دانشگاه تربیت مدرس

دانشکده **علوم ریاضی**

رساله دوره دکترای آمار

پیوند احتمالاتی رکوردها و تحلیل آماری داده‌های پیوند یافته

توسط

افشین فلاح

استاد راهنما

دکتر محسن محمدزاده

دی ۱۳۸۸

تقدیم به همسرم و فرزندم احسان

قدردانی

سپاس خدایی را که سخنوران در ستودن او بمانند و شمارشگران شمردن نعمتهای او نتوانند. خدایی که پای اندیشه‌ی تیزگام در شناسایی او لنگ است و سَرِ فکرت ژرف، رو به دریای معرفتش بر سنگ[1].

از استاد راهنمای گرامی جناب آقای دکتر محسن محمدزاده، که در طی این سالهای طولانی همواره از الطاف برادرانه‌ی ایشان بهره‌مند بوده و افتخار شاگردی ایشان را داشته‌ام، کمال تشکر را دارم و برای ایشان سلامتی و بهروزی آرزومندم.

از سایر اساتید دوران تحصیل که ذکر نام این بزرگواران در این تنگنا نمی‌گنجد و اعضای هیأت علمی گروه آمار دانشگاه تربیت مدرس، بخاطر زحماتشان متشکرم.

از تمامی دوستان دانشگاهی، که در طی دوران کارشناسی ارشد و دکتری، به دوستی آنان مفتخر بوده‌ام، بخاطر همه چیز متشکرم.

از همسرم و فرزندم احسان، که بیش از هر کس دیگری در خلال دوران دکتری من متقبل سختی و دشواری شده‌اند، صمیمانه سپاسگزارم. بی‌شک طی این مسیر طولانی بدون کمک‌ها و فداکاری‌های بی‌دریغ آنان ممکن نبود.

افشین فلاح، دی ماه ۱۳۸۸

[1] برگرفته از خطبه‌ی اول نهج البلاغه، ترجمه دکتر سید جعفر شهیدی

چکیده

وقتی اطلاعات مختلف مربوط به واحدهای جامعه در چند مجموعه داده یا فایل قرار دارند، بکارگیری تنها یکی از این فایل‌ها به معنی از دست دادن اطلاعات تکمیلی موجود در سایر فایل‌ها است. بنابراین یکپارچه ساختن اطلاعات پراکنده‌ی افراد یک جامعه در مجموعه داده‌های مختلف، می‌تواند برای دسترسی به اطلاعات کامل و غیر تکراری واحدهای جامعه بسیار سودمند باشد. برای این منظور لازم است رکوردهای یکسان درون یک مجموعه داده یا بین مجموعه داده‌های متفاوت، شناسایی و پیوند داده شوند. این کار که پیوند رکوردها نامیده می‌شود، معمولاً به دو صورت تعیینی و احتمالاتی صورت می‌پذیرد. در این رساله پیوند احتمالاتی رکوردها و تحلیل آماری بر مبنای داده‌های پیوند یافته، مورد مطالعه قرار گرفته است. در فصل ۱ مفاهیم اولیه‌ی پیوند رکوردها معرفی شده‌اند. در فصل ۲ مبانی نظری پیوند رکوردها، مدلهای احتمالاتی آن و قواعد پیوند مختلف از دیدگاههای بسامدی و بیزی مورد بحث و بررسی قرار گرفته‌اند. در فصل ۳ پیوند احتمالاتی رکوردهای فارسی که به دلیل ویژگی‌های خاص زبان فارسی دارای مشکلات و پیچیدگی‌های زیادی می‌باشد، مورد بحث قرار گرفته و راهکارهایی برای حل برخی از دشواری‌های آن ارائه شده است و نحوه‌ی بکارگیری آنها در قالب دو مثال کاربردی به نمایش گذاشته شده است. فصل ۴ به تحلیل رگرسیونی با داده‌های پیوند یافته اختصاص دارد. نشان داده شده است که به دلیل وجود خطاهای انطباق، برآوردگرهای کمترین توانهای دوّم ضرایب رگرسیونی در این حالت لزوماً بهینه نیستند. سپس برای تحلیل رگرسیونی با داده‌های پیوند یافته، روشی مبتنی بر لحاظ نمودن توزیع متغیر پاسخ و با تأکید بر رهیافت بیزی پیشنهاد شده و کارایی روش پیشنهادی در یک مطالعه‌ی شبیه‌سازی با سایر روشهای موجود مقایسه شده است. همچنین روشی برای تحلیل رگرسیون لوژستیک با داده‌های پیوند یافته برای متغیر پاسخ گسسته و دوحالتی، با لحاظ نمودن آمیخته‌ای

از توزیع‌های برنولی و استفاده از الگوریتم EM ارائه و بر اساس آن یک برآوردگر ماکسیمم درستنمایی تکراری برای ضرایب رگرسیونی پیشنهاد شده است. کارایی برآوردگر پیشنهادی و تأثیر خطاهای انطباق بر آن نیز در یک مطالعه شبیه‌سازی مورد ارزیابی قرار گرفته است. نهایتاً خلاصه‌ی یافته‌های پژوهشی این رساله به همراه نتایج و پیشنهادات ارائه شده است.

واژه‌های کلیدی: پیوند رکوردها، رگرسیون با داده‌های پیوند یافته، توزیع‌های آمیخته، الگوریتم EM، رهیافت بیزی.

فهرست مندرجات

لیست اشکال

لیست جداول

لیست جداول ح

فصل ۱

مفاهیم اولیه

۱.۱ مقدمه

اطلاعات توصیف کننده‌ی هر واحد جامعه مانند افراد، مکانها، اتفاقات و غیره، رکورد[۱] نامیده می‌شود و شامل اطلاعات جزئی‌تری بنام فیلد[۲] می‌باشد. به عنوان مثال، مجموعه‌ی اطلاعات فردی شامل فیلدهای نام، نام خانوادگی، آدرس و اطلاعاتی از این دست، یک رکورد را تشکیل می‌دهند. مجموعه رکوردهای افراد یک جامعه نیز یک مجموعه داده یا فایل را تشکیل می‌دهند. هنگامی که اطلاعات مختلف در مورد یک موضوع یا افراد یک جامعه در چند فایل قرار دارند، استفاده از یکی از این فایلها به معنی از دست دادن اطلاعات موجود در سایر فایلها است. بنابراین یکپارچه ساختن اطلاعات پراکنده‌ی افراد یک جامعه در مجموعه داده‌های مختلف، می‌تواند بسیار سودمند باشد. در این راستا لازم است رکوردهای یکسان در مجموعه داده‌های متفاوت یا رکوردهای تکراری در یک مجموعه داده، به نحوی شناسایی و فایلی حاوی اطلاعات کامل و غیر تکراری تهیه شود. شناسایی

[۱] Record

[۲] Field

رکوردهای یکسان درون یک مجموعه داده یا بین مجموعه داده‌های متفاوت، پیوند رکوردها[۳] نامیده می‌شود. غالباً مقایسه‌ی رکوردهای موجود در یک فایل با رکوردهای موجود در فایلهای دیگر برای شناسایی رکوردهای مشترک به منظور ساختن یا بهنگام سازی یک فایل پایه برای جامعه و افزایش اطلاعات موجود در مورد واحدهای آماری، با بکارگیری مجموعه داده‌های متعدد و شناسایی و حذف رکوردهای تکراری، انجام می‌شود. چنانچه برای هر رکورد شناساگرهای یکتایی مانند شماره‌ی شناسایی در فایل مشخصات دانشجویان، موجود و عاری از خطا باشند، شناسایی رکوردهای یکسان با مقایسه این شناساگرهای یکتا امکان پذیر خواهد بود. اما چون فایل‌های داده اغلب توسط افراد یا سازمانهای مختلف و با اهداف متفاوت تهیه می‌شوند، معمولاً شناساگرهای متفاوتی در نظر گرفته می‌شوند، به گونه‌ای که با یکدیگر قابل مقایسه نیستند. بنابراین سعی می‌شود از فیلدهای مشترک رکوردها که متغیرهای شناساگر[۴] نامیده می‌شوند، برای بررسی تشابه رکوردها استفاده شود. اما در عمل به تناسب میزان دقت در مراحل جمع‌آوری، ثبت و ضبط اطلاعات، گاهی فیلدهای رکوردها آغشته به خطا هستند. بنابراین روشهای پیوند رکوردها را می‌توان به دو دسته کلی قطعی[۵] و احتمالاتی[۶] تقسیم نمود. وقتی هر رکورد دارای فیلدهای عاری از خطا باشد، از روشهای قطعی که شامل الگوریتم‌های مقایسه‌ی مبتنی بر تکنیک‌های رایانه‌ای برای شناسایی دقیق رکوردهای یکسان هستند، استفاده می‌شود. اما گاهی در عمل تعیین چنین شناساگرهایی در رکوردهای فایل‌های مختلف مقدور نمی‌باشد و اطلاعات برخی فیلدهای مشترک رکوردها ناقص بوده یا به دلایل مختلف قابل مقایسه نیستند. در اینگونه موارد مدلهای احتمالاتی پیوند رکوردها در نبود چنین شناساگرهایی، از فیلدهای مشترک بین رکوردها برای قضاوت در مورد تشابه آنها استفاده می‌کنند. روشهای قطعی تنها قادرند انطباق رکوردهایی را تشخیص دهند که فیلدهای تشکیل

---

[۳] Record Linkage

[۴] Identifier Variables

[۵] Deterministic

[۶] Probabilitistic

دهنده‌ی آنها بطور دقیق و کامل همخوانی داشته باشند. این روشها هیچگونه مؤلفه‌ی تصادفی را در نظر نمی‌گیرند و از این رو بکارگیری آنها به معنی عدم وجود خطا در مراحل مختلف جمع‌آوری، ثبت و مقایسه‌ی رکوردها است. در حالی که معمولاً خطاهای مختلفی در تهیه‌ی فایل‌های اطلاعاتی حجیم رخ می‌دهند. به عنوان مثال زوج رکورد مربوط به دو کارگاه صنعتی که در جدول ۱.۱.۱ نشان

جدول ۱.۱.۱: اطلاعات دو رکورد از مجموعه داده‌های کارگاههای ایران.

ردیف	نام کارگاه	آدرس	کد فعالیت	تعداد کارکنان
۱	رادارخانه پل ذغال	جاده کندوان پل ذغال	۶۴۲۰	۲
۲	راهدار خانه پل ذغال	جاده اصلی کندوان روستای پل ذغال	۷۵۱۳	۳

داده شده است، علی‌رغم برخی تفاوتهای ظاهری، به واحد آماری یکسانی تعلق دارند. در رکورد ردیف ۱، کلمه‌ی راهدارخانه در فیلد نام کارگاه در اثر یک اشتباه تایپی و حذف حرف «ه» به صورت رادارخانه ثبت شده است. فیلد آدرس در رکورد ردیف ۲، با افزودن دو کلمه‌ی «اصلی» و «روستای» به همان فیلد در رکورد ردیف ۱، حاصل می‌شود و لذا مختصات دقیقتری از محل کارگاه مورد نظر را در اختیار قرار می‌دهد. تفاوت بین فیلدهای کد فعالیت و تعداد کارکنان در این دو رکورد نیز به ترتیب ناشی از تغییر در کدهای ثبتی و توسعه‌ی کارگاه هستند. این تفاوتهای ظاهری یا ناشی از خطاهای تصادفی و غیر قابل کنترل هستند یا در طول زمان و به دلیل تغییر در وضعیت رکوردها ایجاد شده‌اند. به هر حال الگوریتم‌های قطعی رایانه‌ای تنها قادر به بازشناسی انطباق‌های دقیق و کامل هستند و به دلیل لحاظ نکردن مولفه‌ی خطا، توان بازشناسی اینگونه انطباق‌ها را ندارند. این در حالی است که الگوریتم‌های احتمالی براساس میزان شباهت بین دو رکورد و با در نظر گرفتن مولفه‌ی خطا در ثبت و مقایسه‌ی رکوردها، در مورد انطباق یا عدم انطباق زوج رکوردها

در سطح خاصی از اطمینان ابراز نظر می‌کنند.

پیوند رکوردها اولین بار توسط نیوکمب و همکاران (۱۹۵۹) به عنوان یک مسأله آماری و برای ردیابی بیماریهای ارثی مورد استفاده قرار گرفت. فلگی و سانتر (۱۹۶۹) نشان دادند که روش مبتنی بر نسبتهای درستنمایی که بطور تجربی توسط نیوکمب و همکاران توسعه داده شده بود، مطابق با نظریه کلاسیک آزمون فرض‌های آماری است. بر این اساس آنها نظریه‌ای مستحکم را ارائه نمودند، که از آن زمان تا کنون بطور گسترده مورد توجه قرار گرفته است. آرمسترانگ و مِی‌دا (۱۹۹۲، ۱۹۹۳)، بلین (۱۹۹۳) منابع تغییرپذیری در پیوند رکوردها را مورد ارزیابی قرار داد. بلین و روبین (۱۹۹۵) مساله برآورد نرخ انواع خطاها را برای ارزیابی پیوند رکوردها مورد بررسی قرار دادند. وینکلر (۱۹۸۹b و ۱۹۸۹c) روشهایی برای اصلاح نقیصه‌ی حاصل از عدم منظور نمودن وابستگی‌های شرطی فیلدها در پیوند رکوردها، بر اساس روشهای احتمالاتی، ارائه کردند. کوپاس و هیلتون (۱۹۹۰) در زمینه مدلهای آماری مورد استفاده در پیوند رکوردها مطالعاتی را انجام دادند. نیوکمب و همکاران (۱۹۹۲ و ۱۹۷۵) نحوه‌ی پیوند رکوردهای حاوی اطلاعات شخصی را مورد بحث قرار دادند. جارو (۱۹۸۹ و ۱۹۹۵)، وینکلر (۱۹۹۳، ۱۹۹۴، ۱۹۹۵ و ۱۹۹۸) و تیبودیو (۱۹۹۳) مسأله‌ی برآورد پارامترهای مدل فلگی–سانتر و امکان بهبود آنها را مورد مطالعه قرار دادند. لارسن و روبین (۲۰۰۱) و لارسن (۲۰۰۴) استفاده از توزیعهای آمیخته در پیوند رکوردها را مورد بررسی قرار دادند. دو و راهم (۲۰۰۲) رهیافتی برای ساختن ترکیبی انعطاف پذیر از روشهای موجود پیوند رکوردها به منظور دستیابی به نتایج بهتر ارائه نمودند. الگوریتم‌ها و معیارهای اندازه‌گیری میزان تشابه نویسه‌ها توسط کوداس و همکاران (۲۰۰۶) مورد بررسی قرار گرفت. وینکلر (۱۹۹۲) و گوماتان و همکاران (۲۰۰۲) روشهای مختلف پیوند رکوردها را به منظور شناسایی نقاط ضعف و قدرت آنها، بصورت تجربی و بر اساس مجموعه داده‌های مختلف مورد ارزیابی و مقایسه قرار دادند. اسچرون و وینکلر (۱۹۹۱، ۱۹۹۳ و ۱۹۹۷) و لاهیری و لارسن (۲۰۰۵) تحلیل رگرسیونی داده‌های پیوند یافته را مورد توجه قرار داده و برآوردگرهایی از نوع کمترین توانهای دوم برای

ضرایب رگرسیونی ارائه نمودند. فورتینی و همکاران (۲۰۰۱) پیوند رکوردها را به عنوان یک مساله استنباط آماری بر اساس رهیافت بیز مورد مطالعه قرار دادند. وینکلر (۲۰۰۲) پیوند رکوردها با استفاده از شبکه‌های بیزی را مورد مطالعه قرار داد. جین و مرورتا (۲۰۰۳) چگونگی افزایش کارایی، در پیوند رکوردهای فایلهای حجیم و مشکلات ناشی از بزرگ بودن فایلها را مورد بحث قرار دادند. اسچورل (۲۰۰۵) روشی برای لحاظ نمودن وابستگی‌های شرطی درون فیلدها بر مبنای ایجاد تغییراتی در الگوریتم EM ارائه نمود. لارسن (۲۰۰۲ و ۲۰۰۵) پیوند رکوردها با رهیافت بیز سلسله مراتبی را مطرح نمود. وینکلر (۲۰۰۶) چالشهای پیش روی روشهای مختلف پیوند رکوردها و مسائل باز در این زمینه را مورد بحث قرار داد. چمبرز (۲۰۰۹) مسأله‌ی برآورد ضرایب رگرسیونی را با بکارگیری ماتریسهای جایگشت تصادفی مورد مطالعه قرار داد.

این رساله به موضوع پیوند احتمالاتی رکوردها و استنباط آماری بر مبنای داده‌های پیوند یافته، اختصاص دارد. در ادامه‌ی این فصل ضمن معرفی زمینه‌های مختلف کاربرد پیوند رکوردها، مفاهیم اولیه‌ی مرتبط با آن ارائه می‌شوند. سپس در فصل ۲ مبانی نظری پیوند رکوردها، مدلهای احتمالاتی، قواعد پیوند و روشهای مختلف برآورد پارامترها در مدلهای احتمالاتی، از دیدگاههای بسامدی و بیزی، مطرح شده‌اند. همچنین برخی ملاحظات کاربردی مهم در پیوند رکوردها مانند نحوه‌ی تعیین آستانه‌های مدل و نرخ خطاهای انطباق، ارائه شده‌اند. فصل ۳ به پیوند احتمالاتی رکوردهای فارسی که به دلیل ویژگی‌های خاص این زبان دارای مشکلات و پیچیدگی‌های زیادی می‌باشد و تاکنون مورد مطالعه قرار نگرفته است، اختصاص یافته و برای حل برخی از دشواری‌های آن از جمله تأثیر نامطلوب داده‌های گمشده بر کارایی الگوریتم پیوند، راهکارهایی ارائه شده و از طریق مطالعه‌ی شبیه‌سازی مورد ارزیابی قرار گرفته است. سپس روشها و راهکارهای ارائه شده برای دو مثال کاربردی در خصوص سرشماری‌های کارگاهی، سرشماری عمومی نفوس و مسکن و طرح آمارگیری از هزینه و درآمد خانوار کشور، بکار گرفته شده و نتایج حاصل مورد ارزیابی قرار گرفته‌اند. در فصل ۴ تحلیل رگرسیونی با داده‌های پیوند یافته مورد بررسی قرار گرفته و

برآوردگرهای کمترین توانهای دوم ضرایب رگرسیونی و واریانسهای آنها ارائه شده‌اند. سپس روشی برای تحلیل رگرسیونی با داده‌های پیوند یافته با لحاظ نمودن توزیع آمیخته‌ی متغیر پاسخ و با تأکید بر رهیافت بیزی پیشنهاد شده است. آنگاه کارایی روش پیشنهادی در یک مطالعه‌ی شبیه‌سازی مورد ارزیابی قرار گرفته و با کارایی روشهای دیگر مقایسه شده است. سپس تحلیل رگرسیون لوژستیک با داده‌های پیوند یافته صفر و یک، مطرح و نحوه‌ی برآورد ماکسیمم درستنمایی ضرایب رگرسیونی با الگوریتم EM ارائه شده است. کارائی برآوردگرهای پیشنهادی و تأثیر خطاهای انطباق بر آنها نیز در یک مطالعه شبیه‌سازی دیگر مورد ارزیابی قرار گرفته است. در انتها نتایج این رساله به همراه پیشنهادات ارائه شده‌اند.

۲.۱ کاربردهای پیوند رکوردها

از جمله دلایل گسترش چشمگیر کاربردهای پیوند رکوردها در زمینه‌های مختلف زندگی بشری، یکی شکل‌گیری فایلهای بزرگی است که لازم است در طول زمان به هنگام شوند و دیگری پیشرفت در تجهیزات رایانه‌ای ثبت، نگهداری و انتقال داده‌ها می‌باشد. محققان بسیاری از روشهای مختلف پیوند رکوردها برای اهداف مختلف سود جسته‌اند، که به چند نمونه از این کاربردها که اغلب در سطح کلان و بین المللی هستند، اشاره می‌شود.

۱- اداره آمار آلمان در تلاش است که به جای سرشماری به روش سنتی، از این پس اطلاعات مورد نیاز خود را بصورت ثبتی و با جمع‌آوری اطلاعاتی که در ادارات و سازمانهای مختلف موجودند، بدست آورد. انگیزه اصلی برای اجرای سرشماری‌های ثبتی[۷] آن است که در آلمان پس از جنگ جهانی دوم سه بار سرشماری نفوس و مسکن در سالهای ۱۹۵۰، ۱۹۶۱ و ۱۹۷۰ با موفقیت انجام شد. اما سرشماری بعدی که قرار بود در بهار سال ۱۹۸۱ اجرا شود، توسط دادگاه

[۷] Registration-Based Census