،۱۵۶ ، ۶۷ -)می توان نتیجه گیری کرد،هرسه تا دارای خطاهای چاپی می باشند(داده های وارد شده بااعداداضافی یایک علامت « - » اضافی).بااستفاده از یک رویکرد متوالی با استفاده کردن از شیوه مشتاق، الگوریتم با انتخاب نمونه ای در هر مرحله که سبب بیشترین کاهش درواریانس کلی می گردد،نمونه به نمونه اندازه به صورت متوالی کاهش می دهد.

مراجع

[1] K.R.Venugopal,K.G.srinivasa and L.M,patnaik,soft computing for data mining applications,pp. 85-87,2009.
[2] IH,Frank “An instruction to data mining”,pp.55-59,2007.
[3] Brunk.c,Kelly.j,”an integrated system for data mining “,proc.3 ,int.1.conf.on knowledge discovery and data minig,pp.135-138,1997.
[4] Javad azimi,Reza Davoodi,Morteza analoui ,“fast Convergence Clustring ensemble,In Conference of Data mining and Data warehouse”,pp.14-16,2006.
[5] W. OCK .Day,”consensus methods as tools for data analysis.in classification and related method for data analysis” ,Elsiver scince publishers B.V ,pp.317-320,1988.
[6] Brunk.c,Kelly.j,”an integrated system for data mining “,proc.3 ,int.1.conf.on knowledge discovery and data minig,pp.135-138,1999.
[7] David Hand,Heikki Mannila,”Principles of data mining”,pp.246-248,2001.
[8] J.Han, M.Kamber,”Date Mining: Goncepts and Techinqeus”, pp.54-56,2000.

آکادمی داده
dataacademy.ir