مدل‌سازی بیزی بر اساس داده‌های حاصل از اینترنت اشیاء

فرزاد اسکندری[۱]، سیما نقی زاده اردبیلی[۲] سروش پاک‌نیت[۳]

تاریخ دریافت: ۹۹/۸/۳۰
تاریخ پذیرش: ۹۹/۱۲/۲۷

چکیده:

اینترنت اشیاء با دارا بودن قابلیت بسیار بالا برای بهره‌ور نمودن کسب‌وکارها در حوزه‌های مختلف از جمله صنایع به‌عنوان انقلاب آتی در فناوری اطلاعات و ارتباطات معرفی‌شده است. این بهره‌وری در زمینه بروز نوآوری و ارائه قابلیت‌های نو برای کسب‌وکارها است. صنایع مختلف در خصوص اینترنت اشیاء واکنش‌های مختلفی را نشان داده‌اند اما آنچه واضح است این است که اینترنت اشیاء در تمامی کسب‌وکارها و صنایع دارای کاربرد است. این کاربردها در برخی صنایع مانند بهداشت و حوزه سلامت و یا حمل‌ونقل پیشرفت چشمگیری داشته اما در صنایع دیگر همچون کشاورزی و دامداری درحال‌توسعه است. در واقع تولید داده‌ها بر مبنای اینترنت اشیاء از ارکان اصلی در حوزه مه داده‌ها و علم داده‌ها خواهد بود. لذا استفاده از مفاهیم و مدل‌های آماری که در علم داده‌ها مورداستفاده قرار می‌گیرند به‌خوبی می‌توانند در این‌گونه داده‌ها مورداستفاده قرار گیرند. از جمله مدل‌های آماری معتبر آمار بیزی برای مه داده‌ها است که مبنای استفاده در این پژوهش قرارگرفته است. در این پژوهش ضمن معرفی مفاهیم مهم و معتبر که در حوزه مه داده‌ها مورداستفاده قرار می‌گیرند به‌طور خاص اصول آمار بیزی برای مه داده‌ها و به‌طور مشخص برای داده‌های حاصل از اینترنت اشیاء توضیح داده‌شده است. به‌صورت کاربردی نیز در دو حوزه رفتار اجتماعی افراد برای علاقه‌مندی به استفاده از وسیله نقلیه و ترافیک شهری بررسی‌شده است که نتایج معتبری از نظر علمی و کاربردی در برداشته است.

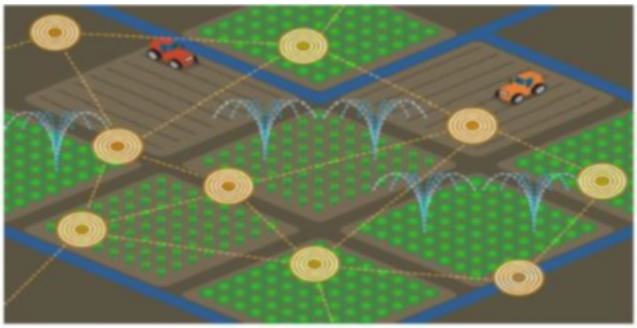
واژه‌های کلیدی: اینترنت اشیاء، نظریه بیزی، رده‌بندی، مصورسازی داده‌ها.

۱ مقدمه

طی سال‌های گذشته همگام با همه دنیا، اینترنت اشیاء موردتوجه محققان کشورمان قرارگرفته است. فناوری اینترنت اشیاء نقش بسیار مهمی در دنیای کارآفرینان بازی می‌کند. کسب‌وکارهای متعددی بر محور این فناوری راه‌اندازی شده‌اند، درحالی‌که این مفهوم و این فناوری در ابتدای راه خود قرار دارد و هر روز بیش‌ازپیش تغییر و تحولات جدیدی در آن رخ می‌دهد. استفاده از این فناوری برای کارآفرینان و محققین خلاق ایرانی یک فرصت گران‌بها به شمار می‌رود که می‌تواند به بهبود فضای کسب‌وکار و اشتغال‌زایی در کشور کمک شایانی نماید. در این پژوهش، پس از بررسی متون مرتبط با علوم و تکنولوژی‌های اینترنت اشیاء به تحلیل و مدل‌سازی آماری مجموعه داده مربوط به‌موقعیت جغرافیایی افراد (گردآوری‌شده توسط شرکت مایکروسافت) بر اساس روش رگرسیون لوژستیک می‌پردازیم. مباحثی مانند ناهمگونی اشیاء، تبادل اطلاعات بهینه‌سازی مصرف انرژی، مدیریت داده و حفظ امنیت و حریم خصوصی مؤلفه‌های مهم در تحلیل و توسعه اینترنت اشیاء از منظرهای گوناگون هستند. می‌توان رشد اینترنت اشیاء را در شهرهای بزرگ دنیا مانند ریودوژانیرو، بیجینگ و دهلی‌نو مشاهده نمود، در جایی که هزاران سنسور کیفیت هوا، ترافیک و سیستم‌های آب و فاضلاب را مانیتور می‌کنند. دولتمردان با استفاده از تکنولوژی‌های اینترنت اشیاء و تحلیل داده می‌توانند منابع را به درستی مدیریت کرده و رشد اقتصادی قابل‌توجهی در فعالیت‌ها و تصمیم‌های خود داشته باشند. اینترنت اشیاء در بخش سلامتی نسبت به بخش‌های دیگر تأثیر بیشتری داشته است. بر طبق پیش‌بینی‌ها حدود ۴۰ درصد از اقتصاد جهان از برنامه‌های کاربردی اینترنت اشیاء در حوزه سلامت تأثیر خواهد گرفت. معیار دسترس‌پذیری برنامه‌های کاربردی در حوزه سلامت رشد خیره‌کننده‌ای در سراسر جهان داشته است. یکی از کاربردهای دیگر اینترنت اشیاء در بخش کشاورزی است که یک حوزه بکر و حاصلخیز برای برنامه‌های کاربردی در این حوزه است. انتظار می‌رود در سال ۲۰۵۰ در حدود دو بیلیون نفر از جمعیت جهان به چالش حاد در مواد غذایی دچار شوند و سرمایه‌گذاری در حوزه کشاورزی با استفاده از اینترنت اشیاء می‌تواند راه‌حلی برای این مشکل باشد.

[۱] استاد گروه آمار دانشگاه علامه طباطبایی (نویسنده مسئول). askandari@atu.ac.ir
[۲] استادیار سازمان سنجش آموزش کشور s_naghizadeh12@yahoo.com
[۳] دانش‌آموخته رشته رایانه دانشگاه علامه طباطبایی .pakniat.sh@gmail.com

اینترنت اشیاء با شناخت قطعات و محصولات جعلی باعث ایجاد امنیت و آرامش در زمینه سرویس‌های هوایی نیز می‌شود. صنعت حمل‌ونقل هوایی در مقابل خطر قطعات تائید نشده و مشکوک بسیار آسیب‌پذیر است؛ بنابراین قطعات غیراستاندارد به‌شدت امنیت یک هواپیما را به خطر می‌اندازد. همچنین از جمله کاربردهای اینترنت اشیاء در صنعت خودرو می‌توان به کاربرد تجهیزات هوشمند در جهت مشاهده و گزارش پارامترهای متفاوت از فشار داخل تایرها گرفته تا تخمین فاصله از سایر وسایل در حال حرکت در جاده اشاره کرد.



شکل ۱. نمایی از وضعیت کشاورزی که بر اساس سنسورها کنترل می‌شود

در حوزه‌های مختلف دیگر مانند صنعت تولید، صنعت نفت و گاز، صنعت حمل‌ونقل، زراعت و تولیدمثل، رسانه و صنعت سرگرمی، صنعت بیمه، شبکه بازیافت، شبکه هوشمند برق، معادن و استخراج مواد معدنی، خانه هوشمند و نظارت بر آزمون‌های سراسری و انتخابات نیز اینترنت اشیاء بسیار کاربرد دارد. از ارکان اصلی ارتقا کیفیت اینترنت اشیاء می‌توان از پلتفرم‌ها، فناوری شبکه موبایل، پردازش و ذخیره‌سازی داده‌های ابری و همچنین تجزیه‌وتحلیل و امنیت آنان را بیان نمود. همچنین شرکت‌های برتر در زمینه اینترنت اشیاء اینتل، مایکروسافت، گوگل، IBM، سامسونگ، اپل، گارتنر، Oracle و جنرال الکتریک هستند.

در این پژوهش می‌خواهیم به‌وسیله آمار بیزی داده‌های حاصل از اینترنت اشیاء را که در بانک‌های اطلاعاتی ذخیره می‌شوند مورد تجزیه‌وتحلیل مناسب قرار دهیم. در واقع در این پژوهش به‌طور مشخص می‌خواهیم ابتدا یک شناسایی مطلوب از علوم و تکنولوژی‌های مرتبط با موضوع اینترنت اشیاء انجام دهیم، سپس بر اساس الگوریتم‌های خوشه‌بندی و همگن‌سازی مه‌داده‌ها و با استفاده از مدل تحلیلی آمار بیزی به ارزیابی داده‌ها بپردازیم. لازم به ذکر است برای ایجاد سرویس‌ها در هر زمان و مکان تعداد زیادی از اشیاء از طریق حسگر هوشمند به اینترنت متصل می‌شوند که باید با هویت یکتای خود با یکدیگر ارتباط برقرار کنند. این رفتار و ارتباط با موضوع اینترنت اشیاء در شکل ۲ نشان داده‌شده است. در بخش بعدی با این تفکر به ارائه مبانی نظری و تحلیل داده‌ها پرداخته می‌شود.

شکل ۲. الگوریتم "هر" که اینترنت اشیاء با آن در ارتباط است

۲ پیشینه‌ی پژوهش

در سال ۲۰۰۲ مجله‌ی علم داده‌ها توسط کمیته‌ی اطلاعات علوم و فناوری از شورای بین‌المللی علوم راه‌اندازی شد. در این کمیته به موارد مختلفی از جمله بررسی روش‌های رده‌بندی پرداخته‌شده است. روش‌های رده‌بندی در زمینه‌های مختلفی کاربرد دارد که از این میان می‌توان به تشخیص کلاه‌برداری، بازاریابی هدف، پیشگویی کارایی، تولید محصول و تشخیص بیماری اشاره کرد. بسیاری از روش‌های رده‌بندی توسط محققین حوزه‌های یادگیری ماشین، تشخیص الگو و آمار پیشنهادشده‌اند. از جمله‌ی این روش‌ها می‌توان به روش‌های درخت تصمیم، نزدیک‌ترین همسایه و ماشین بردار پشتیبان اشاره کرد. در این خصوص می‌توان به پژوهش [۱۵] اشاره نمود. [۱۱] نیز کارایی تعدادی از رده‌بندهای شبکه‌ی بیزی را مقایسه نمودند؛ اما در خصوص داده‌های حاصل از اینترنت اشیاء می‌توان به کارهای [۳] اشاره نمود که در آن به اصول و مفاهیم اینترنت اشیاء پرداخته است. البته [۱۴] نیز یک مطالعه مروری بر مباحث مربوط به اینترنت اشیاء انجام دادند. [۷] به مدل‌سازی وقایع پیچیده و داده‌های حاصل از اینترنت اشیاء پرداخته و نتایج معتبری را به دست آورده است. در زمینه داده‌های حاصل از ترافیک می‌توان به [۶] اشاره نمود. [۱۰] با استفاده از مفاهیم یادگیری شبکه‌های بیزی به مطالعه داده‌های مربوط به رخدادهای تصادفی پرداخته‌شده است. در خصوص موضوع انرژی‌های جریانی کانتر و ال‌هالمیری در سال ۲۰۱۰ مطالعه‌ای را انجام داده‌اند. در خصوص توان پیشامدهای حاصل از وقایع لخام (۲۰۰۸) یک بررسی دقیق را انجام و نتایج معتبری را به دست آورده است. برای ترکیب مدل‌های بیزی و داده‌های حاصل از اینترنت اشیاء [۱۶] مطالعه دقیقی را ارائه داده‌اند.

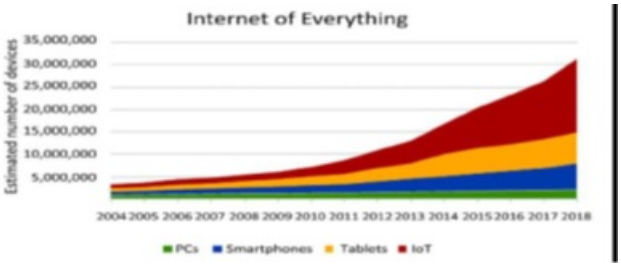
۳ پردازش جریانی در اینترنت اشیاء

یک سیستم اطلاعاتی امروزی باید قابلیت پردازش حجم بالای داده و همچنین استخراج اطلاعاتی که وابسته به محدودیت زمانی از مجموعه داده هستند را داشته باشد. روش پردازش جریانی به‌طور پیوسته داده مرحله دوم را بافاصله مشخص بعد از تولید داده اولیه وارد فرآیند پردازش می‌کند. در این مدل نه‌تنها تجمیع داده‌ها مهم است بلکه تحلیل آن‌ها با توجه به مدت‌زمانی مشخص نیز باید صورت پذیرد. پردازش جریانی به‌عنوان یک بخش اساسی در کاربردهای اینترنت اشیاء مطرح می‌شود که باید قابلیت‌هایی نظیر مقیاس‌پذیری، قابلیت دسترسی بالا، تحمل‌پذیری خطا برای مدیریت داده با حجم زیاد که به‌صورت دائم تولید می‌شود را داشته باشد. به‌عنوان یک طرح کلی از جایگاه این مدل، شکل ۳ اهمیت آن را در معماری اینترنت اشیاء نشان می‌دهد. پردازش جریانی یک جریان دنباله‌ای از داده است که توسط زمان مرتب‌شده است. این روش یک الگوی پردازش داده است که جهت رسیدن به کمترین تأخیر پردازش عمل می‌کند. در زیر به برخی از ویژگی‌های مجموعه داده جریانی می‌پردازیم.

الف) الزام ویژگی به‌موقع بودن نیازمند قابلیت گردآوری، انتقال، پردازش و ارائه مجموعه داده جریان به‌صورت بی‌درنگ است. توجه کنید که پردازش و ارتباطات در لحظه صورت می‌گیرد؛

ب) داده نویز یا معیوب تقریباً در هر سیستمی وجود دارد؛ بنابراین ویژگی تصادفی بودن در زنجیره دریافت و انتقال داده جریانی وجود دارد؛

ج) تا زمانی که منبع داده فعال است داده به سیستم پردازش جریانی ارسال می‌شود؛ بنابراین باید توجه داشت سیستم طراحی‌شده باید قابلیت پردازش حجم بی‌پایان از داده را دارا باشد. اکثر سیستم‌های پردازش جریانی، داده جریانی را پس از انتقال حذف می‌کنند؛ بنابراین داده ویژگی فرار بودن را دارا است و هر داده به‌عنوان داده جدید تلقی می‌شود.



شکل ۳. نمای پردازش جریانی یک جریان دنباله‌دار

۴ مدل‌سازی با استفاده از آمار بیزی

یک پردازشگر که وقایع را ثبت می‌نماید از مجموعه‌ای از شبکه‌های پردازشگر که خود از چندین فرستنده‌ی رخداد ساده تشکیل‌شده‌اند، ساخته‌شده است. اگر برای هر یک از این رخدادهای ساخته‌شده، هر جریان حالت (i,t) را با $f_{i,t}$ نمایش دهیم، $f_{i,t}$ به حالت گره‌های پیش از زمان t که آن را والدین حالت (i,t) می‌نامیم و با $Pa_{(i,t)}$ نشان می‌دهیم، بستگی دارد. مقادیر $f_{i,t}$، گره‌ها یا به عبارتی متغیرهای تصادفی موجود در شبکه‌ی بیزی هستند و وجود یال میان متغیرها، تعیین‌کننده‌ی وجود یا عدم وجود وابستگی می‌باشد. از رابطه‌ی هر متغیر با متغیرهای پیشین خود در شبکه‌ی بیزی، متغیرها دارای یک ساختار رتبه‌ای خواهند بود. مجموعه حالت‌های والد به شکل

$$F_{Pa_{(i,t)}} = \{f_{j,s} : (j,s) \in Pa_{(i,t)}\}$$

خواهد بود. مطابق با نظریه بیزی برای توزیع همه‌ی گره‌های موجود در شبکه‌ی بیزی که آن را F می‌نامیم، خواهیم داشت

$$P(F) = \prod_{i,t} P(f_{i,t}|F_{Pa_{(i,t)}})$$

که توزیع شرطی برای آن با استفاده از

$$P(f_{i,t}|F_{Pa_{(i,t)}}) = \frac{P(f_{i,t}, F_{Pa_{(i,t)}})}{F_{Pa_{(i,t)}}}$$

محاسبه می‌شود.

برای تعیین توزیع آماری داده‌های حاصل از اینترنت اشیاء و برآورد توزیع توأم احتمال در شبکه‌ی بیزی، احتمال $P(f_{i,t}, F_{Pa_{(i,t)}})$ را به‌وسیله‌ی مدل آمیزه‌ای نرمال یعنی ترکیب وزن‌داری از چندین توزیع نرمال، به‌صورت زیر مدل می‌نمایند. در تابع

$$P(f_{i,t}, F_{Pa_{(i,t)}}) = \sum_{m=1}^{M} a_m g_m \left(f_{i,t}, F_{Pa_{(i,t)}} | \mu_m, C_m\right)$$

M تعداد گره‌ها و $g_m \left(f_{i,t}, F_{Pa_{(i,t)}} | \mu_m, C_m\right)$ توزیع نرمال چندمتغیره‌ی برای گره mام با بردار مقادیر میانگین μ_m، $(N_p+1)\times 1$ بعدی و ماتریس کوواریانس C_m، $(N_p+1)\times(N_p+1)$ بعدی و a_m ضریب توزیع است. توزیع شرطی $P(f_{i,t}|F_{Pa_{(i,t)}})$ با استفاده از $P(f_{i,t}, F_{Pa_{(i,t)}})$ به دست می‌آید و برآورد $\hat{f}_{i,t}$ از طریق $F_{Pa_{(i,t)}}$ با استفاده از روش کم‌ترین توان‌های دوم خطا قابل‌محاسبه است. به‌طورکلی، برای نوشتن یک رابطه آماری در نظر بگیرید X یک متغیر تصادفی یا یک بردار تصادفی چندبعدی است، آنگاه مؤلفه‌های توزیع احتمالی مدل آمیخته‌ی نرمال به شکل زیر خواهد بود:

$$P(x|\Theta) = \sum_{l=1}^{M} a_l p_l(x|\theta_l)$$

که در آن $\Theta = (a_1,\cdots,a_M,\theta_1,\cdots,\theta_M)$ و M پارامترهای مدل هستند و $\sum_{l=1}^{M} a_l = 1$ است. هر $P(\cdot)$ یک توزیع احتمال پارامتری شده به‌وسیله‌ی $\theta_l = (\mu_l,\Sigma_l), l = 1,\cdots,M$ است. برای برآورد پارامترهای مدل آمیخته نرمال روش‌های معمول پاسخگو نیست لذا از الگوریتم امید ریاضی بیشینه‌سازی استفاده می‌کنیم.

۵ الگوریتم EM در برآورد پارامترهای شبکه‌های بیزی

الگوریتم امیدریاضی-بیشینه‌سازی (EM)، یک روش تکراری برای به دست آوردن بیشینه درستنمایی است و نواقص روش‌های سنتی به دست آوردن درستنمایی را که اغلب به فرم بسته‌ی توزیع نیاز دارند، جبران می‌نماید. اجرای الگوریتم امیدریاضی- بیشینه‌سازی (EM) اساساً شامل دو حالت است.

الف) اول زمانی است که داده‌ها به دلیل مشکلاتی چون محدودیت در فرآیند مشاهده، دارای پارامترهای پنهان هستند.

ب) دیگری زمانی اتفاق می‌افتد که بهینه‌سازی تابع درستنمایی به جهت تحلیلی دشوار است اما با فرض وجود ارزش برای متغیرهای اضافی به‌جز متغیرهای پنهان می‌توان آن را ساده‌سازی کرد که البته مورد دوم بیشتر اتفاق می‌افتد.

با استفاده از الگوریتم EM می‌توان پارامترهای $\{a_m,\mu_m,C_m\}_{m=1}^{M}$ را از داده‌های پیشین برآورد کرد. الگوریتم امیدریاضی-بیشینه‌سازی، چهارچوبی است که احتمال را بیشینه می‌نماید و یا به عبارتی برآورد پسین پارامترها در یک مدل آماری را بیشینه می‌نماید. در ارتباط با مدل‌سازی مبتنی بر احتمال، الگوریتم امیدریاضی-بیشینه‌سازی EM کار را با یک مجموعه اولیه از پارامترها شروع کرده به تکرار ادامه می‌دهد تا این‌که به مدل بهینه دست یابد؛ یعنی اگر مقصود را ارائه‌ی مدل برای هر رده در نظر بگیریم، الگوریتم تا زمانی که رده همگرا شود یا تغییرات کوچک باشند (کمتر از آستانه‌ی مشخص‌شده)، ادامه می‌یابد. فرض کنید M، مجموعه داده‌های مشاهده‌شده از یک توزیع باشد که مجموعه‌ی داده‌ی کامل برای آن به‌صورت $Z = (X,Y)$ خواهد بود و توزیع توأم متغیرها یا بردار تصادفی $Z = (X,Y)$ شکل زیر را داراست:

$$p(z|\Theta) = p(y|x,\Theta)p(x|\Theta)$$

که در آن Y مجموعه داده‌های پنهان و Θ مجموعه پارامترهای موجود در توزیع $p(z|\Theta)$ است. رابطه‌ی فوق اغلب از تابع توزیع حاشیه‌ای $p(x|\Theta)$ و فرض متغیرهای پنهان و برآورد پارامترها به دست می‌آید. در موارد دیگر مانند مقادیر گم‌شده در نمونه‌های یک توزیع، باید یک رابطه‌ی توأم میان مقادیر گم‌شده و مقادیر مشاهده‌شده فرض کرد؛ بنابراین، برای تابع توزیع جدید، یک تابع درستنمایی تعریف می‌شود

$$l(\Theta|z) = l(\Theta|x,y)p(x,y|\Theta)$$

که درستنمایی داده‌های کامل نامیده می‌شود. توجه کنید که این تابع در حقیقت یک متغیر تصادفی است زیرا مقادیر گم‌شده‌ی y، غیرمعلوم و تصادفی هستند؛ بنابراین، می‌توان به‌جای $l(\Theta|x,y) = h_{x,\Theta}(y)$ برای بعضی توابع $h_{x,\Theta}(\cdot)$ در نظر گرفت که در آن x و Θ مقدار ثابت و y یک متغیر تصادفی است. نسخه‌ی اصلی درستنمایی $l(\Theta|x)$ تابع درستنمایی داده‌های ناکامل نامیده می‌شود.

گام E: الگوریتم EM مقدار مورد انتظار داده‌های کامل لگاریتم درستنمایی $\log p(X,Y|\Theta)$ را نسبت به داده‌های نامعلوم Y به‌شرط داده‌های معلوم X و برآورد پارامترهای حاضر پیدا می‌کند؛ بنابراین، تعریف می‌شود

$$Q(\Theta,\Theta^{(i-1)}) = E\left[\log p(X,Y|\Theta)|X,\Theta^{(i-1)}\right]$$

که در آن $\Theta^{(i-1)}$ها پارامترهای فعلی هستند که امید در نظر گرفته می‌شوند و Θها پارامترهای جدید هستند که آن‌ها را برای افزایش Q بهینه می‌سازند. نکته‌ی کلیدی برای فهم مطلب این است که x و $\Theta^{(i-1)}$ ثابت هستند. Θ یک متغیر نرمال و متغیر تصادفی y به‌وسیله‌ی توزیع

$$f(y|x,\Theta^{(i-1)})$$

به شکل تابع درمی‌آید. سمت راست معادله می‌تواند به شکل زیر بازنویسی شود

$$E\left[\log p(X,Y|\Theta)|X,\Theta^{(i-1)}\right] = \int_{y\in Y} \log p(X,Y|\Theta) f(y|x,\Theta^{(i-1)})dy$$

توجه کنید که $f(y|x,\Theta^{(i-1)})$ توزیع حاشیه‌ای داده‌های مشاهده نشده است و به هردوی داده‌های مشاهده‌شده‌ی x و y و پارامترهای فعلی وابسته است و Y فضای مقادیر y است. در بهترین حالت‌ها، این توزیع حاشیه‌ای، یک بیان تحلیلی ساده برای پارامترهای $\Theta^{(i-1)}$ و شاید داده‌هاست. در بدترین حالت‌ها، این توزیع به‌سختی قابل به دست آوردن است. گاهی در حقیقت چگالی زیر استفاده می‌شود

$$f(y,x|\Theta^{(i-1)}) = f(y|x,\Theta^{(i-1)})f(x|\Theta^{(i-1)})$$

اما این روی مراحل بعدی تأثیر نمی‌گذارد زیرا عامل اضافه و $f_x(x|\Theta^{(i-1)})$ به Θ وابسته نیستند. به‌عنوان یک قیاس، فرض کنید که یک تابع $h(\cdot,\cdot)$ با دو متغیر داریم. $h(\theta,Y)$ را در نظر بگیرید که θ یک ثابت و Y یک متغیر تصادفی است که از توزیع $f_Y(y)$ به‌دست‌آمده است. پس

$$E[h(\theta,Y)] = \int_y h(\theta,Y)f_Y(y)dy$$

در مرحله اول از الگوریتم محاسبه امید تابع موردنظر است که با $Q(\Theta,\Theta')$ نمایش داده می‌شود. مرحله‌ی دوم از الگوریتم EM، بیشینه ساختن امیدی است که در مرحله‌ی اول ساخته‌شده و داریم

$$\Theta^{(i)} = arg\max_{\Theta} Q(\Theta,\Theta^{(i-1)})$$

به تعداد لازم می‌توان الگوریتم را تکرار نمود که بعد از هر تکرار میزان درستنمایی افزایش خواهد یافت و رفتار الگوریتم چنان است که همگرایی آن به یک بیشینه‌ی نسبی خواهد بود. برای این منظور، یک مدل احتمالاتی را در نظر گرفته و عبارت درستنمایی داده‌های ناکامل برای این چگالی حاصل از داده‌های x به‌وسیله‌ی

$$\log(l(\Theta|x)) = \log\prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N}\log\left(\sum_{j=1}^{M} a_j p_j(x_j|\theta_j)\right) \quad (1)$$

ارائه خواهد شد. با توجه به رابطه (۱) بهینه‌سازی آن دشوار خواهد بود زیرا شامل لگاریتم مجموع است. اگر x را به‌عنوان داده‌ی ناکامل در نظر بگیریم در آن صورت حضور اجزای داده‌های پنهان $y = \{y_i\}_{i=1}^{N}$ برای اطلاع از آن است که کدام چگالی هر جزء داده‌ها را تولید کرده است. به همین دلیل فرض می‌شود برای هر $y_i \in 1,\cdots,M$ و اگر iامین نمونه به‌وسیله‌ی kامین جزء ترکیبی تولیدشده باشد، $y_i = k$. لذا تابع درست نمایی دوباره به شکل زیر نوشته می‌شود.

$$\begin{aligned}\log l(\theta|x) &= \log\left(p(x,y|\theta)\right)\\ &= \sum_{i=1}^{N}\log\left(p\left(x_i|y_i(p(x,y|\theta))\right)\right)\\ &= \sum_{i=1}^{N}\log\left(\sum_{i=1}^{M} a_{y_i}p_{y_i}(x_i|\theta_{y_i})\right)\end{aligned}$$

و می‌توان آن را بهینه نمود. برای این منظور ابتدا باید عبارتی برای توزیع داده‌های مشاهده نشده در نظر گرفت. با فرض آن‌که پارامترهای موجود در تابع بردار $\Theta^g = (a_1^g,\cdots,a_M^g,\theta_1^g,\cdots,\theta_M^g)$ پارامترهای مناسبی برای درستنمایی $l(\Theta^g|x,y)$ هستند، لذا تحت Θ^g، $p_j(x_i|\theta_j^g)$ به‌آسانی برای هر i و j محاسبه می‌شوند. علاوه بر پارامترهای ترکیب، a_j می‌تواند به‌عنوان احتمال پیشین هر جزء ترکیب پنداشته شود که

$$a_j = p(C_j).$$

بنابراین، با استفاده از قانون بیزی می‌توان محاسبه نمود که

$$p(y_i|x_i,\Theta^g) = \frac{a_{y_i}^g p_{y_i}(x_i|\theta_{y_i}^g)}{p(x_i|\Theta^g)} = \frac{a_{y_i}^g p_{y_i}(x_i|\theta_{y_i}^g)}{\sum_{k=1}^{M} a_k^g p_k(x_i|\theta_k^g)}$$

و

$$p(y|x,\Theta^g) = \prod_{i=1}^{N} p(y_i|x_i,\Theta^g)$$

که در آن $y = (y_1,\cdots,y_N)$ یک نمونه از داده‌های مشاهده نشده‌ی مستقل است. حالا با نگاهی به معادله‌ی فوق دیده می‌شود که در این مورد چگالی حاشیه‌ای موردنظر به‌وسیله‌ی فرض وجود متغیرهای پنهان و حدس پارامترهای اولیه برای توزیع آن‌ها، به‌دست‌آمده‌اند. در این مورد، معادله‌ی بالا به شکل زیر درمی‌آید

$$\begin{aligned}Q(\Theta,\Theta^{(g)}) &= \sum_{y\in Y}\log\left(l(\Theta|X,Y)\right)p(y|x,\Theta^g)\\ &= \sum_{y\in Y}\sum_{i=1}^{N}\log\left(a_{y_i}p_{y_i}(x_i|\theta_{y_i})\right)\prod_{j=1}^{N}p(y_j|x_j,\Theta^g)\end{aligned}$$

و با اندکی محاسبه و برای $l \in 1,\cdots,M$، خواهیم داشت

$$\begin{aligned}&\sum_{y_1=1}^{M}\sum_{y_2=1}^{N}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p(y_j|x_j,\Theta^g) =\\ &\left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N}p(y_j|x_j,\Theta^g)\right)p(l|x_i,\Theta^g)\\ &= \prod_{j=1,j\neq i}^{N}\left(\sum_{y_i=1}^{M}p(y_j|x_j,\Theta^g)p(l|x_j,\Theta^g)\right) = p(l|x_j,\Theta^g)\end{aligned}$$

به دلیل این‌که $\sum_{i=1}^{M} p(i|x_j,\Theta^g) = 1$ می‌توان نوشت

$$\begin{aligned}Q(\Theta,\Theta^{(g)}) &= \sum_{l=1}^{M}\sum_{i=1}^{N}\log\left(a_l p_l(x_i|\theta_l)\right)p(l|x_i,\Theta^g)\\ &= \sum_{l=1}^{M}\sum_{i=1}^{N}\log(a_l)p_(l|x_i,\Theta^g)\\ &+ \sum_{l=1}^{M}\sum_{i=1}^{N}\log\left(p_l(x_i|\theta_l)\right)p(l|x_i,\Theta^g)\end{aligned}$$

گام M: در مرحله دوم از الگوریتم برای بیشینه ساختن عبارت، می‌توان جمله‌ی شامل a_l و جمله‌ی شامل θ_1 را به‌طور مستقل بیشینه ساخت، زیرا آن‌ها به هم وابسته نیستند. برای پیدا کردن عبارت a_l، ضریب لاگرانژ λ با محدودیت $\sum_l a_l = 1$ معرفی می‌شود. پس

$$\frac{\partial}{\partial a_l}\left[\log(a_l)p_(l|x_i,\Theta^g) + \lambda\left(\sum_l a_l - l\right)\right] = 0$$

و

$$\sum_{i=1}^{N}\frac{1}{a_l}p_(l|x_i,\Theta^g) + \lambda = 0$$

با جمع هر دو اندازه روی l، $\lambda = -N$ در نتیجه

$$a_l = \frac{1}{N}\sum_{i=1}^{N}p(l|x_i,\Theta_g)$$

برای بعضی توزیع‌ها، به دست آوردن عبارت تحلیلی برای θ_l به‌عنوان تابع‌هایی از هر چیز دیگری، ممکن است.

اکنون با توجه به اینکه هرکدام از اجزاء به‌عنوان یک متغیر تصادفی چند متغیری پیوسته است و می‌تواند به‌صورت بزرگ نمونه‌ای دارای توزیع نرمال d

متغیری باشد و ازآنجایی‌که آمیزه‌ای از توزیع‌ها در نظر گرفته‌شده‌اند، برآورد پارامترهای مدل عبارت است از

$$a_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g)$$
$$\mu_l^{new} = \mu_l = \frac{\sum_{i=1}^{N} x_i p(l|x_i,\Theta^g)}{\sum_{i=1}^{N} p(l|x_i,\Theta^g)}$$
$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)\left(x_i-\mu_l^{new}\right)\left(x_i-\mu_l^{new}\right)^T}{\sum_{i=1}^{N} p\left(l|x_i,\Theta^g\right)} \quad (۲)$$

که در آن $l = 1,\cdots,M$ و N اندازه‌ی مجموعه داده‌ها است؛ اما در الگوریتم EM، M یک پارامتر از پیش تعیین‌شده است و الگوریتم ممکن است به بیشینه‌ی محلی و یا مرز فضای پارامتر همگرا شود.

۶ پیش‌بینی بیزی برای آمیزه‌ای متناهی از مدل‌ها

هدف اصلی استنباط در شبکه‌های بیزی، برآورد مقادیر هدف گره‌ها به‌شرط مقادیر مشاهده‌شده‌ی گره‌ها است. فرض نمایید زوج (E,F) یک افراز از گره‌های شبکه‌ی بیزی به زیرمجموعه‌های گسسته باشد و (x_E,x_F) افراز متناظر متغیرها یا بردارهای تصادفی باشد، احتمال حاشیه‌ای می‌تواند به شکل زیر فرمول‌بندی شود

$$p(x_E) = \sum_{x_F} p(x_E,x_F)$$

بنابراین مطابق با نظریه بیزی، احتمال شرطی برابر است با

$$p(x_F|x_E) = \frac{p(x_F,x_E)}{p(x_E)} = \frac{p(x_F,x_E)}{\sum_{x_F} p(x_E,x_F)}$$

که می‌تواند برای هر x_F محاسبه شود. تحت قاعده‌ی کم‌ترین توان‌های دوم خطا، برآورد بهینه‌ی x_F برابر خواهد شد با

$$\hat{x_F} = E(x_F|x_E)$$

برای به دست آوردن پیش‌بینی بهینه‌ی $\hat{x_F}$ تحت متغیرهای پیوسته چند متغیری پیش‌بینی بهینه برای $\hat{x_F}$ تحت معیار میانگین حداقل توان‌های دوم خطا، برابر است با

$$\hat{x} = \sum_{l=1}^{M} \beta\mu_{lF|E} \quad (۳)$$

۷ ملاک ارزیابی با استفاده از خوشه‌بندی آمار بیزی

یکی از شاخه‌های یادگیری خودکار و ناراهنمایی ده، خوشه‌بندی است که در حین آن نمونه‌های مشابه در یک خوشه قرار می‌گیرند. پس مطابق با تعاریف متعارف در نمونه‌گیری آماری، اشیای موجود در یک خوشه از نمونه‌های با بیشترین شباهت و اشیای موجود در خوشه‌های متفاوت بیشترین تفاوت‌ها را خواهند داشت. برای اندازه‌گیری این شباهت‌ها که مطلوب، بیشینه بودن آن در بین خوشه‌ها و کمینه بودن آن در هر خوشه می‌باشد، معیاری نیاز است که این معیار اغلب فاصله در نظر گرفته می‌شود و به این نوع خوشه‌بندی، خوشه‌بندی مبتنی بر فاصله گویند.

اگرچه معیاری برای یافتن بهترین خوشه‌بندی وجود ندارد اما برای یافتن خوشه‌هایی مشابه از اشیاء در بین نمونه‌های ورودی، معیارهایی به جهت ارزیابی خوب بودن خوشه‌بندی همچون فشردگی ناحیه و مفهوم آنتروپی در نظریه اطلاع، استفاده‌شده است. از مهم‌ترین مسائل در ارزیابی همگنی، تعداد خوشه‌ها است که بسته به نوع الگوریتم می‌تواند از پیش تعیین‌شده یا متغیر باشد و باید در نظر داشت که گاهی ممکن است با افزایش تعداد خوشه‌ها، از بهینگی کاسته شود. یک روش برای ارزیابی مدل، ضمن ارائه‌ی الگوریتم خوشه‌بندی C میانگین خوشه‌بندی احتمالی است که تابع آن به‌صورت زیر است

$$J_m = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ij}^m ||x_i - c_j||^2 \quad (۴)$$

در عبارت (۴)، m یک عدد حقیقی بزرگ‌تر از ۱ است. C تعداد خوشه‌ها و N تعداد نمونه‌ها است. u_{ij} نیز میزان تعلق نمونه‌ی iام به خوشه‌ی jام را نشان می‌دهد. با برابر صفر قرار دادن مشتق تابع هدف خواهیم داشت

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(||x_i-c_j||^2/||x_i-c_k||^2\right)^{1/(m-1)}}$$

بر اساس ملاک (۴) می‌توان خوب بودن مدل را مورد ارزیابی قرار داد.

۸ مفهوم شباهت

برای تعیین شباهت دو رخداد از مدل احتمالی برای رخدادها استفاده می‌شود. ازآنجاکه ساختار رتبه‌ای در شبکه‌ی بیزی برقرار است، برای محاسبه‌ی شباهت دو گره برای ساختن مدل، یافتن والدین هر متغیر تصادفی لازم است. از مدل احتمالی گره‌ها به شکل زیر استفاده می‌شود

$$sim(C_i,C_j) := \frac{a_i^* a_j}{l_{ij}} \quad (۵)$$

که در آن، شباهت دو گره در شبکه‌ی بیزی، از ضرب وزن گره‌ها یعنی aها و تقسیم مقدار حاصل بر فاصله‌ی دو گره یعنی l_{ij} -که برابر با تعداد یال‌های در مسیر دو گره است- حاصل می‌شود. برای دو رخداد e_1 و e_2 مجموعه‌ای از حالات مختلف تحت عدم قطعیت، برابراست با

$$C_{e1} = (c_{e11},\cdots,c_{e1m}) \quad \text{و} \quad C_{e2} = (c_{e21},\cdots,c_{e2m})$$

برای هر $c_{e1i} \in C_{e1}$، یک c_{e2j} یافت می‌شود که بیشترین شباهت را داشته باشد، در واقع $\max(sim(c_{e1i}, c_{e2k}))$ به‌عنوان بیشترین شباهت بین C_{e1} و C_{e2} به شکل زیر محاسبه می‌شود

$$Q(C_{e1}, C_{e2}) = \frac{n}{k+l-n} \sum_{i=1}^{m} \beta(c_{e1i}) sim(c_{e1i}, c_{e2j})$$

به همین ترتیب شباهت بین C_{e2} و C_{e1} به شکل زیر محاسبه می‌شود

$$Q(C_{e2}, C_{e1}) = \frac{n}{k+l-n} \sum_{i=1}^{n} \beta(c_{e2i}) sim(c_{e2i}, c_{e1j})$$

که در آن‌ها، β وزن مشاهدات است؛ یعنی $\beta(c_{e1i})$ وزن مشاهدات در زمینه‌ی C_{e1} و $\beta(c_{e2i})$ وزن مشاهدات در زمینه‌ی C_{e2} نسبت به کل مشاهدات را نشان می‌دهد. در نتیجه شباهت بین C_{e1} و C_{e2} از میانگین دو رابطه‌ی فوق به‌صورت زیر به دست می‌آید

$$sim(C_{e1i}, C_{e2k}) = \frac{Q(C_{e1}, C_{e2}) + Q(C_{e2}, C_{e1})}{2}$$

برای ارزیابی فشردگی خوشه‌ها، ابتدا احتمال تعلق c_{ei} به خوشه‌ی C_h با استفاده از رابطه‌ی زیر به دست می‌آید

$$\hat{P}(C_h|ce_i) = \frac{sim(c_{ei}, C_h)}{\sum_j sim(c_{ei}, C_j)}$$

اگر c_{ei} که مجموعه حالت‌های ممکن برای والدین هر رخداد است، در خوشه‌بندی اصلی در خوشه قرار بگیرد، آنگاه آنتروپی نرمال‌شده با استفاده از رابطه‌ی زیر قابل‌محاسبه است

$$H_{norm}(ce_i) = \frac{-\sum_{h=1}^{k} \hat{P}(C_h|ce_i) \log_2^{\hat{P}(C_h|ce_i)}}{\log_2^k}$$

با استفاده از معنای آنتروپی در نظریه اطلاع، هرچقدر مقدار $H_{norm}(ce_i)$ به صفر نزدیک‌تر باشد، مقدار فشردگی بهتر است؛ بنابراین $H_{norm}(ce_i)$ برای همه‌ی c_{ei}ها محاسبه می‌شود و بهترین خوشه‌بندی را که خوشه‌بندی با کمترین مقدار $H_{norm}(ce_i)$ است به دست می‌آید.

در این مطالعه برای اندازه‌گیری میزان تشابه یا تمایز دو بردار تصادفی $(\overrightarrow{x})$ و $(\overrightarrow{y})$ با توزیع یکسان و با ماتریس واریانس کوواریانس S که به‌صورت ذیل تعریف می‌شود و به فاصله ماهالانویس شهرت دارد.

$$d(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(\overrightarrow{x} - \overrightarrow{y})^T S^{-1} (\overrightarrow{x} - \overrightarrow{y})}$$

می‌توان استفاده نمود. لازم به ذکر است درصورتی‌که ماتریس واریانس کوواریانس S یک ماتریس همانی باشد، فاصله ماهالانوبیس به فاصله‌ی اقلیدسی تبدیل می‌شود.

۹ مدل متوسط‌گیری بیزی

برای آن‌که بتوان برآورد بهتری را به شیوه بیزی به دست آورد و این‌که از تمام اطلاعات پیشین به درستی استفاده نمود، رویکرد متوسط‌گیری بیزی موردتوجه قرارگرفته است. فرض کنید Q کمیت موردنظر باشد، آنگاه می‌توان برای توزیع‌های پسین H مدل مختلف تولیدشده از فرآیندهای پردازش رخدادها، خوشه‌بندی و شبکه‌های بیزی و با استفاده از متوسط‌گیری بیزی با شرط معلوم بودن داده‌های D، نوشت.

$$P(Q|D) = \sum_{k=1}^{H} P(Q|M_k, D) P(M_k|D) \qquad (۶)$$

که در آن $P(Q|M_k, D)$ توزیع پسین Q تحت مدل M_k با داده‌های مفروض و $P(M_k|D)$ احتمال پسین مدل یا وزن مدل نامیده می‌شود. با توجه به اینکه رابطه (۶) دارای فرم بسته نخواهد شد لذا از الگوریتم زنجیر مارکوف مونت‌کارلو استفاده خواهد شد. با فرض اینکه مدل

$$P(\Theta_k|D, M_k) = \frac{P(D|\Theta_k, M_k) P(\Theta_k|M_k)}{P(D|M_k)}$$

که در آن $P(D|M_k)$ تابع درستنمایی حاشیه‌ای بوده و از رابطه‌ی زیر به دست می‌آید

$$P(D|M_k) = \int P(D|\theta_k, M_k) P(\theta_k|M_k) d\theta_k$$

توزیع پسین هر مدل یا وزن مدل‌های مختلف، با استفاده از

$$P(M_k|D) = \frac{P(D|M_k) P(M_k)}{P(D)}$$

قابل‌محاسبه است که در آن $P(M_k)$ توزیع پیشین مدل k است و اگر هیچ ارجحیتی برای هیچ‌یک از مدل‌ها وجود نداشته باشد، توزیع پیشین همه‌ی مدل‌ها یکسان فرض می‌شود؛ اما برای محاسبه‌ی $P(D|M_k)$ می‌توان از اعتبارسنجی متقابل استفاده نمود.

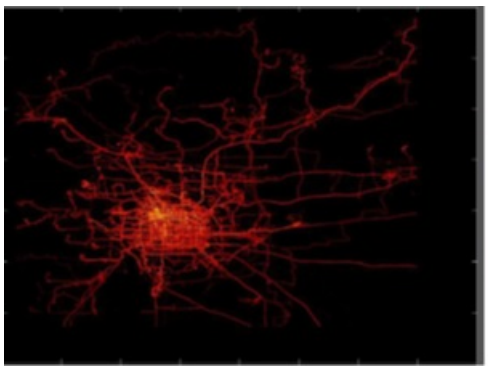
۱۰ کاربرد مدل‌های آماری بیزی برای اینترنت اشیاء

در این قسمت برای تفسیر نظریه‌های فوق به بررسی دو مثال واقعی پرداخته می‌شود

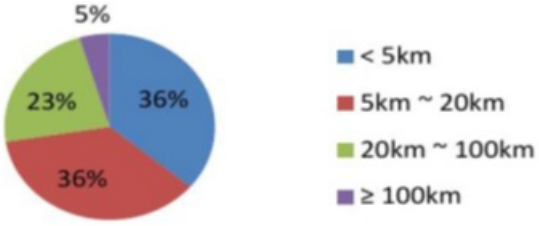
الف- مجموعه داده GPS

این مجموعه داده توسط گروه تحقیقاتی مایکروسافت در آسیا توسط ۱۸۲ کاربر در طول ۵ سال از آوریل ۲۰۰۷ تا آگوست ۲۰۱۲ به‌صورت ماهانه گردآوری‌شده است. این موضوع بر اساس مطالعات [۲۲، ؟] انجام‌شده است. یک مسیر GPS از این مجموعه داده‌ها به‌وسیله دنباله‌ای از نقاط که برچسب

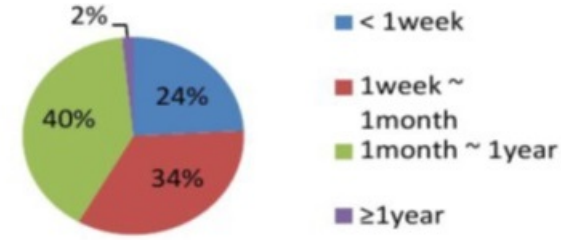
زمانی دارند به کار گرفته‌شده است و هر نقطه دارای سه مقدار عددی است که نشان‌دهنده‌ی طول جغرافیایی، عرض جغرافیایی و ارتفاع نسبت به سطح دریا است. این مجموعه داده شامل ۱۷۶۲۱ مسیر در کشور چین است که درمجموع فاصله‌ی ۱۲۹۲۹۵۱ کیلومتر داشته و در طول مدت ۵۰۱۷۶ ثانیه گرفته‌شده است. این مسیرها توسط دستگاه‌هایی که مجهز به دستگاه GPS بوده‌اند مانند تلفن همراه عکس‌برداری شده و به‌عنوان یک نمونه آماری در نظر گرفته‌شده‌اند. بررسی اولیه انجام‌شده اعلام می‌دارد که ۹۱/۵ درصد از داده‌ها دارای ساختار هستند و می‌توان برای آن‌ها یک چگالی در نظر گرفت و در ساختار یک چگالی نمایش داد. به‌عنوان‌مثال برای هر ۱ الی ۵ ثانیه یا هر ۵ الی ۱۰ متر در هر نقطه می‌توان یک نمونه در نظر گرفت. این مجموعه داده که از کاربران گرفته‌شده است علاوه بر این‌که شامل فعالیت‌های روزانه مانند رفتن به خانه، محل کار و موارد این‌چنینی است همچنین شامل فعالیت‌های ورزشی و تفریحی و نیز خرید کردن، رستوران رفتن، دیدن مناظر جالب، دوچرخه‌سواری و پیاده‌روی را شامل می‌شود. این مجموعه از داده‌ها می‌توانند در شاخه‌های تحقیقاتی نظیر تشخیص فعالیت کاربران، تشخیص الگوی متحرک، شبکه اجتماعی بر مبنای مکان، امنیت و حریم خصوصی مکانی و سیستم‌های توصیه‌گر مکانی مورداستفاده قرار گیرند. اگرچه این مجموعه داده در بیش از ۳۰ شهر در چین گردآوری‌شده است در کشورهای اروپایی و کشور آمریکا نیز در برخی از شهرهای آن داده گردآوری‌شده است که بیشترین نمونه از شهر بیجینگ در چین است که در اختیار ما قرارگرفته است. مطابق شکل ۴ می‌توان وضعیت داده‌ها را در نقشه این شهر به‌صورت نقشه حرارتی مشاهده کرد. شکل ۴ نشان‌دهنده توزیع مسیرها بر اساس فاصله است. شکل ۵ مدت‌زمان مؤثر مسیرها بوده، شکل ۶ توزیع زمانی کاربران که داده را گردآوری کرده‌اند و شکل ۷ توزیع کاربران در مسیرها است.



شکل ۴. نمایی از نقشه حرارتی شهر بیجینگ در چین بر مبنای اینترنت اشیاء



شکل ۵. مدت‌زمان مؤثر مسیرها



شکل ۶. توزیع زمانی کاربران

شکل ۷. توزیع کاربران در مسیرها

پس از انجام نمونه‌برداری مشخص شد کاربران داده‌ها را برحسب نوع حمل‌ونقلی که انجام می‌دهند، برچسب‌گذاری کرده‌اند که شامل حرکت به‌وسیله‌ی رانندگی کردن، به‌وسیله اتوبوس، دوچرخه‌سواری و پیاده‌روی می‌شود. نتایج این برچسب‌گذاری در جدول ۱ ارائه‌شده است. هر پوشه در مجموعه داده‌ها شامل فایل‌های GPS است که به‌صورت یک رشته اعداد تبدیل‌شده است. هر داده که به‌صورت رشته تبدیل‌شده است شامل یک مسیر بوده و نام آن بر اساس زمان گذاشته‌شده است. هر خط شامل موارد زیر است: عرض جغرافیایی، طول جغرافیایی، ارتفاع که از ۷۷۷− کمتر نامعتبر است، تاریخ بر اساس روزها و قسمت اعشاری که از تاریخ ۱۸۹۹/۱۲/۳۰ محاسبه‌شده است، تاریخ به‌صورت رشته و زمان به‌صورت رشته. به‌عنوان‌مثال:

۳۹.۹۰۶۶۳۱, ۱۱۶.۳۸۵۵۶۴, ۰, ۴۹۲,۴۰۰۹۷ · ۵۸۶۴۵۸۳۳۳, ۲۰۰۹−۱۰−۱۱, ۱۴:۰۴:۳۰

جدول ۱. توزیع تنوع کاربران در اینترنت اشیاء

نوع ترابری	فاصله برحسب کیلومتر	مدت‌زمان برحسب ساعت
پیاده‌روی	۱۰۱۲۳	۵۴۶۰
دوچرخه‌سواری	۶۴۹۵	۲۴۱۰
اتوبوس	۲۰۲۸۱	۱۵۰۷
تاکسی/خودروی شخصی	۳۲۸۰۶	۲۳۸۶
قطار	۳۶۲۵۳	۷۴۵
هواپیما	۲۴۷۸۹	۴۰
موارد دیگر	۹۴۹۳	۴۰۴

تعداد کاربران	۱۷۸
تعداد مسیرها	۱۷۶۲۱
تعداد نقاط	۲۳۶۶۷۸۲۸
مجموع مسیرها	۱۲۵۱۶۵۴ کیلومتر
مجموع مدت‌زمان	۴۸۲۰۳ ساعت
روزهای مؤثر	۱۰۴۱۳

در این مطالعه می‌خواهیم با استفاده از مدل‌های آماری پیشرفته و به‌وسیله اینترنت اشیاء مکانی که بیشترین علاقه‌مندی را برای افراد ایجاد می‌نماید تعیین نموده و m مسیری که بیشترین تردد و بازدید از دنباله‌ها را داشته است به دست آوریم. ابتدا تاریخچه رفت‌وآمد را بر اساس مدل‌سازی آماری دنباله مسیرها و مکان‌ها مدل کرده و بر این ساختار مدل استنتاجی برای هدف موردنظر اعمال می‌شود. روش موردنظر بر روی ۱۰۷ کاربر در بازه زمانی یک سال بر اساس داده GPS به‌دست‌آمده اعمال می‌شود. دو معیار مهم در این روش نقاط ثابت و تاریخچه مکانی است که به ترتیب ناحیه‌ای است که کاربر در یک‌زمان مشخص در یک ناحیه به گشت‌زنی پرداخته و تاریخچه مکانی دنباله‌ای از نقاط ثابت که بر اساس زمان‌های ورود و خروج نقاط ثابت ثبت‌شده‌اند. ساختار داده‌ها دو مرحله برای ساخت دارد.

الف) ایجاد یک درخت و انطباق مدل آماری برای آن‌ها بر اساس شناسایی GPSها که معمولاً از روش‌هایی که بر اساس نوع خوشه‌بندی داده‌ها انجام‌شده و در ساختار سلسله مراتبی استفاده می‌شود.

ب) ایجاد و ساخت گراف آماری در هر سطح و مکان که بر اساس شرایط مکانی یال‌های بین رأس‌های آن سطح در درخت متصل می‌شود.

معماری سیستم پیشنهادی برای داده‌ها نیز از چهار قسمت تشکیل می‌شود:

- مدل‌سازی تاریخچه مکانی
- مکان‌های موردعلاقه و کاوش در دنباله‌ها
- قسمت دانش
- سیستم توصیه‌گر

در این پژوهش علاوه بر پیش‌بینی وضعیت سیستم بر اساس نقاط به دنبال تحلیل رفتارهای پیچیده‌تر مانند حالت حمل‌ونقل که شامل دوچرخه‌سواری، پیاده‌روی، سوارشدن به اتوبوس، مترو و موارد مشابه هستیم. مشخص کردن وضعیت کاربر در حالت حمل‌ونقل بسیار مشکل بوده و وابسته به برچسب‌گذاری کاربر است، حتی با اندازه‌گیری سرعت نیز نمی‌توان به‌طور مشخص از وضعیت کاربر اطلاع حاصل کرد. به دلیل آن‌که کاربر در محیطی قرار دارد که عواملی مانند ترافیک، شرایط آب و هوایی و وضعیت‌های دیگر وجود دارد که باید به‌طور تقریبی حالت حمل‌ونقل کاربر را بررسی کرد. در این پژوهش بر اساس یادگیری راهنمایی‌ده پیش‌بینی در مورد حالت حمل‌ونقل کاربر بر اساس مجموعه داده GPS برای ۶۵ نفر در طول بازه زمانی ده ماه زده می‌شود. در این پژوهش ابتدا به شناسایی مجموعه‌ای از ویژگی‌های بامعنی که در شرایط ترافیک می‌تواند کمک بسزایی در بهبود کارایی الگوریتم استنتاج کند می‌پردازیم؛ مدلی بر اساس گراف می‌سازیم که در مرحله‌ی پس از پردازش عملکرد الگوریتم استنتاج را بهبود می‌بخشد. معماری این موضوع در شکل ۸ ارائه‌شده است.

شکل ۸. الگوریتم استنتاج از مجموعه داده‌های آموزشی

در بخش بعدی به روش تحلیلی داده‌ها پرداخته می‌شود.

ب- **داده‌های ترافیک**

در رویکرد جدید، اشیاء از جمله خودروها در حال مجهز شدن به حسگرهایی باقدرت پردازش بالا هستند. از جمله کاربردهای اینترنت اشیاء در صنعت خودرو می‌توان به ارتقای کیفیت و کنترل خودرو توجه نمود. همچنین این تجهیزات به کاهش انتشار آلاینده‌های محیطی به‌خصوص آلاینده‌های الکترونیکی و نیز جلوگیری از انتشار آلاینده‌های محیط‌زیست کمک می‌کنند. در ادامه با شرح یک مسئله‌ی واقعی در مورد اینترنت اشیاء، یک مدل ترافیکی را مورد ارزیابی قرار می‌دهیم. با توجه به مطالب گفته‌شده در این مقاله استفاده از مدل‌های آماری می‌توان داده‌های حاصل از اینترنت اشیاء را مورد ارزیابی قرار داد تا امکان کنترل با استفاده از تکنولوژی‌های جدید در حوزه‌های کنترل زیست‌محیطی همچون کنترل ترافیک و کنترل آلودگی هوا و ... به وجود آید. در مسئله‌ی تحلیل داده‌های تولیدشده از اینترنت اشیاء، موضوعات تکنیکی همچون حجم داده‌ها، سرعت تولید آن‌ها، عدم قطعیت و ناتمامی رخدادهای موجود در داده‌های تولیدی محیط شهر، پیچیدگی‌ها و الزام‌هایی را به کار می‌افزاید. در این مطالعه داده‌هایی مورد ارزیابی قرار می‌گیرند که قبلاً در [۱] و [۲] مورد ارزیابی قرارگرفته و در یکی از مطالعات بین‌المللی در شهر آرهاس دانمارک، در پروژه‌ای به نام CityPluse باهدف کمک به مدیران شهری در ارائه‌ی خدمات به شهروندان به کار گرفته‌شده‌اند. در آن پروژه، شرایط برای گسترش کاربردهایی ایجادشده که به‌طور پیوسته، کاربران توانایی مشاهده‌ی آنچه در شهر در حال رخ دادن است و نحوه‌ی تأثیر آن بر شهروندان، توریست‌ها، شرکت‌ها و مدیریت شهری را دارند. ولی هدف استفاده از داده‌های ترافیک این مجموعه داده است و تنها از ۱۰۰۰ مشاهده‌ی آن که بررسی‌های صورت گرفته در یک هفته از یک کد شبکه‌ی پردازنده‌ی رخداد می‌باشد استفاده‌شده است. با به‌کارگیری مجموعه داده‌های ترافیک موجود در داده‌های شهر آرهاس یک مسئله‌ی اینترنت اشیاء را تا ارائه‌ی یک مدل نهایی برای پیش‌بینی ترافیک به کار گرفتیم. در این پژوهش فرض بر این است که متغیرهای موجود در این مجموعه داده، برای ۱۰۰۰ رکورد داده، عبارت از متوسط زمان ترافیک، میانه زمان ترافیک، تعداد وسیله‌ها و متوسط سرعت وسیله‌ها بر اساس شماره IDی گزارشگر موجود در شبکه‌ی پردازنده‌ی رخدادهای پیچیده و در هر ۵ دقیقه ضبط‌شده، هستند. مراحل پردازش داده‌های حاصل از اینترنت اشیاء برای انجام کار ۵ مرحله است که شامل بررسی اولیه داده‌ها، خوشه‌بندی داده‌ها، مدل‌سازی شبکه بیزی، تعیین مدل نهایی، پیش‌بینی ترافیک و ارزیابی و سنجش اعتبار مدل پیشنهادی است. در این مطالعه مشخص گردید داده‌های حاصل از اینترنت اشیاء به سه خوشه تبدیل می‌شوند که می‌توان برای هر خوشه مدل بیزی را مورداستفاده قرار داده و برای استفاده از یک ساختار همگنی، می‌توان روش متوسط‌گیری بیزی را در ایجاد مدل نهایی به کار برد. بررسی‌ها نشان می‌دهد که ارتباط بین متغیرها در هر سه خوشه مانند یکدیگر است و این موضوع باعث می‌شود تا انتخاب مدل متوسط‌گیری بیزی با اطمینان بیشتری مورداستفاده قرار گیرد. میزان مخاطره مورد انتظار به شیوه بیزی برای سه خوشه به ترتیب برای خوشه‌ی اول برابر است با ۱۰٫۹۱۸۸۸، برای خوشه‌ی دوم برابر ست با ۹٫۰۲۳۲۳۲ و برای خوشه‌ی سوم برابر است با ۱۲٫۷۸۱۹۱ است که با استفاده از مدل متوسط‌گیری بیزی این مقدار به ۱۲٫۱۹۰۴ می‌رسد. با توجه به محاسبات انجام‌شده برای تعیین مقدار احتمال‌های پسین بر مبنای متوسط‌گیری بیزی، دیده می‌شود که فرم بسته‌ای ایجاد نمی‌شود. لذا از روش‌های شبیه‌سازی استفاده می‌شود. درواقع با استفاده از زنجیر مارکوف، ابتدا یک دنباله از نمونه‌های مستقل $\theta_k^{(t)} : t = ۱,\cdots,T$، از θ_k از توزیع $P(\theta_k|M_k)$ به‌وسیله‌ی نمونه‌گیری از θ_k به دست می‌آید، سپس رابطه‌ی تقریب پیشنهادی به‌صورت زیر اعلام می‌شود

$$\hat{P}(y_i|M_k) = \frac{۱}{T}\sum_{t=۱}^{T} p(y_i|\theta^{(t)}k, M_k) \qquad (۷)$$

جدول ۲. برآورد پارامترها در مدل آماری بیزی بر مبنای متوسط‌گیری بیزی و با استفاده از تقریب (۷) برای خوشه سوم

متوسط زمان اندازه‌گیری شده	۱۰۶	۱۱۰	۱۱۸	۱۲۲	۱۲۸
احتمال‌های شرطی	۰٫۰۱۴۹۴۶۱۲۴	۰٫۲۷۷۷۱۹۸۴۷	۰٫۰۱۴۹۴۶۱۲۴	۰٫۰۱۴۹۴۶۱۲۴	۰٫۰۵۸۷۴۱۷۴۵
متوسط زمان اندازه‌گیری شده	۱۳۲	۱۳۵	۱۳۶	۱۳۷	۱۳۹
احتمال‌های شرطی	۰٫۰۵۱۴۴۲۴۷۵	۰٫۰۷۲۳۴۰۲۸۵	۰٫۰۲۲۲۴۵۳۹۵	۰٫۰۲۹۵۴۴۶۶۵	۰٫۰۲۲۲۴۵۳۹۵
متوسط زمان اندازه‌گیری شده	۱۴۰	۱۴۱	۱۴۲	۱۴۳	۱۴۴
احتمال‌های شرطی	۰٫۰۲۹۵۴۴۶۶۵	۰٫۰۵۸۷۴۱۷۴۵	۰٫۰۲۹۵۴۴۶۶۵	۰٫۰۶۶۰۴۱۰۱۵	۰٫۰۶۶۰۴۱۰۱۵
متوسط زمان اندازه‌گیری شده	۱۴۵	۱۴۶	۱۴۷	۱۴۸	۱۴۹
احتمال‌های شرطی	۰٫۰۲۲۲۴۵۲۹۵	۰٫۰۶۶۰۴۱۰۱۵	۰٫۰۲۲۲۴۵۳۹۵	۰٫۰۰۷۶۴۶۸۵۴	۰٫۳۶۸۴۳۹۳۵
متوسط زمان اندازه‌گیری شده	۱۵۰				
احتمال‌های شرطی	۰٫۰۱۲۹۴۶۱۲۴				

سایر موارد نیز می‌تواند به این طریق محاسبه و بر اساس آن شانس رخداد پیشامد موردنظر برای مدل ترافیکی پیشنهادی در کاربرد موردتوجه قرار گیرد. به‌عبارت‌دیگر اینترنت اشیاء می‌تواند وقوع ترافیک را بر اساس ثبت داده‌ها بر اساس متغیرهایی مانند افزایش متوسط زمان اندازه‌گیری شده، کاهش سرعت و یا افزایش مدت‌زمان اندازه‌گیری شده و تعداد وسیله‌ها در یک مکان، تشخیص دهد.

۱۱ بحث و نتیجه‌گیری

مطالعات انحام شده در این بخش نشان داد که استفاده از نظریه آمار بیزی می تواند بخوبی ضمن ارایه یک مدل تحلیلی مناسب برای اینتر نت اشیا نتایج معتبری را نیز به صورت کار بردی در حوزه های مختلف ارایه دهد که متناسب با ساختار جدید تولید داده ها در حالتهای غیر نظام مند می تواند عمل نماید. به شیوه رویکرد مدل یشنهادی در این مقاله می توان چار چوب مناسبی برای مفاهیمی مانند پیش بینی تحلیل و استخراج مقادیر پاسخ حاصل از ورود یهای به مدل در داده های حاصل از اینتر نت اشیا ایجاد نمود که در بسیاری از محیط های کسب و کار مورد نیاز است.

مراجع

[۱] اسکندری، فرزاد (۱۳۹۹). کاربرد مدل‌های آماری در ساختار اینترنت اشیاء. طرح پژوهشی اتمام یافته. دانشگاه علامه طباطبایی.

[۲] ارشدی، افروز (۱۳۹۷). تحلیل مدل‌های بیزی چند متغیری. پایان‌نامه کارشناسی ارشد. دانشگاه علامه طباطبایی.

[3] Amir Vahid Dastjerdi, R. B. (2016). *Internet of Things: Principles and Paradigms*. Elsevier, Todd Green.

[4] Baldi, P., S. P. W. D. (2014). *Searching for exotic particles in high-energy physics with deep learning*. Nature communications. 4308.

[5] Berthelsen, K. and J. Muller (2003). likelihood and non-parametric bayesian mcmc inference for spatial point processes based on perfect simulation and path sampling. *Scandinavian J. Statist*, **30**, 549–564.

[6] Castillo, E., Menéndez, J. M., and Sánchez-Cambronero, S. (2008). Predicting traffic flow using Bayesian networks. *Transportation Research Part B: Methodological*, **42**, 482–509.

[7] Chuanfei, X., Shukuan, L., Lei, W., and Jianzhong, Q. (2010). Complex event detection in probabilistic stream. *Web Conference (APWEB), 2010 12th International Asia-Pacific*, 361–363.

[8] Dietrich, D. and et al. (2014). *Data Science Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC Education Services, Canada, Wiley J.

[9] Dolev, S., Kopeetsky, M., and Shamir, A. (2011). RFID authentication efficient proactive information security within computational security. *Theory of Computing Systems*, **48**, 132–149.

[10] Etzion, O., Niblett, P., and Luckham, D. C. (2011). Event processing in action, Manning Greenwich. Heckerman, David, A tutorial on learning with Bayesian networks Learning in graphical models. *Springer*, 301–354.

[11] Heckerman, D.; Geiger, D. C. D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

[12] Kunz, T. and Alhalimi, R. (2010). Energy-efficient proactive routing in MANET: Energy metrics accuracy. *Ad Hoc Networks*, **8**, 755–766.

[13] Luckham, D. (2008), The power of events: An introduction to complex event processing in distributed enterprise systems. *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, 3–3.

[14] Madakam, Somayya and Ramaswamy, R and Tripathi, Siddharth (2015). Internet of Things (IoT): A literature review. *Journal of Computer and Communications*, **42**, 164.

[15] NIPS. Baik, S. Bala, J. (2004). A decision tree algorithm for distributed data mining: Towards network intrusion detection. *Lecture Notes in Computer Science*, **3046**, 206 – 212.

[16] Pascale, A. and Nicoli, M. (2011). Adaptive Bayesian network for traffic flow prediction, *Pearl, Judea Bayesian networks*, 177–180.

[17] Samaranayake, S., Blandin, S., and Bayen, A. (2011). Learning the dependency structure of highway networks for traffic forecast, *IEEE Conference on Decision and Control and European Control Conference*, 5983–5988.

[18] Schwaab, J. and Olthof, S. (2015). Internet of things, using sensors for good: How the internet of things can improve lives. *Deutsche Gesellschaft für Internationale Zusammenar beit (GIZ) GmbH (2015)*.

[19] Tawara, N., Ogawa, T., Watanabe, S., and Kobayashi, T. (2012). Fully Bayesian inference of multi-mixture Gaussian model and its evaluation using speaker clustering, Acoustics, Speech and Signal Processing (ICASSP). *IEEE International Conference*, 5253–5256.

[20] Wang, Y. and Kuang, L. (2015). A traffic prediction method based on complex event processing and adaptive Bayesian networks. *International Academic Research Conference, 3-6 August*, University of London.

[21] Wang, Y. and Zhang, X. (2014). A proactive parallel complex event processing method for large-scale intelligent transportation systems, *International Journal of Multimedia and Ubiquitous Engineering*, **9**, 111–122.

[22] Zhou, Y., Johansen, A. M., and Aston, J. A. (2012). Bayesian model comparison via path-sampling sequential Monte Carlo. *Statistical Signal Processing Workshop (SSP), IEEE*, 245–248.

[23] Zheng, Y., Li., Q., Chen, Y., Xie, X. and Ma, W.Y (2008). Understanding mobility based on GPS data. *In Proceedings of ACM conference on Ubiquitous Computing*, 312–321.

[24] Zheng, Y., Zhang, L., Xie, X. and Ma, W.Y (2008). Mining interesting locations and travel sequences from gps trajectories. *Proceedings of International conference on World Wild Web, ACM Press (2009)*, 791–800.