آموزش ماشین یادگیرنده برای تحلیل داده های توزیع شده بر اساس تبدیلات موجکی و داده کاوی تجمعی

حیدر مختاری فریور [۱]، حبیب ایزدخواه [۲]

[۱] پژوهشگر، دانشگاه تبریز ، تبریز ، mokhtari@pgu.ac.ir
[۲] دانشیار گروه علوم کامپیوتر ، دانشگاه تبریز ، تبریز ، izadkhah@tabrizu.ac.ir

چکیده

تحلیل انواع داده های توزیع شده غالبا متضمن تجمیع آنهاست. در اغلب اوقات اینکار بدلایل مختلف امکان پذیر نمی نماید و ضرورت وجود روشهای جایگزین برای تجمع نتایج تحلیل بجای تجمیع اصل داده ها اجتناب ناپذیر است. این مقاله روشی را جهت تحلیل محلی داده ها و تجمیع نتایج آنها داده ها بدون نیاز به تجمیع اصل آنهاست. اینکار با استفاده از برآورد تابعی برای یادگیری ماشین جهت تحلیل های محلی و تجمیع مرکزی آنها با استفاده از تبدیلات موجکی و داده کاوی تجمعی بصورت بلادرنگ ارئه می نماید. در این روش، سامانه های یادگیرنده و تحلیل گر خودمختار برای برآوردهای اطلاعاتی معرفی و توصیف میگردند

واژه های کلیدی

داده کاوی، داده های توزیع شده –داده کاوی تجمعی- ماشین های یادگیرنده .

مقدمه

تحلیل های اطلاعاتی صرف نظر از نوع و مقیاس و شکل و محتوای آن، بر پایه داده هایی از جنس و ابعاد مختلف و حتی پراکندگی و تنوع ساختاری انجام می گیرد. این داده ها در صورتی که بشکل موثر در کنار هم قرار نگیرند و تاثیر متقابل آنها بر همدیگر بدرستی سنجیده نشود، منبع مناسبی برای تحلیل های گوناگون و پیچیده و چند بعدی نخواهند بود. بلافاصله می توان نتیجه گرفت که تحلیل قوی نیازمند فراهم شدن یکپارچگی داده هاست به نحوی که امکان استخراج مدل های مختلف اطلاعاتی و تحلیلی میسر گردد.

اما تجمیع داده ها در شرایط واقعی گاه با سختی ها و دشواری های فراوانی همراه است. این دشواری ها منبعث از سیاستگذاری های دارندگان داده ها و یا هزینه های تجمیع آنهاست. در دنیای واقعی داده ها بندرت شکل یکپارچه داشته و مهیای عملیات تحلیل و تبیین مدلهای نهفته در آن نیست. داده ها در حاضر بشکل توزیع شده در منابع مختلف و نقاط مختلف انبارشی و نگهداری است. این نوع داده ها، داده های بزرگ نامیده میشوند.

گام نخست تعریف دقیق مسئله و سپس بررسی روشها و ابزار های مختلف برای یکپارچه سازی داده ها برای انجام عملیات تحلیلی است به نحوی که بتوان متد ها و روش های مورد بحث برای تجمیع و تحلیل را به الگوریتمهایی قابل استفاده در ماشینهای تحلیلگر و تصمیم گیرنده بصورت اتوماتیک در مقیاسهای بسیار بزرگ، تبدیل کرد و استفاده نمود [۲].

ساختارهای داده ای و روشهای تحلیل

داده های بزرگ[۱] اصطلاحی است که به مجموعه ای گسترده از داده ها که به اشکال مختلف و در محل های گوناگون موجود هستند، با حفظ بعضی ویژگی های مشترک اطلاق می شود. فعالیتهای مختلف در حوزه های گوناگون سبب تولید انبوهی از داده هایی میشود که گاه بدلیل عدم برخورداری از انسجام ساختاری کافی بلا استفاده بوده و گاه به دلیل نیاز به تجمیع آنها برای انجام تحلیل های مختلف متضمن هزینه های انتقال داده ای بسیار بالا و گاه نیز بدلیل محرمانگی و یا سیاستهای دارندگان آنها عملا امکان تجمیع و تحلیل یکپارچه آنها وجود ندارد .

بسیاری از داده های غیر متمرکز که توسط بخشهای مختلف در کشور تولید می شوند و نقش به سزایی در ارزیابی عملکردهای مختلف در حوزه های گوناگون را دارند فاقد ساختار متشکل بوده و عملا برای تحلیل های مفید در دسترس نیستند. مثالهایی در این زمینه برای فهم موضوع مفید خواهد بود .

۱. فرض کنیم محققینی علاقمند به یافتن وابستگی بیماری های خاص با الگوهای آب وهوایی کشور باشند. این محققین به پایگاه داده ای بزرگی از بیماریهای خاص در مرکز کنترل بیماریها وهمچنین به پایگاه داده ای محیط واقلیم در سازمان هواشناسی

[۱] Big Data

دسترسی دارند. اما این دو پایگاه داده ای در دو محل متفاوت قرار دارند وبرای تحلیل داده ای با نرم افزارهای متداول ومعمولی تحلیل داده ای، نیازمند تجمیع هر دو پایگاه در یک محل منفرد می باشد که ممکن است غیر عملی باشد.

۲. سازمان عمده مالی (مانند بانکها، شرکت های بیمه، صادر کنندگان کارت های اعتباری و ...) برای ممانعت از تقلب وجعل در سیستمهای همدیگر نیازمند همکاری می باشند. آنها ناگزیر از اشتراک گذاری وتسهیم الگوهای داده ای مربوط به جعل وتقلب هستند، اما علاقمند به تسهیم واشتراک گذاری اطلاعات حساس نیستند. بنابراین ترکیب پایگاه های داده ای آنها با یکدیگر عملا امکان پذیر نیست. روشهای های داده کاوی موجود، راه حلی برای چنین وضعیتی ارائه نمی کنند.

۳. سازمان های دفاعی/ امنیتی وظیفهٔ نظارت ومونیتورینگ بعضی از موارد را بعهده دارند. چندین سامانه حسگر در حال نظارت وجمع آوری داده ها از وضعیت موجود هستند. تحلیل وآنالیز سریع اطلاعات وارده وواکنش وپاسخ فوری در این وضعیت ضروری است. جمع آوری تمام اطلاعات در یک سایت مرکزی وتحلیل آنها نیازمند مدت زمان زیادی بوده واین راه حل ورویکرد برای سیستمهای مدرن نظارتی با تعداد زیادی حسگر ووظایف واکنش سریع نسبت به عوامل بیرونی، قابل استفاده ومقیاس پذیر نیست .

٤. شرکت های بزرگ برای توسعه موفق استراتژی تجاری خود، نیازمند تحلیل رکوردهای تراکنشی مشتریان خود می باشند. این داده ها در هزاران مرکز استقرار پخش شده اند وجمع آوری وانبارش کل این اطلاعات در یک سایت مرکزی برای تحلیل وداده کاوی با سامانه های تجاری داده کاوی مستلزم زمان زیادی بوده وبطول می انجامد .

۵. نهادهای نظارتی و امنیتی در جهت نظارت و تشخیص تخلفات مالی اعم از پولشویی، اختلاس و غیره نیازمند دسترسی همزمان به داده های دستگاههای مختلف و تجمیع آنها در جهت انجام تحلیل های فراگیر می باشد اما تجمیع چنین داده هایی بدلیل فرادستگاهی بودن آنها تقریبا غیر ممکن است .

داده کاوی توزیع شده[۲] راه حل های خوبی برای توصیف، مدل سازی و تبیین چنین مسائلی را ارائه می کند. اما اغلب این الگوریتم ها مشتمل بر تحلیل موضعی اطلاعات و تولید یک مدل داده ای بر اساس تجمیع نتایج تحلیل موضعی و بخشی است. متاسفانه رویکرد محلی برای تحلیل موضعی، ممکن است منجر به تولید مدلهای داده ای عمومی مبهم یا نادرست گردد. بویژه در حالت کلی زمانی که سایت های مولد داده یا اطلاعات متفاوت دارای مشاهدات با مجموعه ویژگیهای متفاوت باشند، این موضوع بحرانی تر می شود. بنابراین توسعه یک متدلوژی با پایه های مستحکم برای پرداختن به این موضوع بسیار حائز اهمیت است.

سایتهای داده ای ممکن است ناهمگون باشند بدین معنی که هر سایت، داده ها را دقیقا بر اساس مجموعه ای از ویژگی های مورد نظر خود نگهداری نماید. می توان فرض کرد هر سایت، پایگاه داده ای را با انواع مختلف اطلاعات نگه داری می کند .حساسیت های مربوط به حفظ محرمانگی اطلاعات در اغلب موارد مانع از ترکیب کلیه این اطلاعات در یک پایگاه داده واحد می شود. اما روش داده کاوی جمعی این امکان را فراهم می آورد که، الگوهای داده ای و اطلاعاتی با در نظر گرفتن محرمانگی و سیاست های امنیتی و اطلاعاتی، از پایگاه داده ای منفرد استخراج شوند.

مثال دیگر داده های ناهمگون توزیع شده عبارت است از اطلاعات موضعی حاصل از کارکرد های مجموعه بزرگی از حسگر های توزیع شده (محیطی یاعملیاتی) که دارای پارامترها و شاخص های اندازه گیری و مشاهده متفاوت می باشند. ممکن است هیچ مانع سازمانی در مقابل تمرکز و تجمیع داده ای وجود نداشته باشد، اما محدودیت های زمانی و پردازشی مربوط به تجمیع اطلاعات، روش داده کاوی جمعی را در مقابل سایر روشهای متمرکز ممتاز می سازد . بطور کلی ویژگی های مشاهده شده در سایت های مختلف متفاوت می باشد[۲].

سابقه تحقیق

در راستای تحلیل داده های توزیع شده، سه مبحث داده کاوی، تبدیلات موجکی[۳] و تحلیل داده های توزیع شده مورد بررسی قرار می گیرد. در این مقاله روشی برای تحلیل داده های توزیع شده بر اساس داده کاوی جمعی موجکی[۴] ارائه می شود . داده کاوی جمعی، توسعه ای از داده کاوی توزیع شده[۵] است که مسائل حاصل از داده های ناهمگون و غیر همجنس را در مجموعه مشاهدات توزیع شده مد نظر قرار می دهد.

بعبارتی داده کاوی جمعی به بررسی تحلیل اطلاعاتی و مدل سازی موضعی و کلی و تعیین نوع ارتباطات اطلاعات با همدیگر، مشتمل بر داده های تولید شده در سایتهای داده ای با اطلاعات ناهمگون که توسط روش داده کاوی توزیع شده پشتیبانی نمی شود، می پردازد.

داده کاوی توزیع شده، روشهای یافتن الگوی داده ای در مجموعه داده های توزیع شده و محیط های محاسباتی را مورد بحث قرار می دهد و

---

[۲] Distributed Data Mining
[۳] Wavelet Transformation
[٤] CDM (Collective Data Mining)
[٥] DDM (Distributed Data Mining)

امکان آنالیز داده های توزیع شده را با حداقل نیاز به تبادل داده ای فراهم می کند. عموما ، الگوریتم های داده کاوی توزیع شده، با آنالیز داده های موضعی مرتبط با مدل عمومی و بر اساس ترکیب نتایج حاصل از آنالیز موضعی شروع می شود.

در حالت کلی هنگامی که سایت های متفاوت دارای مشاهداتی با مجموعه ویژگی های متفاوت است، رویکرد ساده برای آنالیز داده ها ممکن است مبهم یا حتی ناصحیح باشد و منجر به نادرستی نتایج در مدل عمومی شود. اما پس از بررسی روش داده کاوی توزیع شده، بررسی روش داده کاوی جمعی، نشان میدهد داده کاوی جمعی متدلوژی قوی و محکمی را جهت بررسی این موضوع از طریق پرداختن به پایگاه های داده ای توزیع شده ناهمگون با فضای ویژگی های متمایز ارائه می نماید.

پایه و اساس داده کاوی جمعی، مشاهده ای است که در آن هر تابع امکان نمایش در فرم توزیع شده با استفاده از مجموعه مناسب توابع پایه ای را داراست[۲۶]. (هارموث، ۱۹۷۲)

بر اساس تئوری ارتباطات، انتقال موثر اطلاعات از طریق استفاده از توابع متعامد انجام پذیر است.

تکنیک های آنالیز موجکی، ابزار قدرتمندی برای ایجاد مجموعه توابع پایه ای متعامد برای استفاده در داده کاوی جمعی فراهم می کند. رگرسیون چند متغیره پارامتری یک تکنیک پرکاربرد تحلیل داده ای آماری است که بعنوان یک الگوریتم یادگیری نظارت شده بکار می رود.

پایه های داده کاوی جمعی

. آنچه که در این بخش نشان داده می شود خلاصه کارهایی است که در رگرسیون چند متغیره توزیع شده بکار می رود. در روابط رگرسیون چند متغیره بین اعضای مختلف دامنه و اعضای مربوطه در فاصله (عناوین کلاس یا مقادیر تابع خروجی که با Y نشان داده می شود) مطلوب می باشند.

هدف عبارتست از آموزش تابع $\hat{f}: X^L \to Y$ از مجموعه داده ها

$$\Omega = \{(\mathbf{x}_{(1)}, y_{(1)}), (\mathbf{x}_{(2)}, y_{(2)}), \cdots, (\mathbf{x}_{(N)}, y_{(N)})\}$$

ایجاد شده توسط تابع زیرین $\hat{f}: X^L \to Y$ بنحوی که $\hat{f}$ ، f را برآورد کند. اعضای فردی دامنه-L $\mathbf{x} = x_1, x_2, \cdots, x_L$ چند تایی بوده و x_l ها مربوط به ویژگیهای فردی دامنه هستند. بنیان داده کاوی جمعی بر اساس این واقعیت می باشد که با استفاده از مجموعه توابع پایه ای مناسب، هر تابع می تواند در یک فرم توزیع معرفی گردد.

فرض کنید Θ مجموعه ای احتمالا نامتناهی از توابع پایه ای باشد. شاخصی را همراه با هرکدام از توابع در Θ همراه سازید و تابع پایه ای kام را در Θ با Ψ_k و مجموعه ای از تمامی این شاخص های توابع پایه ای را مانند Θ_i مشخص کنید. تابع $f(\mathbf{x})$ می تواند بشکل زیر بیان شود .

$$f(\mathbf{x}) = \sum_{k \in \Theta_I} w_k \Psi_k(\mathbf{x})$$

که w_k ضریبk امین تابع پایه ایست. هدف، ایجاد تابع $\hat{f}(\mathbf{x})$ می باشد که $f(\mathbf{x})$ را از مجموعه داده های زیر برآورد می کند .

$$\hat{f}(\mathbf{x}) = \sum_{k \in \hat{\Theta}_I} \hat{w}_k \Psi_k(\mathbf{x})$$

که $\hat{\Theta}_I$ نشان دهنده زیر مجموعه Θ_I و $\hat{w}_k$ برآورد تقریبی ضریب w_k است.

برای رگرسیون چند متغیره توزیع شده با استفاده از داده کاوی جمعی موجکی گام بعدی عبارتست از محاسبه ضرایب معنا دار موجکی. $\hat{w}_k$ اگر تابعی، تعداد زیادی ضرایب پایه ای معنادار داشته باشد، برای محاسبه نمایش یکامتعامد، زمان بشکل نمایی درمیاید. (در تعداد ویژگیها). برای داشتن هزینه محاسباتی چندجمله ای برای ضرایب، دو شرط بایدبرقرار شود:

۱. نمایش تنک که در آن اغلب ضرایب صفر یا قابل اغماضند.
۲. ارزیابی تقریبی ضرایب معنادار.

در اغلب مدلهای رگرسیون چند متغیره ، غیر خطی بودن نوعا کران دار می ماند. بنابراین تمامی ویژگی ها بطور غیر خطی با هر ویژگی دیگر تداخل نمی کنند. معمولا قابل قبول است فرض شود که تعداد ویژگی هایی که بطور غیر خطی با هر ویژگی داده شده تعامل می کند، توسط تعدادی مقدار ثابت محدود می شود. اگر چنین نباشد پس مسئله کاملا غیر خطی است و احتمالا حتی در مورد الگوریتم داده کاوی مرکزی پیش از داده کاوی توزیع شده مشکل پیش خواهد آمد.

این نیاز، ریشه ای عمیق در موارد زمان محاسباتی چند جمله ای و احتمالی و قابلیت یادگیری تقریبی دارد (کوشیلویتز و منصور،۱۹۹۱) [۴۶]. غیر خطی بودن محدود به ما این اطمینان را می دهد که نمایش یکامتعامد تنک خواهد بود که در نتیجه اولین شرط محاسبه زمان چند جمله ای را برآورد می کند. شرط دوم با این واقعیت همراه است که تنها نمونه ای از دامنه برای محاسبه ضرایب پایه ای قابل دسترس می باشد. تا آنجا که اندازه نمونه ما منطقی باشد، این موضوع ایجاد مشکل نخواهد کرد. برای نشان دادن منطق پشت این مشاهده، فرض کنید که هر دو طرف معادله به $\Psi_j(\mathbf{x})$ ضرب شود و منجر به معادله

$$f(\mathbf{x})\Psi_j(\mathbf{x}) = \sum_{k \in \Theta_I} w_k \Psi_k(\mathbf{x})\Psi_j(\mathbf{x})$$

گردد، اگر مجموعه داده نمونه با Ω نشان داده شود آنگاه با جمع هر دو طرف روی تمام اعضای Ω معادله زیر نتیجه میشود :

$$\sum_{\mathbf{x} \in \Omega} f(\mathbf{x})\Psi_j(\mathbf{x}) = \sum_{\mathbf{x} \in \Omega} \sum_{k \in \Theta_i} w_k \Psi_k(\mathbf{x})\Psi_j(\mathbf{x})$$

ازآنجاییکه $\sum_{\mathbf{x} \in \Omega} \Psi_j(\mathbf{x})\Psi_j(\mathbf{x}) = |\Omega|$ پس $\Psi_j(\mathbf{x})\Psi_j(\mathbf{x}) = 1$ که $|\Omega|$ حجم نمونه بوده و به قرار زیر است .

$$\frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} f(\mathbf{x})\Psi_j(\mathbf{x}) = \omega_j + \sum_{k \in \Theta_I, k \notin j} \omega_k \sum_{\mathbf{x} \in \Omega} \frac{\Psi_k(\mathbf{x})\Psi_j(\mathbf{x})}{|\Omega|}$$

اگر میانگین جامعه روی کل دامنه برابر صفر باشد پس میانگین نمونه با افزایش اندازه نمونه باید به صفر نزدیک شود. از انجاییکه $\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\Psi_k(\mathbf{x})\Psi_j(\mathbf{x})$ میانگین نمونه است، نتیجه می شود که در مورد اندازه های نمونه بزرگ (نوعا در مورد مسائل مربوط به داده کاوی) جمله آخر باید به صفر میل کند. از این بررسی نتیجه مهم زیر حاصل می گردد که "*ضرایب موجکی محاسبه شده بر روی مجموعه بقدر کافی بزرگ نمونه ها، می تواند ضرایب واقعی را بخوبی برآورد نماید.*" [۲۵]-[۲۱].

تدوین مدل و آزمایش

در این بخش برای نشان دادن و معین کردن روندهای اجرایی روش داده کاوی جمعی-رگرسیون چند متغیره بر حسب

۱. انتخاب مناسب توابع موجکی
۲. تعداد جملات متقابل و جملات مرتبه بالا به نسبت تعداد ویژگیها
۳. اندازه نمونه برای یک مسئله مفروض

از چندین مجموعه داده بزرگ (تا ۱۰۰ مگابایت) استفاده می کنیم. علاوه بر اینکه ما مقیاس پذیری را نشان می دهیم، متریک پایه که برای اندازه گیری این اجرا بکار می بریم، مقدار باقیمانده o_i تعریف شده توسط

$$y_i - \hat{y}_i = o_i$$

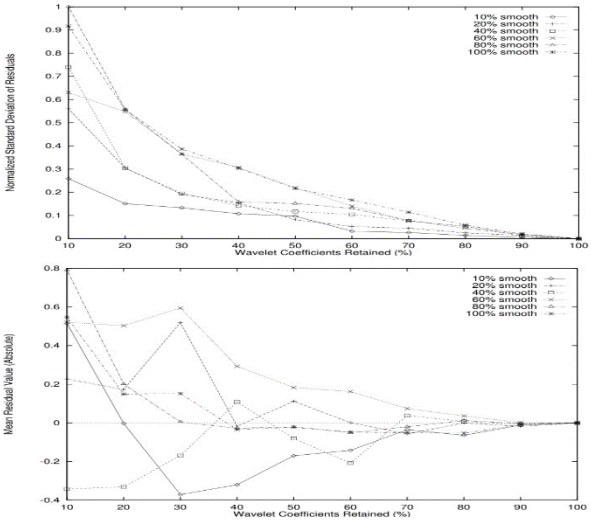
میباشد که y_i تابع عملی یا مقدار ویژگی وابسته می باشد. بعنوان مثال $\hat{y}_i$ و i برآورد آن مقدار ایجاد شده توسط مدل رگرسیون می باشد. این مشابه مقدار باقیمانده محاسبه شده در رگرسیون پارامتری کلاسیک می باشد اما این تفاوت کلی را دارد که نمونه های مورد استفاده از داده ها خارج از نمونه هستند ونه از نمونه های مورد استفاده برای ساختن مدل رگرسیون می باشد. مقدار باقیمانده متوسط بر روی مجموعه داده های آزمایشی

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad ,$$

نزدیک به صفر باقی خواهد ماند مگر اینکه یک انحراف در مدل معرفی کند. انحراف معیار استاندارد مقدار باقیمانده

$$\sigma_0 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(o_i - \bar{o}_i)}$$

تفکری را در مورد توزیع باقیمانده حول میانگین فراهم می آورد. [۳۶]-[۳۱].



شـکل ۱ دقت مدل بخاطر ضـرایب کمتر موجک که بعنوان مشـخصـه نمونه ویژگی تولید شده و با هموارتر شدن تابع موجک استوارتر میشود، کاهش می یابد.

همانگونه که ما صرفا به اجرای نسبی مدلها علاقمندیم، انحراف معیار استاندارد مقادیر باقیمانده برای مجموعه ای از مدلها که قرار است مقایسه شوند به نحوی نرمال میشوند که انحراف معیار استاندارد بیشینه یک باشد.

روشهای موجکی دارای انواع زیادی هستند. هر مجموعه ای از توابع موجکی متعامد ممکن است بطور بالقوه برای داده کاوی جمعی پایه موجکی بکار رود. موضوع خیلی مهم در انتخاب یک تابع موجکی خاص برای یک مسئله این است که نمایش ویژگی بر حسب ضرایب موجکی تا چه حد کم است. با کمتر شدن نمایش گر ضرایب موجکی برای یک مدل موضعی با صحت معلوم مورد نیاز است. در این تحقیق موجک های هار بدلیل اینکه شکل ریاضی نسبتا ساده آنها شفافیت لازم در توسعه روشهای رگرسیون چند متغیره توزیع شده را فراهم میآورد، انتخاب شدند. موجک های هار کمترین وهمواترین اعضای خانواده موجکی هستند که به آنها تعلق دارند. (باربارا بورک هوبارد، ۱۹۹۸) [۱۰].

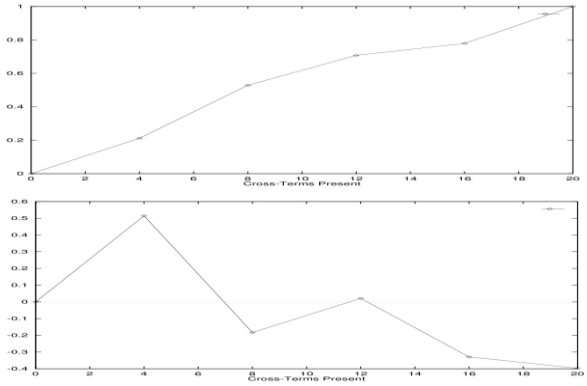
برای نشان دادن اهمیت انطباق تابع موجک با ویژگی ها و خواص داده، مسئله را از ابتدا شروع کرده و رگرسیون چند متغیره توزیع شده بر مبنای هار به مجموعه سری مجموعه داده ها بکار می بندیم که درجه متغیری (اندازه از) مناسب بودن برای استفاده از موجک ها را فراهم می کند.

تابعی با ۱۵ جمله خطی و چهار جمله متقاطع (کلا ۱۹ جمله) بر اساس مجموعه ویژگی باندازه ۱۵، برای این منظور بکار برده می شود. نمونه ویژگی بصورت تصادفی با احتمال یکنواخت روی بازه $[-۱۰.۰,۱۰.۰]$ ایجاد شده اند. همواری داده ها با معرفی احتمالی که

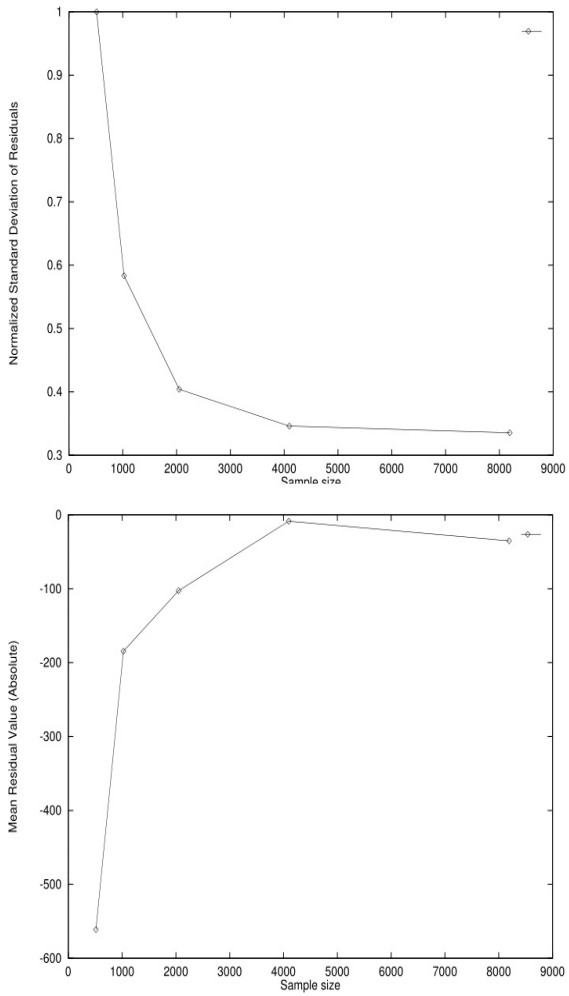
مقدار نمونه هر ویژگی از یک نمونه تا نمونه دیگر تغییر خواهد کرد، متغیر است. هموار ترین داده با احتمال ٪۱۰۰ تولید شد که مقادیر نمونه ویژگی از یک نمونه تا نمونه دیگر تغییر خواهد کرد. (با محدود کردن شانس اینکه مولد اعداد تصادفی همان مقادیر را برگرداند.(کمترین داده هموار تنها ٪۱۰ احتمال مقادیر نمونه را دارا بود که از یک نمونه تا نمونه دیگر تغییر میکرد.

موارد آزمون با درجه هموارى بین ۱۰ الی ٪۱۰۰ برای ارزیابی اثر روی دقت مدل کلی بعنوان تابعی از درصد ضرایب موجکی باقیمانده (نگه داشته شده) در مدلهای موضعی اجرا شدند.

شکل (۲) نشان می دهد که برای یک درصد معلوم (داده شده) ضرایب موجکی نگه داشته شده، مجموعه داده های با درجه هموارى کمتر بسمت دقت بیشتر میل می کنند چرا که آنها با معادلات موجکی سازگارترند. همچنین نشان میدهد که دقت مدل کلی همانگونه که با انحراف معیار استاندارد مقدار باقیمانده نرمال شده مقدار باقیمانده متوسط اندازه گیری میشود. با افزایش ضرایب موجکی بدست آمده درصد آن نیز افزایش می یابد. در شکل ۲ تاثیر تعداد متغیر جملات متقابل نسبت به تعداد ویژگیهای ثابت نشان داده می شود .



شکل ۲ باقیمانده انحراف استاندارد بصورت خطی افزایش و باقیمانده میانگین بازای افزایش تعداد جملات متقاطع نزدیک صفر باقی میماند



شکل ۳ : محاسبات مجموعه داده های بزرگ با با احتساب بخشهای مختلف ضرایب

اجرای الگوریتم رگرسیون داده کاوی جمعی

با فرض N نمونه از هر ویژگی، الگوریتم تجزیه بسته موجکی، Nعمل محاسباتی برای هر کدام از $\log_2^N$ تجزیه، انجام می دهد که منجر به پیچیدگی زمانی $O(N\log N)$ می شود. اگر M ویژگی در یک پارتیشن وجود داشته باشد، آنگاه $O(MN\log N)$ ، هزینه زمانی برای محاسبه ضرایب در مورد تمامی ویژگیها در آن پارتیشن نیاز خواهد بود .

انحراف استاندارد نرمال	باقیمانده میانگین	درصد ضرایب حفظ شده
۰.۲۵۲	۰.۰۰۶	
۰.۲۸۱	۰.۰۱۵	
۰.۴۰۱	۰.۰۰۷	
۱.۰۰۰	۰.۰۲۳	

بمحض اینکه تمامی ضرایب موجکی تعیین شده متمرکز شدند، ممکن است لازم باشد که ضرایب جملات متقاطع نیز محاسبه شود. با فرض اینکه N نمونه وجود داشته دارد، و با فرض اینکه r درصد از ضرایب

موجکی برای هر ویژگی حفظ می شود و با فرض اینکه جملات متقابل بستگی به ویژگی هایی در تقریبا p پارتیشن دارد آنگاه بدترین حالت هزینه محاسباتی $O((rN)^p)$ خواهد بود.

الگوریتم رگرسیون همانگونه که برای اینکار بکار رفته است، توسط زمان لازم برای ترتیب دادن معادلات همزمان حکم می کند. با فرض L رگرسور مستقل شامل متغیر های ساختگی برای جملات قطع کننده در صورت نیاز، تنظیم معادلات نیازمند هزینه زمانی $O(NL^2)$ میباشد .

آنالیز ناپیوسته خطی[۶] نوع و فرم دیگری از یادگیری هدایت شده مربوط به رگرسیون چند متغیره است. با فرض معلوم بودن دو جامعه که برای آن همان ویژگی ها اندازه گیری می شود، که از هر جامعه با عضویت معلوم نمونه برداری می شود برای ایجاد قانون تصمیم گیری استفاده می شود. (ج. فریمن، ۱۹۷۰)[۲۸].

شاهدات با اعضای جامعه نامعلوم با احتمال بالا ممکن است با استفاده از قانون تصمیم گیری بصورت صحیح طبقه بندی گردد.

تحلیل جدا سازی خطی (ناپیوسته) در داده کاوی جمعی

فیشر به یک تساوی بین آنالیز ناپیوسته خطی و رگرسیون چند متغیره اشاره نموده است. داخل مدل رگرسیون شبه متغیرهای نشانگر کلاسهای جامعه ای بعنوان متغیر های وابسته بکار گرفته می شود. در مورد مسئله دو کلاسی:

$$f(\bar{x}_i) = \begin{cases} c_1 & \text{باشد 1 کلاس از ام } i \text{ مشاهده وقتی} \\ c_2 & \text{باشد 2 کلاس از ام } i \text{ مشاهده وقتی} \end{cases}$$

فیشر پیشنهاد کرد که مقادیر شبه متغیر از قرار زیر باشد :

$$c_1 = \frac{n_2}{n_1 + n_2} \quad , \quad c_2 = \frac{-n_1}{n_1 + n_2}$$

که n_1, n_2 تعداد مثالهای آموزشی بترتیب از کلاسهای ۱ و ۲ هستند.

از جایگاه تئوری، تفاوت بین رگرسیون چند متغیره و آنالیز ناپیوسته خطی فیشر آنست که در مورد رگرسیون چند متغیره فرض می شود که متغیر های مستقل دقیقا با هر متغیر مندرج در متغیر وابسته باشد در حالیکه در آنالیز ناپیوسته خطی متغیر وابسته (کلاس) دقیقا معلوم است و هرگونه تغییر پذیری در متغیر های مستقل مندرج است

از منظر کاربردی، تفاوت مهم بین رگرسیون چند متغیره و آنالیز ناپیوسته خطی چگونگی استفاده از مدل است، نه داده کاوی جمعی پایه و نه تکنیکهای رگرسیونی برای ایجاد مدل استفاده شده است. (ر.آ. فیشر،۱۹۳۶) [۳۹].

در مورد آنالیز ناپیوسته خطی، نتایج بکار بستن مدل رگرسیون در مجموعه ای از ویژگیها در مشاهده کلاس نامعلوم با یک مقدار مرزی تصمیم مقایسه می شود که بعنوان قسمتی از حالت یادگیری یا یاد دادن تعیین می شود.

با استفاده از مقادیر شبه متغیر پیشنهاد شده توسط فیشر در مورد مقادیر کلاس و با این فرض که کوواریانس های ویژگی بین جامعه ها بطور قابل ملاحظه متفاوت نیستند، مقادیر مرزی تصمیم ۰/۰ می گردد. آنچه که اختلاف قابل ملاحظه در کوواریانس ویژگی بین جامعه در چارچوب داده کاوی جمعی مقرر می کند و اینکه این مسئله چگونه ممکن است در مورد جامعه های معلوم مورد ارزیابی قرار بگیرد به راه حل آتی موکول می شود.

بخش بعدی کاربرد آنالیز ناپیوسته خطی بر روی مجموعه داده محموعه داده (۱) که شاخص پرکاربرد برای آمار و یادگیری ماشین است توصیف می کند .

جدول ۲: نتایج اعتبار سنجی بر روی داده های محموعه داده (۱)

درصد دقت	طبقه بندی ناصحیح	طبقه بندی صحیح	مورد
۱۰۰/۰	۰	۴۸	
۸۴/۰	۷	۴۱	
۸۴/۰	۷	۴۱	
۹۰/۳	۱۴	۱۳۰	ترکیبی

تحلیل جداسازی (ناپیوسته) داده کاوی جمعی خطی(داده های مجموعه (۱))

در این بخش، آنالیز ناپیوسته خطی توزیع شده بر روی مجموعه داده های محموعه داده (۱) بکار رفته است که مشتمل بر اندازه گیریهای چهار ویژگی سه گونه داده مجموعه داده (۱) میباشد. مجموعه داده شامل ۱۵۰ مثال است. ۵۰ مورد برای هر کلاس یا هر گونه از هر کلاس، دو نمونه بطور تصادفی حذف شده است. ٪۵۰ مورد برای هرگونه بطور نمونه از هر کلاس نو نمونه حذف شده است. ۱۴۴ نمونه باقیمانده به سه گروه ۴۸ نمونه ای ۱۶ نمونه از هر گروه برای تسهیل اعتبار سنجی 3-fold مدل تقسیم شده است.

برای اهداف این نمایش، فرض میشود هر ویژگی که در یک پارتیشن جداگانه ساکن باشد و بردار ستونی با نشان کلاس تنها برای ایجاد مدلهای عمومی مورد نیاز است. بنابراین به هر سایتی ارسال نمی شود. بدلیل اینکه داده های مجموعه داده (۱) نمایانگر سه کلاس می باشد نه دو، یک مرحله دیگر در فرآیند مدل سازی مورد نیاز است. مدلهای اولیه رگرسیون، که بین هر جفت کلاسها تفکیک قائل می شوند، ایجاد می شوند. پس از آن مدل ها در کمیته برای انتخاب طبقه بندی درست برای کلاسهای نامعلوم مشاهداتی استفاده می شود. در حالت گره (هر

[۶] (LDA)-Linear Discriminant Analysis

مدل یک کلاس مختلف انتخاب کند) نزدیک ترین مقدار کلاس برآورد شده به مقدار کلاس مشخص برای هر مدل، جهت انتخاب طبقه بندی بکار می رود.

نتایج اعتبارسنجی موارد آزمون شده در جدول (۲) نشان داده شده است. بطور متوسط مدل های ایجاد شده بر اساس داده کاوی جمعی موجکی موارد آزمون شده خارج از نمونه با ٪۹۰.۳ زمان را بطور صحیح طبقه بندی کرده است. مثالهایی از دقت های گزارش شده برای روشهای مرکزی در جدول (۳) ارائه می شده است.

بدلیل اینکه هیچ جمله مشترک یا جملات مرتبه بالا وجود ندارد، برای ایجاد مدل کلی تنها چهار ضریب موجکی از هر پارتیشن مورد نیاز است. بنابراین مخارج کلی، ارتباط کمی بیشتر از ٪۴ مورد نیاز برای تمرکز تمامی ۹۶ نمونه در هر مجموعه آموزشی در یک سایت بود. مضافا در مورد این مسئله، هزینه ارتباطی چهار ضریب موجکی مستقل از اندازه نمونه می باشد. این یک نتیجه سطح بالا از سازگاری بین نمایش گسسته از متغیر کلاس و توابع موجکی هار است.

جدول ۳: گزارش دقت روش طبقه بندی برای داده های .

روش	دقت-درصد	منبع
خوشه بندی ISODATA)	۰/۸۸	(Freemen, 1970) [14]
کاهش داده NN	۹۶/۷ – ۰/۹۳	(Gates, 1972) [15]
NN	۹۵-۹۷	(Duda & Hart, 1973) [11]
خوشه بندی پارتیشنی	۸۹/۳۳۷	(Duda & Hart, 1973) [11]
خوشه بندی سلسله مراتبی	۹۰/۰	(Duda & Hart, 1973) [11]
درختی	۹۷/۳۳	(Duda & Hart, 1973) [11]

نتیجه گیری و پیشنهادات

این تحقیق روشی برای یارگیری ماشین برای تحلیل داده های توزیع **شده** از طریق رگرسیون چند متغیره توزیع شده با استفاده از داده کاوی جمعی موجکی ارائه می کند. تکنیک رگرسیون پارامتری چند متغیره توزیع ارائه شده در اینجا، اطلاعات موضعی را بر حسب ضرایب یک نمایش تابع اولیه متعامد فرا میگیرد، تعداد کمی (نسبت به اندازه نمونه) از ضرایب قابل ملاحظه را به یک سایت مرکزی انتقال میدهد و یک مدل کلی از آن مجموعه کوچک ضرایب قابل ملاحظه را بطور مستقیم ایجاد می کند.

در کاربرد رگرسیون پارامتری چند متغیره توزیع شده، تکنیک های موجکی برای تولید یک پایه متعامد که یک نمایش توزیع شده تنک یک تابع بعنوان ضرایب تابع پایه ای نشان داده شده اند.

با استفاده از این ضرایب برای تبادل اطلاعات مدل موضعی به یک سایت مرکزی مستلزم به کوچکی ٪۱۰ هزینه ارتباطی میباشد نیازمند به تجمیع یک مجموعه داده های متمرکز خواهد بود.

اهمیت انتخاب توابع موجکی که با ویژگی های داده، کاهش در دقت مدل بعنوان تعداد نسبی افزایش جملات متقابل غیر خطی، و افزایش در دقت مدل با اندازه نمونه سازگار هستند، نشان داده شده بودند. کاربرد روش شناسی داده کاوی موجکی مبتنی بر موجک برای آنالیز متمایز خطی، یک تکنیک مرتبط با رگرسیون چند متغیره نیز ارائه شده اند. یک کاربرد در داده های محموعه داده (۱) با این فرض که هر ویژگی در یک دیتابیس قرار دارد، نشان داد که دقت طبقه بندی مشابه به روشهای مرکزی می باشد. مسائل جداسازی خطی مانند محموعه داده (۱) برای رفتار با موجک های هار بکار رفته در این تحقیق بدلیل طبیعت گسسته ویژگی های کلاس تناسب ویژه ای داشت. هزینه های ارتباطی برای این مسئله نشان داده شده بودند که بطور مستقیم با تعداد ویژگیهای مستقل در اندازه نمونه توابع جداساز و مستقل متناسب بودند.

این روش بطور یکپارچه یادگیری ماشین و تئوری ارتباطات را با روشهای آماری بکار رفته در رگرسیون چند متغیره پارامتری برای آماده سازی یک تکنیک موثر داده کاوی برای تحلیل اطلاعاتی جهت استفاده در حوزه یک داده توزیع شده و محیط محاسباتی ترکیب مینماید. این عملیات منجر به یادگیری ماشین های تصمیم گیر و تصمیم ساز بر اساس تحلیل های خودکار بر روی داده های توزیع شده بصورت بلادرنگ میگردد. این ماشین ها بعنوان سامانه های مکانیزه تحلیل و برآورد اطلاعاتی جایگاه مهمی در ماموریت های مختلف بویژه در موقعیت هایی که نیاز به تحلیل بلادرنگ و پیوسته روی جریان دریافت اطلاعاتی بصورت مستمر وجود دارد کارایی خواهند داشت .

مثالهای فراوانی برای این کارکردها میتوان عرضه کرد. فرض کنیم برای یافتن تصویری خاص از یک متهم و یا مظنون نیاز به بررسی همزمان حجم بسیار زیادی از داده هایی که احتمال ارتباط با شخص مورد نظر را دارد، وجود داشته باشد. این داده میتوانند از جنس و شکل گوناگون برخوردار باشند. مانند تصاویر دوربین های شهری، رکوردهای ثبت شده در محلهای اقامتی، رکوردهای ثبت مسافرتی، مکالمات تلفنی، مراودات پیام کوتاه و متنی یا شبکه های مجازی و غیره باشد .

در این صورت شروع و پردازش همزمان این حجم داده ای، بویژه در موقعیتهای خاص زمانی توسط عوامل انسانی بسیار مشکل و احتمالا ناممکن خواهد بود. در حالی که با استفاده از ماشین های آموزش دیده این عملیات، پردازش های و تحلیل های روی مجموعه داده های متنوع و مختلف توزیع شده در بانک های مختلف در حد اقل زمان و با کمترین هزینه جابجایی اطلاعات امکان پذیر است.

تحقیقات آتی دو مسیر مشخص متمایز، کشف بیشتر استفاده از تکنیکهای موجکی در این زمینه و گسترش این تکنیک های داده کاوی جمعی به دیگر مسائل آموزشی دامنه واقعی، دنبال خواهد کرد.

کار ارائه شده در این تحقیق بر اساس موجک های هار و پایه بسته موجک هار-والش میباشد. موجک های متعامد پررتبه (هموارتر) و دیگر پایه های موجکی، مانند چند حلی یا ترتیب پلی ممکن است عملکرد بهبود یافته ای در بعضی موارد را فراهم آورد. توانایی پیش مشخص سازی مجموعه داده ها بر حسب پایه تابع موجک مناسب در حال حاضر در حال تحقیق و بررسی است. تحقیق بر روی گسترش تکنیک های داده کاوی جمعی موجکی با دامنه واقعی برای مدلهای شبکه عصبی یادگیر در حال انجام است.

مراجع و منابع

[۱]. حیدر مختاری فریور ،محمود افشاری، (۱۳۹۴). برآورد تابع رگرسیونی با نقطه مشخص تغییر کننده با استفاده از موجک های کرانه ای. (هشتمین همایش ملی تخصصی آمار-۶-۷ اسفند ۱۳۹۴ دانشگاه پیام نور تهران- مجموعه مقالات ص ۸۴۷-۸۵۸)

[۲]. حیدر مختاری فریور ،محمود افشاری، (۱۳۹۴). کاربرد داده کاوی در برآورد رگرسیونی چند متغیره توزیع شده و یادگیری ماشین بر اساس موجک ها. (هشتمین همایش ملی تخصصی آمار-۶-۷ اسفند ۱۳۹۴ دانشگاه پیام نور تهران- مجموعه مقالات ص ۸۵۹-۸۷۵)

[۳]. حیدر مختاری فریور ، عبدالمجید مصلح، محمود افشاری، (۱۳۹۵). بررسی الگوی های تحلیل اطلاعات بر روی داده های توزیع شده با مشخصه بلادرنگ. همایش ملی روش های تحلیل اطلاعات- 16 خرداد 1395، دانشگاه اطلاعات و امنیت ملی(وزارت اطلاعات) . مجموعه مقالات

[4] Afshari, M. Estimation of Hazard Function for Censoring Random Variable bu Using Wavelet Decomposition and Evaluate of MISE, AMSE with Similation. Journal of Data Analysis and Information Processing, 2, 1-5, 2014, http://dx.doi.org/10.4236/jdaip.2014.21001

[5] Afshari, M. Threshholding Kernel-regression estimation and empirical distribution of Wavelet coefficients for regularity regression function. World applied Sciences Journal , 4(4):605-609.2006

[6] Afshari, M., Doosti, H., Niroomand, H. A. Wavelets based estimation of time stochastic process. J the derivative of a density for a discrete Journal of Sciences, Islamic republic of Iran 17:75-81, 2006 , 4(4):605-609.

[7] Afshari, M., Doosti, H., Niroomand, H. A. Wavelets for nonparametric stochastic regression with mixing stochastic process. Communication of statictics-Theory and methods. Volume 37:373-385.

[8] B. Carnahan, H. A. Luther, and J. O. Wilkes. Applied numerical methods. John Wiley and Sons, Inc., New York, 1969.

[9] B. Flury and H. Riedwyl. Multivariate Statistics A Practical Approach . Chapman and Hall, New York, 1988.

[10] Barbara Burke Hubbard. The World According to Wavelets. A. K. Peters, Ltd., Wellesley, MA, 1998.

[11] D. Cheung, V. Ng, A. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. IEEE Transaction on Knowledge and Data Engineering, 8(6):911–922, 1996.

[12] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5:81–102, 1978.

[13] D. Wolpert. Stacked generalization. Neural Networks, 5:241–259, 1992.

[14] E. Blake, E. Keogh, and C.J. Merz. U.C.I. Repository of machine learning databases, 1998.

[15] Eric Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer, part 1. IEEE Computer Graphics and Applications, 5(3):76–84, May 1995.

[16] Eric Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer, part 2. IEEE Computer Graphics and Applications, 5(4):75–85, June 1995.

[17] F.J. Provost and K. Venkateswarlu. A survey of methods for scaling up inductive learning algorithms. Data Mining and Knowledge Discovery, 3(2):131–169, June 1999.

[18] Fazal Majid i. XWPL, the X Wavelet Packet Laboratory. http://users.math.yale.edu/users/majid/manual/node31.html, 1995

[19] Frederick Mosteller and John W. Tukey. Data Analysis and Regression. AddisonWesley, Menlo Park, CA, 1977.

[20] G. W. Gates. The reduced nearest neighbor rule. IEEE Transactions on Information Theory, 18(3):431–433, 1972.

[21] H. Kargupta, B. Park, D.E. Hershberger, and E. Johnson. Collective data mining: a new perspective toward distributed data mining. Technical Report EECS99001, Washington State University, Department of Electrical Engineering and Computer Science, 1999. To be published in the book “Advances in Distibuted and Parallel Knowledge Discovery.” Eds: Hillol Kargupta and Philip Chan.

[22] H. Kargupta, E. Johnson, E. Riva Sanseverino, H. Park, L. D. Silvestre, and D. Hershberger. Scalable data mining from distributed, heterogeneous data, using collective learning and gene expression based genetic algorithms. Technical Report EECS98001, School of Electrical Engineering and Computer Science, Washington State University, 1998.

[23] H. Kargupta, I. Hamzaoglu, and B. Stafford. Scalable, distributed data mining using an agent based architecture. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, Proceedings of Knowledge Discovery And Data Mining, pages 211–214, Menlo Park, CA, 1997. AAAI Press.

[24] H. Kargupta, I. Hamzaoglu, B. Stafford, V. Hanagandi, and K. Buescher. PADMA: Parallel data mining agent for scalable text classification. In Proceedings Conference on High Performance Computing ’97, pages 290–295. The Society for Computer Simulation International, 1996.

[25] H. Kargupta. Distributed knowledge discovery: A brief overview. In The Proceedings of the Spring 1999 Symposium of the Institute for operations Research and the Management Sciences (INFORMS), May 1999.

[26] H.F. Harmuth. Transmission of Information by Orthogonal Functions. SpringerVerlag, New York, 1972.

[27] J. Chattratichat, J. Darlington, Y. Guo, S. Hedvall, M. Kohler, A. Saleem, J Sutiwaraphun, and D. Yang. Toward scalable learning with nonuniform class and cost distribution: A case study in credit card fraud detection. In Proceeding of the Fourth International Conference on

Knowledge Discovery and Data Mining, pages 74–81, Menlo Park, CA, 1997. AAAI Press.

[47] V. Cho and B W¨uthrich. Toward real time discovery from distributed information sources . In Xingdong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, Research and Development in Knowledge Discovery and Data Mining, number 1394 in Lecture Notes in Computer Science : Lecture Notes in Artificial Intelligence, pages 376–377, New York, 1998. SpringerVerlag. Second PacificAsia Conference,PAKKD98, Melbourne, Australia , April 1998.

[48] W. Lam and A. M. Segre. Distributed data mining of probabilistic knowledge . In Proceedings of the 17th International Conference on Distributed Computing Systems, pages 178–185, Washington, 1997. IEEE Computer Society Press.

[49] W. Lee, S. Stolfo, and K. Mok. A data mining framework for adaptive intrusion detection. To appear in the Proceedings of the 1999 IEEE Symposium on Security and Privacy, IEEE Computer Society Press, 1999.

Knowledge Discovery and Data Mining, page Not available. AAAI Press, September 1998.

[28] J. J. Freemen. Experiments in discrimination and classification. Patterns Recognition, 1:207–218, 1970.

[29] J. M. Aronis, V. Kolluri, F. J. Provost, and B. G. Buchanan. The world: Knowledge discovery from multiple distributed data bases. Technical Report ISL966, Intelligent Systems Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, 1996.

[30] J. W. Longley. An appraisal of least squares programs for the electronic computer from the viewpoint of the user. Journal of the American Statistical Association, 62:819–841, 1967.

[31] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. In Book, lectures, Morgan Kaufmann, 2012

[32] John C. Hull. Option, Futures, and Other Derivatives . Prentice Hall, Upper Saddle River, NJ, 1997.

[33] K. Yamanishi. Distributed cooperative bayesian learning strategies. In Proceedings of COLT 97, pages 250–262, New York, 1997. ACM

[34] M. V. Wickerhauser. Adapted Wavelet Analysis from Theory to Software. A. K. Peters Ltd., 1994.

[35] N.I.S.T. Statistical reference datasets. http://www.nist.gov/itl/div898/strd/ Linear Regression – Longley.

[36] P. Chan and S. Stolfo. Toward parallel and distributed learning by metalearning. In Working Notes AAAI Work. Knowledge Discovery in Databases, pages 227–240. AAAI, 1993.

[37] P. Chan and S. Stolfo. Toward scalable learning with nonuniform class and cost distribution: A case study in credit card fraud detection. In Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining, page Not available. AAAI Press, September 1998.

[38] P. Chan and S. Stolfo. Experiments on multistrategy learning by metalearning. In Proceeding of the Second International Conference on Information Knowledge Management, pages 314–323, 1993.

[39] R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179–188, 1936.

[40] R. Carmona, WenLiang Hwang, and Bruno Torresani. Practical TimeFrequency Analysis , volume 9. Academic Press, San Diego, 1998.

[41] R. Grossman, S. Bailey, S. Kasif, D. Mon, A. Ramu, and B. Malhi. The preliminary design of papyrus: A system for high performance, distributed data mining over clusters, metaclusters and superclusters. Fourth International Conference of Knowledge Discovery and Data Mining, New York, New York, Pages 37–43, 1998.

[42] R. O. Duda and D. E. Hart. Pattern classification and scene analysis. John Wiley and Sons, New York, 1973.

[43] R. Subramonian and S. Parthasarathy. An architecture for distributed data mining. Fourth International Conference of Knowledge Discovery and Data Mining, New York, New York, Pages 44–59, 1998.

[44] Ronald J. Brachman Tej Anand i. The process of knowledge discovery in databases. In Book, Advances in knowledge discovery and data mining, 1996

[45] S. Kushilevitz and Y. Mansour. Learning decision rees using fourier spectrum. In Proc. 23rd Annual ACM Symp. on Theory of Computing, pages 455–464, 1991.

[46] S. Stolfo et al. Jam: Java agents for metalearning over distributed databases. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, Proceedings Third International Conference on