**مقاله مروری**

# روش شناسی پیوند داده بر اساس الگوریتم احتمالاتی: یک مقاله مروری

***عرفان ایوبی[۱،۲]، کامیار منصوری[۳،۴]، محمد گل ماهی[۵]، عذرا رمضانخانی[۶]، علیرضا موسوی جراحی[۷]***

[۱] دانشجوی دکترای تخصصی اپیدمیولوژی، دانشکده پزشکی، دانشگاه علوم پزشکی زابل، زابل، ایران

[۲] دانشجوی دکترای تخصصی اپیدمیولوژی، گروه اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی تهران، تهران، ایران

[۳] دانشجوی دکترای تخصصی اپیدمیولوژی، دانشکده پزشکی، دانشگاه علوم پزشکی کردستان،سنندج، ایران

[٤] دانشجوی دکترای تخصصی اپیدمیولوژی، گروه اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی ایران، تهران، ایران

[٥] تحلیلگر سیستمهای کامپیوتری، مرکز تحقیقات سرطان دانشگاه علوم پزشکی تهران

[٦] دکتری تخصصی پژوهش، پژوهشکده علوم غدد درون ریز و متابولیسم، دانشگاه علوم پزشکی شهید بهشتی

[۷] دانشیار اپیدمیولوژی، گروه پزشکی اجتماعی، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی

***نشانی نویسنده مسئول:** علیرضا موسوی جراحی، دانشیار اپیدمیولوژی، گروه پزشکی اجتماعی، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی
E-mail: rmosavi@yahoo.com

**وصول:۹۴/۶/۶، اصلاح:۹۴/۸/۱۲، پذیرش:۹۴/۹/۲۷**

## چکیده

**امروزه با پیشرفت تکنولوژی و توسعه پژوهش در کشور، شاهد شکل گرفتن بانک های اطلاعاتی بزرگ و ارزشمند هستیم. لازمه ردیابی اطلاعات افراد در این مجموعه های گرانبها، بکارگیری روشهای جدید واکاوی داده های مرتبط می‌باشد. با این روش‌ها اطلاعات بسیار مفیدی را می‌توان درباره تشخیص، سبب شناختی و پیش آگهی پیامدهای مختلف بدون انجام مطالعات پرهزینه فراهم کرد. گوناگونی در جمع‌آوری و تعاریف فیلد های حاوی داده های سلامت در بانک‌های اطلاعاتی مختلف، نیاز به آشنایی با روش شناسی پیوند داده ها را بیشتر میکند. هدف از این مقاله مروری، آشنایی با روش‌شناسی پیوند داده ها براساس روش‌های احتمالاتی می‌باشد. تعریف پیوند داده در دو روش قطعی و احتمالاتی ارائه خواهد شود و در ادامه مبانی روش شناختی پیوند داده احتمالاتی مانند ارزیابی کیفیت داده ها، ارزیابی همسان بودن رکوردها و محاسبه وزن همسان بودن رکوردها از دو بانک اطلاعاتی به همراه تعیین سطح تصمیم گیری برای همسان بودن آنها بحث خواهد شد. در قالب یک مثال عملی روش‌شناسی پیوند داده احتمالاتی با استفاده از داده های بانک اطلاعات ثبت سرطان و مرگ و میر نشان خواهد داده شد.**

**واژه‌های کلیدی:** *پیوند داده، رویکرد احتمالاتی، ثبت سرطان، ثبت مرگ.*

## مقدمه

### ۱-۱. سابقه متون

پیوند رکوردها یا ارتباط داده‌ها اولین بار توسط نیوکامب و همکاران (۱) به‌عنوان یک مسئله‌ی آماری و برای ردیابی بیماری‌های ارثی مورد استفاده قرار گرفت. امروزه ارتباط الکترونیکی داده‌ها به‌طور وسیعی به‌وسیله محققین و دست‌اندرکاران بهداشتی مورد استفاده قرار می‌گیرد. از مهم‌ترین دلایل گسترش چشمگیر کاربردهای پیوند رکوردها در حوزه سلامت، یکی شکل‌گیری فایل‌های بزرگی است که لازم است در طول زمان به‌روز شوند و دیگری پیشرفتی است که در تجهیزات رایانه‌ای ثبت، نگهداری و انتقال داده‌ها حاصل گردیده است. در

اپیدمیولوژی به فراوانی از علم پیوند داده‌ها استفاده شده که مهم‌ترین مورد آن ایجاد ارتباط بین یک پیامد خاص (مثلاً مرگ در اثر یک علت خاص) با یک عام مواجه می‌باشد. به‌طور مثال با وجود بانک داده ثبت بیماران سرطانی و بانک داده مرگ‌ومیر می‌توان با کمک پیوند داده میزان بقاء بیماران مبتلا به یک سرطان خاص را در شرایط ایده آل محاسبه کنیم. در ادامه به چند نمونه از این کاربردها اشاره می‌شود :

در مطالعه‌ای در هلند به‌منظور بررسی تأثیر غربالگری سرطان پستان، داده‌های مربوط به ۹۰۰۰۰ خانم ۶۹-۴۹ ساله که در برنامه غربالگری سال‌های ۱۹۹۵ - ۱۹۹۰ شرکت کرده بودند، با داده‌های ثبت سرطان سال‌های ۹۶ -۱۹۸۶ این کشور، با روش ارتباط داده‌ها مرتبط شدند (۲).

در مطالعه‌ای با استفاده از ارتباط داده با روش احتمالاتی، احتمال بروز سرطان پستان بعد از بیماری‌های خوش‌خیم مورد بررسی قرار گرفت. در این مطالعه داده‌های کوهورت بیماران مبتلا به ضایعات خوش‌خیم با داده‌های سیستم ثبت سرطان پیوند داده شد و بدین ترتیب پتانسیل بدخیمی ضایعات خوش‌خیم و مدت‌زمانی که برای بدخیم شدن این ضایعات لازم بود، تعیین گردید (۳).

در مطالعه دیگری رابطه بین مصرف هورمون‌های جایگزین در یائسگی و بروز سرطان پستان با استفاده از روش ارتباط داده‌ها موردبررسی قرار گرفت. در این مطالعه زنانی که طبق سیستم ثبت تجویز داروها، بین سال‌های ۸۷-۱۹۷۶ هورمون‌های مذکور را مصرف کرده بودند با داده‌های سیستم ثبت سرطان سال‌های ۹۰-۱۹۶۰ پیوند داده شدند (٤).

**۲-۱. مفاهیم اولیه**

اطلاعات توصیف‌کننده هر واحد جامعه مانند افراد، مکان‌ها، اتفاقات و ...، رکورد (record) نامیده می‌شود. هر رکورد شامل اطلاعات جزیی‌تری بنام فیلد (field) می‌باشد. به‌عنوان مثال، مجموعه‌ی اطلاعات شناساگر فردی شامل فیلدهای نام، نام خانوادگی، آدرس و ... می‌تواند باشد. مجموعه رکوردهای افراد یک جامعه، یک فایل (file) را تشکیل می‌دهند. به مجموعه چندین فایل، بانک داده (dataset) گفته می‌شود. هنگامی‌که داده‌های موضوعی (داده‌هایی که یک صفت خاص را برای یک فرد شامل می‌شود) برای یک فرد در چند مجموعه متفاوت داده یا فایل قرار دارند، یکپارچه ساختن اطلاعات پراکنده می‌تواند موجب جامعیت مطلب و بسیار سودمند باشد و چه‌بسا محدود کردن اطلاعات صرفاً به یکی از این مجموعه داده‌ها ممکن است موجب از دست دادن اطلاعات موجود در سایر مجموعه داده‌ها و ارائه گزارش ناقصی از موضوع شود. 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 غیرتکراری تهیه شود. شناسایی رکوردهای یکسان درون یک مجموعه داده (فایل) یا بین مجموعه داده‌های متفاوت، پیوند رکوردها (record linkage) یا ارتباط داده‌ها (data linkage) نامیده می‌شود.

دو نوع کلی الگوریتم پیوند وجود دارد: قطعی(deterministic) و احتمالی (probabilistic). هر دو این روش‌ها به‌طور مناسبی در تحقیقات قبلی انجام شده است (۵-۱۵). اینکه کدام روش استفاده شود بستگی به فاکتورهای متعددی دارد که برخی از آن‌ها علمی و برخی دیگر ماهیت ذهنی دارند. در سناریوهای که اطلاعات فراوان وجود دارد و متغیرهای شناسگر از کیفیت خوبی برخوردار هستند روش قطعی پیشنهاد شده است اما در حالت‌هایی که داده‌ها کیفیت مناسبی ندارند و یا در دسترس نیستند روش احتمالی در اولویت می‌باشد (۱٦). اما نکته مهم هنر محقق است که از کدام روش استفاده کند برای مثال در هنگام مطالعه یک بیماری نادر استفاده از روش احتمالی حتی در مواقع که اطلاعات فراوان وجود

**جدول ۱: پیوند داده های دو منبع اطلاعاتی ۱ و ۲ با استفاده از روش قطعی**

| منبع ۱ | | | منبع ۲ | | |
|---|---|---|---|---|---|
| نام | نام خانوادگی | تاریخ تولد | نام | نام خانوادگی | تاریخ تولد |
| *لیلا | احمدیان | ۱۳۴۲ | لیلی | احمدیان | ۱۳۴۲ |
| *علی | یزدان خاه | ۱۳۵۶ | علی | یزدان خواه | ۱۳۵۶ |
| **مرضیه | رشیدی | ۱۳۶۸ | مرضیه | رشیدی | ۱۳۶۸ |

* اطلاعات مربوط به یک فرد میباشد ولی در روش قطعی به دو نفر متعلق است.

** فقط این فرد بصورت قطعی میتواند در هر دو بانک اطلاعاتی به عنوان یک فرد تلقی گردد.

دارد در اولویت است چراکه تلاش می‌شود که همه همسان‌ها شناسایی شوند و حجم نمونه حداکثر شود.

**پیوند قطعی:** الگوریتم قطعی بر این مبناست که آیا زوج مقایسه بر مبنای یک مجموعه از متغیرهای شناساگر توافق دارند یا نه؟ به عبارتی این ارزیابی تابع قانون همه یا هیچ می‌باشد. یک زوج مقایسه به‌عنوان همسان طبقه‌بندی می‌شود که دو رکورد جزءبه‌جزء برای همه‌ی شناساگرها توافق داشته باشند و به همین ترتیب یک زوج مقایسه به‌عنوان غیرهمسان طبقه‌بندی می‌شود اگر روی همه متغیرهای شناساگر توافقی ایجاد نشود. (۸,۱۷).

در این روش لازم است دوسری داده‌های ثبت شده در دو بانک اطلاعاتی در فیلدهای شناساگر (مثل نام یا تاریخ تولد) به‌طور دقیق و کامل همخوانی داشته باشند تا آن دو سری مشخصات به یک فرد منسوب شوند. در این روش هیچ‌گونه مؤلفه تصادفی در نظر گرفته نمی‌شود و ازاین‌رو به‌کارگیری آن با فرض عدم وجود خطا و پایداری در فیلدهای شناساگر می‌باشد. معمولاً خطاها و تغییرات مختلفی در فرایند تهیه و ذخیره‌سازی بانک‌های اطلاعاتی به وجود میاید که استفاده از روش قطعی را در ارتباط داده مخصوصاً در غیاب یک شناسه واحد (مثل کد ملی در ایران) غیرممکن و نا کارآ می‌نماید. در روش قطعی حتی تغییرات خیلی کوچک در ویراستاری فیلد های شناساگر، مانع از شناسایی افراد مشابه در دو فایل میگردد. به عنوان مثال در مورد نام خانوادگی ممکن است دو اسم تفاوت هایی جزئی مثلاً در یک حرف داشته باشند، مانند نام‌های "رسایی" و"رضایی" یا از اسامی کوتاه شده باشد مثل لیلی و لیلا که در روش قطعی موجب عدم ارتباط این دو فیلد می‌گردد.

بدین ترتیب بسیاری از یافته های مربوط به یک شخص که با اختلافات جزئی ثبت شده اند با این روش به هم مرتبط نشده و موجب تورش قابل توجه در پیوند داده ها می شود. این اختلافات جزئی در فیلد های شناساگر از تفاوت لهجه‌های محلی در بیان اسامی، اشتباهات تصادفی در مرحله ورود داده و یا به صورتهای مختلف دیگر می-تواند اتفاق بیفتد. به عنوان مثال در جدول ۱ در پیوند داده های منابع اطلاعاتی ۱ و ۲ که فیلدهای نام، نام خانوادگی و تاریخ تولد به عنوان متغیر شناساگر در نظر گرفته شده اند، در روش قطعی، به علت تفاوت های نگارشی، فقط فرد سوم را مرتبط با هم می شناسد. در صورتی که اطلاعات در دو منبع (بانک اطلاعاتی) متعلق به سه فرد مشخص می‌باشد.

**پیوند احتمالاتی:** الگوریتم های احتمالی بر اساس میزان شباهت بین دو رکورد و با در نظر گرفتن مؤلفه ی خطا در ثبت و مقایسه ی رکوردها، در مورد انطباق یا عدم انطباق زوج رکوردها در سطح خاصی از اطمینان عمل می کنند. در این روش طیف متنوعی از ماهیت و کیفیت در داده ها در نظر گرفته می‌شود تا از همسان بودن دو رکورد، علیرغم تفاوت در چند شناسه اطمینان حاصل گردد. ارتباط داده ها بر اساس احتمال نیاز به عملیات کامپیوتری پیچیده‌تری نسبت به روش قطعی دارد و از مبانی علمی پیچیده‌تری برخوردار بوده و استفاده بیشتری در بخش سلامت دارد. در ادامه در قسمت روش شناسی مفاهیم پایه پیوند احتمالاتی به همراه مثالی کاملاً فرضی از دو سری داده مربوط به سیستم ثبت سرطان و بانک داده

**جدول ۲: قسمتی از دو بانک اطلاعاتی مربوط به ثبت سرطان و مرگ و میر در سطح یک جامعه تعریف شده.**

| | بانک داده ثبت سرطان | | | | بانک داده مرگ ومیر | | | |
|---|---|---|---|---|---|---|---|---|
| شماره رکورد | نام | فامیل | تاریخ تولد | نام مادر | شماره رکورد | نام | فامیل | تاریخ تولد | نام مادر |
| ۷۶۵ | شهرام | احمدی | ۶/۲۵/۲۰۰٤ | راحله | ۷۸۶۵۹ | شهرام | احمدی | ۱۱/۳/۲۰۰۲ | روزا |
| ۲۹۰ | اکبر | بدانلو | ٤/۲۳/۲۰۰۱ | زهرا | ۶۷۸٤۵ | اکبر | بدانلو | ٤/۲۳/۲۰۰۱ | زهرا |
| ۵٤۶ | محمد رضا | پداللو | ۱۱/۳/۲۰۰۲ | روزا | ٤۱٤۳۲ | محمد رضا | پداللویی | ۱۱/۳/۲۰۰۲ | روزا |
| ۲۳۵ | محمد رضا | پدالویی | ۱۱/۳/۲۰۰۲ | روزا | ۲٤۳۸۹ | محمد رضا | پدالویی | ۱۱/۳/۲۰۰۲ | روزا |
| ۵۶۸ | محمد رضا | پناهی | ۶/۲۵/۲۰۰٤ | یاسمن | ۲۳٤۱٤ | محمد رضا | پناهی | ٤/۲۳/۲۰۰۱ | راحله |
| ۶۷۸ | علیرضا | شکوفه | ۵/۱۸/۲۰۰۳ | زهرا | ۹۸۷۶٤ | علیرضا | شکوفه | ۱/۲۷/۲۰۰۰ | زینب |
| ۷۸۱ | اسماعیل | عباسی | ٤/۱۶/۲۰۰۲ | فاطمه | ۵٤۸۷۶ | اسماعیل | عاسی | ٤/۱۶/۲۰۰۲ | زینت |
| ٤۵۰ | عباس | موحدی | ۵/۱۸/۲۰۰۳ | فاطمه | ۵۹۶۸۷ | عباس | موحدی | ۵/۱۸/۲۰۰۳ | فاطمه |
| ۶۲۱ | وحید | موسوی | ۱/۲۶/۲۰۰۰ | زری | ٤۶۷۸۹ | وحید | موسوی | ۱/۲۶/۲۰۰۰ | زری |

مرگ و میر که در جدول ۲ نشان داده شده، جهت تشریح مفاهیم ارتباط داده‌ها با روش احتمالی توضیح داده می-شود.

## ۲. پیوند احتمالاتی

### ۱-۲. مبانی نظری و راهکار های ارتباط داده ها براساس احتمال

مبانی نظری ارتباط داده ها بر اساس احتمالات شامل شناخت سه پارامتر ۱) کیفیت داده ها ۲) شانس موافقت تصادفی و ۳) تعداد موافقت صحیح مورد انتظار در دو سری داده است. بر اساس مقدار این پارامترها و با توجه به استراتژی انتخاب شده احتمال تطبیق و یا ارتباط دو رکورد مشخص می‌گردد.

#### ۱-۱-۲. کیفیت داده‌ها

داده‌ها با شکل، اندازه و کیفیت‌های مختلف، سناریوهای مختلفی برای شکل‌دهی الگوریتم‌های پیوند ایجاد می‌کنند. برای مثال اطلاعات دموگرافیک اغلب شامل خطاهای مربوط به واردکردن داده و خطاهای تایپوگرافیکال می‌باشد. اطلاعات افراد در طول زمان تغییر می‌کند برای مثال با ازدواج و یا تغییر محل سکونت، در برخی مواقع نیز افراد به غلط و تعمدی اطلاعات مربوط به خود را گزارش می کنند. در نهایت این ایدئوسنکرازی، پیوند داده را با مشکل همراه می سازد و نیاز است که قبل از پیوند داده، آنها تمیز و استاندارد شوند.

دشوارترین و وقت گیرترین قسمت یک پروژه پیوند داده، پیش از آنکه به الگوریتم پیوند و امکانات رایانه‌ای وابسته باشد، به کیفیت داده‌ها وابسته است. این پارامتر نشان می‌دهد که تا چه حد اطلاعات یک فیلد، دقیق و با ثبات بوده و اندازه‌گیری‌های صحیح را در یک فرد نشان می‌دهند. خطاهای ورود اطلاعات، تعداد کم فیلدهای مشترک مورد مقایسه، اطلاعات از دست رفته و بی‌دقتی‌های عمدی و غیر عمدی در گردآوری اطلاعات، همگی بر دقت یافته‌ها و کیفیت آنان اثر می‌گذارند. پیوند رکوردهایی که اطلاعات آنها به زبان فارسی ثبت شده است به دلیل ویژگی های خاص نوشتارهای فارسی، عدم وجود استانداردی جهت ثبت اطلاعات و وجود داده های گم شده زیاد و عدم وجود تجربیات علمی، ارتباط داده ها را در زبان فارسی با مشکلات بیشتری مواجه ساخته است.

علاوه بر کیفیت آیتم‌های اطلاعاتی موجود در فیلدهای یک بانک اطلاعاتی، پایداری و ثبات داده‌ها نیز مهم است. این موضوع اشاره به این دارد که تا چه حد یک آیتم اطلاعاتی در طول زمان یا در حین جمع‌آوری از منابع مختلف ممکن است متفاوت باشد. برای مثال شماره تلفن یا آدرس محل زندگی که میتواند به سادگی برای افراد تغییر کند در مقابل آیتم‌های اطلاعاتی مثل کد ملی یا جنسیت می‌باشد که از جمله متغیرهایی با ثبات محسوب می‌شوند.

تکنیک‌های مختلفی برای تمیز سازی داده ها معرفی شده است: برخی از آنها منجر به افزایش تعداد متغیر یا جدا کردن فیلد مورد نظر می‌شود، برخی منجر به تبدیل متغیر به یک شکل خاص دیگر است که در اطلاعات واقعی تغییری ایجاد نمی‌شود. تکنیک‌های دیگری وجود دارند که هدف از آنها تغییر اطلاعات در فیلد مورد نظر مثلا با حذف ارزش‌های نامعتبر، پر کردن ارزش‌های گمشده است (۱۸).

برخی از تکنیک‌های تمیزسازی داده شامل موارد زیر است:

- تغییر فرمت داده‌ها
- حذف نشانه گذاری‌های غیرضروری
- حذف ارزش‌های گمشده و خالی از اطلاعات و پر کردن ارزش‌های گمشده
- تغییر شکل آوایی (phonetic encoding)
- استانداردسازی نام خانوادگی و آدرس
- تصحیح ناهمگنی‌ها

در علم ارتباط داده، کیفیت داده ها را با متغیر m و به‌صورت کمی نشان می‌دهند. m پارامتری است که نشان می‌دهد اگر دو رکورد در دو بانک داده واقعاً متعلق به یک نفر باشد چقدر احتمال وجود دارد که آیتم‌های اطلاعاتی هر فیلد مشابه باشند. مقدار m برای تمام داده های هر فیلد، ثابت است. مثلاً در مورد فیلد نام خانوادگی، ۰/۹۵ =m به این معنی است که احتمال اینکه نام خانوادگی ثبت شده در دو رکورد متعلق به یک نفر از دو منبع داده دقیقاً املای مشابه داشته باشند، ۰/۹۵ است و این احتمال برای همه نام های خانوادگی موجود مقدار ثابتی است. علاوه بر استفاده از روش‌های آماری پیچیده برای برآورد مقدار m در فیلد های مختلف، معمولاً مقدار m بر اساس نظر افراد مجرب و آگاه تعیین می‌شود که برحسب تجربه و به مرور زمان و کار روی پروژه های قبلی قابل برآورد میباشد. جدول ۳ نمونه ای از احتمال m برای فیلد های نام ونام خانوادگی، تاریخ تولد، و نام مادر را که توسط آقای G A Mason در بانک داده انگلیسی گزارش شده نمایش می‌دهد و در این مقاله جهت تشریح مفاهیم مورد استفاده قرار خواهد گرفت (۱۹) لازم به ذکر است که احتمال m برای فیلد های مذکور در بانک های داده ایران با توجه به تفاوت و تنوع نگارشی زبان فارسی ممکن است کمتر و یا بیشتر باشد.

**جدول ۳: نمایش احتمال m برای چهار فیلد گزارش شده توسط آقای G. A. Mason**

| نام فیلد | برآورد احتمال m |
|---|---|
| نام | ۰/۹۵ |
| نام خانوادگی | ۰/۹۵ |
| تاریخ تولد | ۰/۹۸ |
| نام مادر | ۰/۹۸ |

**۲-۱-۲. شانس موافقت تصادفی** (randomly linked)

پارامتر دوم در ارتباط داده‌ها با روش احتمالی، شانس موافقت تصادفی است که احتمال همسان بودن دو رکورد به صورت تصادفی در یک فیلد را نشان می‌دهد. هرچه این احتمال بیشتر باشد شانس یک پیوند واقعی کمتر است. برای مثال اگر تمام افراد دو سری بانک داده، مؤنث باشند و متغیر شناساگر در ارتباط دو بانک داده متغیر جنسیت باشد، احتمال اینکه یک فرد از یک بانک داده با فرد دیگری از بانک داده دیگر به‌صورت تصادفی پیوند داده شود، ۱۰۰ درصد می‌باشد. در نتیجه این فیلد نمی‌تواند کمکی در پیوند دو رکود همسان نماید و برای پیوند داده ها مناسب نیست. به دلیل اینکه در هر زوج تصادفی قطعاً یک تطبیق در آن فیلد وجود دارد، اساساً متغیرهایی مثل جنسیت ارزش محدودی در پیوند داده ها دارند چون انتظار می‌رود حتی در شرایطی که افراد در بانک داده از هر دو جنس باشند و ارتباط دو بانک داده بر اساس متغیر جنسیت باشد، در ۵۰ درصد موارد توافق تصادفی حاصل گردد. در حالی‌که متغیری مثل کد ملی، برای اینکار بسیار مناسب است زیرا انتظار یافتن یک همسان تصادفی برای آن وجود ندارد.

**جدول ٤: احتمال u برای آیتم های اطلاعاتی از بانک اطلاعات مرگ میر در سال ١٣٨٦ سازمان بهشت زهرا شامل ٤٨٠٠٠ رکورد استخراج شده است**

| نام | | نام خانوادگی | | تاریخ تولد | | نام مادر | |
|---|---|---|---|---|---|---|---|
| آیتم داده | احتمال u | آیتم داده | احتمال u | آیتم داده | احتمال u | آیتم داده | احتمال u |
| اسماعیل | ۰/۰۰۶۷۶ | عباسی | ۰/۰۰۰۲۱ | ٤/١٦/٢٠٠٢ | ۰/۰۰۰۰٤ | زینت | ۰/۰۰۰۱٤ |
| علیرضا | ۰/۰۰۵۹۰ | شکوفه | ۰/۰۰۰۱۲ | ١/٢٧/٢٠٠٠ | ۰/۰۰۰۰۲ | زینب | ۰/۰۰۰۲۸ |
| محمد رضا | ۰/۰۰۰۲۳ | پناهی | ۰/۰۰۰۳۳ | ۶/٢۶/٢٠٠٤ | ۰/۰۰۰۰۲ | یاسمن | ۰/۰۰۰۱۳ |
| اکبر | ۰/۰۰۱۲۳ | بدانلو | ۰/۰۰۰۰۶ | ٤/٢٣/٢٠٠١ | ۰/۰۰۰۰٤ | راحله | ۰/۰۰۰۵۶ |
| رضا | ۰/۰۰۱۲۸ | پدالویی | ۰/۰۰۰۰۲ | ١١/٣/٢٠٠٢ | ۰/۰۰۰۰۸ | روزا | ۰/۰۰۰۰۸ |
| عباس | ۰/۰۰۰۵۶ | موحدی | ۰/۰۰۰۲۳ | ۵/١٨/٢٠٠٣ | ۰/۰۰۰۰۶ | زهرا | ۰/۰۰۰۸۷ |
| وحید | ۰/۰۰۰۳٤ | موسوی | ۰/۰۰۰۸۹ | ١/٢۶/٢٠٠٠ | ۰/۰۰۰۰۶ | فاطمه | ۰/۰۰۰۹۲ |
| شهرام | ۰/۰۰۰۱۱ | احمدی | ۰/۰۰۰۶۷ | ۶/٢۵/٢٠٠٤ | ۰/۰۰۰۰۵ | زری | ۰/۰۰۰۰۸ |

**جدول ٥: محاسبه وزن برای فیلد برمبنای اطلاعات جدول قبلی بعلاوه وزن کل برای رکورد های ارتباط داده شده**

| شماره ردیف رکورد در بانک ثبت سرطان | شماره ردیف رکورد در بانک مرگ و میر | $W_i$نام | $W_i$فامیل | $W_i$تاریخ تولد | $w_i$نام مادر | وزن کل $w_t$ |
|---|---|---|---|---|---|---|
| ۷۶۵ | ۷۸۶۵۹ | ١٣/٠٨ | ١٠/٤٧ | -۵/۶٤ | -٤/٣٢ | ١٣/۵٨ |
| ۶۷٨ | ٩٨٧۶٤ | ٩/۵٩ | ٨/٩ | -۵/۶٤ | ١٠/١٤ | ٢٢/٩٩ |
| ۵٤۶ | ۶٧٨٤۵ | ١٢/٠١ | -٤/٣٢ | ١٣/۵٨ | ١٣/۵٨ | ٣٤/٨۵ |
| ٢٣۵ | ٢٤٣٨٩ | ١٢/٠١ | ١۵/۵٤ | ١٣/۵٨ | ١٣/۵٨ | ۵٤/٧١ |
| ۵۶٨ | ٢٣٤١٤ | ١٢/٠١ | ١١/٤٩ | -۵/۶٤ | -۵/۶٤ | ١٢/٢٢ |
| ٢٩٠ | ٤١٤٣٢ | ٧/٣٣ | ١٣/٤٩ | -۵/۶٤ | -۵/۶٤ | ٩/۵٤ |
| ٧٨١ | ۵٤٨٧۶ | ٧/١٣ | ١٢/١٧ | ١٤/۵٤ | -۵/۶٤ | ٢٨/٢٠ |
| ٤۵٠ | ۵٩۶٨٧ | ١٠/٧٣ | ١٢/٠١ | ١٤/٠٠ | ١٠/٠۶ | ٤۶/٧٩ |
| ۶٢١ | ٤۶٧٨٩ | ١١/٤۵ | ١٠/٠۶ | ١٤/٠٠ | ١٣/۵٨ | ٤٩/٠٨ |

احتمال همسان‌های تصادفی با u نمایش داده می-شود. بر خلاف m که برای یک فیلد عددی ثابت محاسبه می‌شود، یک فیلد می‌تواند u های متعددی داشته باشد یا به عبارتی هر آیتم داده‌ای در یک فیلد میتواند یک احتمال u داشته باشد. به عبارتی u نسبت داده‌های با مقدار خاص بر اساس فراوانی مشاهده شده در منبع دقیق اولیه به کل مشاهدات می‌باشد. مثلاً اگر در یک منبع اطلاعاتی با ٣٠٠/٠٠٠ نفر جمعیت ثبت شده، ٣٠ نفر یک فامیل خاص (مثلا ایوبی)را داشته باشند، مقدار u برای آن فامیل خاص (ایوبی) برابر است با نسبت ٣٠ به ٣٠٠٠٠٠ و یا ۰/۰۰۰۱ می‌باشد. انتظار است که برای فیلد تاریخ تولد، هر روز تولد (یک آیتم داده) برای یک سال دارای احتمال u برابر ١ روی ٣۶٠ داشته باشد (با این فرض که توزیع تولد در روزهای مختلف سال یک نواخت باشد). به همین ترتیب آیتم‌های داده‌ای فیلدهای مختلف می‌توانند u های جداگانه‌ای داشته باشند. نمونه‌ای از مقدار u محاسبه شده برای سه آیتم اطلاعاتی از فیلدهای نام، نام خانوادگی، تاریخ تولد، و نام مادر، محاسبه شده بر مبنای داده‌های بانک مرگ و میر در تهران در جدول ٤ نمایش داده شده است. (در ارتباط دو بانک اطلاعاتی مقدار u بر اساس بانک اطلاعاتی که دقیق‌تر است و یا تعداد رکوردهای بیشتری دارد محاسبه می‌گردد).

### ٣-١-٢. تعداد همسان های صحیح مورد انتظار

عامل سوم که در پیوند احتمالی مؤثر است، تعداد همسان هایی است که انتظار می رود در دو بانک اطلاعاتی وجود داشته باشد. برای مثال در پیوند داده‌های مربوط به گواهی ولادت از سال ٢٠٠٤ با اطلاعات نقایص مادرزادی ثبت شده از سال ١٩٩۶، انتظار نداریم هیچ همسان صحیحی از دو سری اطلاعات به‌دست آوریم. در پیوند احتمالی، تعداد همسان‌های صحیح مورد انتظار با

**جدول ٥: محاسبه وزن برای فیلد برمبنای اطلاعات جدول قبلی بعلاوه وزن کل برای رکورد های ارتباط داده شده**

| شماره ردیف رکورد در بانک ثبت سرطان | شماره ردیف رکورد در بانک مرگ و میر | $W_i$نام | $W_i$فامیل | $W_i$تاریخ تولد | $w_i$نام مادر | وزن کل $w_t$ |
|---|---|---|---|---|---|---|
| ۷۶۵ | ۷۸۶۵۹ | ۱۳/۰۸ | ۱۰/٤۷ | -٥/٦٤ | -٤/۳۲ | ۱۳/٥۸ |
| ٦۷۸ | ۹۸۷٦٤ | ۹/٥۹ | ۸/۹ | -٥/٦٤ | ۱۰/۱٤ | ۲۲/۹۹ |
| ٥٤٦ | ٦۷۸٤٥ | ۱۲/۰۱ | -٤/۳۲ | ۱۳/٥۸ | ۱۳/٥۸ | ۳٤/۸٥ |
| ۲۳٥ | ۲٤۳۸۹ | ۱۲/۰۱ | ۱٥/٥٤ | ۱۳/٥۸ | ۱۳/٥۸ | ٥٤/۷۱ |
| ٥٦۸ | ۲۳٤۱٤ | ۱۲/۰۱ | ۱۱/٤۹ | -٥/٦٤ | -٥/٦٤ | ۱۲/۲۲ |
| ۲۹۰ | ٤۱٤۳۲ | ۷/۳۳ | ۱۳/٤۹ | -٥/٦٤ | -٥/٦٤ | ۹/٥٤ |
| ۷۸۱ | ٥٤۸۷٦ | ۷/۱۳ | ۱۲/۱۷ | ۱٤/٥٤ | -٥/٦٤ | ۲۸/۲۰ |
| ٤٥۰ | ٥۹٦۸۷ | ۱۰/۷۳ | ۱۲/۰۱ | ۱٤/۰۰ | ۱۰/۰٦ | ٤٦/۷۹ |
| ٦۲۱ | ٤٦۷۸۹ | ۱۱/٤٥ | ۱۰/۰٦ | ۱٤/۰۰ | ۱۳/٥۸ | ٤۹/۰۸ |

**جدول ٦: نحوه محاسبه احتمال برای ارتباط داده با استفاده از دو بانک اطلاعاتی**

| شماره ردیف رکورد در بانک ثبت سرطان | شماره ردیف رکورد در بانک مرگ و میر | xoبرای فیلد نول | Xiبرای نام | Xi برای نام خانوادگی | Xi برای تاریخ تولد | Xiبرای نام مادر | حاصلضرب Xi ها | احتمال (p) |
|---|---|---|---|---|---|---|---|---|
| ۷۶۵ | ۷۸۶۵۹ | ۰/۰۰۰۰۱۸ | ۸٦۳٦/۳٦ | ۱٤۱۷/۹۱ | ۰/۰۲ | ۰/۰٥ | ٤/۳٥ | ۰/۸۱۳۲۳۲۷۷۸ |
| ٦۷۸ | ۹۸۷٦٤ | ۰/۰۰۰۰۱۸ | ۷۷۲/۳٦ | ۱٥۸۳۳/۳۳ | ۲٤٥۰۰/۰۰ | ۱۱۲٦/٤٤ | ٥۳۲٦۳۱۳/۰۲ | ۰/۹۹۹۹۹۹۸۱۲ |
| ٥٤٦ | ٦۷۸٤٥ | ۰/۰۰۰۰۱۸ | ٤۱۳۰/٤۳ | ۰/۰٥ | ۱۲۲٥۰/۰۰ | ۱۲۲٥۰/۰۰ | ٤٤/۹۸ | ۰/۹۷۸۲٤۹٤۹٦ |
| ۲۳٥ | ۲٤۳۸۹ | ۰/۰۰۰۰۱۸ | ٤۱۳۰/٤۳ | ٤۷٥۰۰/۰۰ | ۱۲۲٥۰/۰۰ | ۱۲۲٥۰/۰۰ | ٤۲۷۲٦۲۹۳/٥٦ | ۰/۹۹۹۹۹۹۹۷۷ |
| ٥٦۸ | ۲۳٤۱٤ | ۰/۰۰۰۰۱۸ | ٤۱۳۰/٤۳ | ۲۸۸۰/۲۸ | ۰/۰۲ | ۰/۰۲ | ٤/۲۳ | ۰/۸۰۸۷۹۸۰٥۳ |
| ۲۹۰ | ٤۱٤۳۲ | ۰/۰۰۰۰۱۸ | ۱٦۱/۰۲ | ۱۱٥۲۱/۱۳ | ۰/۰۲ | ۰/۰۲ | ۰/٦٦ | ۰/۳۹۸۰۱٥٦٥٦ |
| ۷۸۱ | ٥٤۸۷٦ | ۰/۰۰۰۰۱۸ | ۱٤۰/٥۳ | ٤٦۰۸/٤٥ | ۲۳۷٦۹/۹۰ | ۲۳۷٦۹/۹۰ | ۲۷۳٦۷۱/۰۷ | ۰/۹۹۹۹۹٦۳٤٦ |
| ٤٥۰ | ٥۹٦۸۷ | ۰/۰۰۰۰۱۸ | ۱٦۹٦/٤۳ | ٤۱۳۰/٤۳ | ۱٦۳۳۳/۳۳ | ۱٦۳۳۳/۳۳ | ۲۰۳٤٥۸٥/٤۱ | ۰/۹۹۹۹۹۹٥۰۸ |
| ٦۲۱ | ٤٦۷۸۹ | ۰/۰۰۰۰۱۸ | ۲۷۹٤/۱۲ | ۱۰٦۷/٤۲ | ۱٦۳۳۳/۳۳ | ۱٦۳۳۳/۳۳ | ۸٦٦۰۰۹/۹۲ | ۰/۹۹۹۹۹۸۸٤٥ |

مقدار E نشان داده می‌شود. به‌عنوان مثال در یک شرایط فرضی که میزان بقا یک ساله برای بیماران مبتلا به سرطان معده به‌طور متوسط ۷۰ درصد باشد، انتظار می‌رود که اگر در طول سال از ۱۰۰ نفر بیمار ۳۰ نفر فوت نمایند یا به عبارتی اطلاعات ۳۰ درصد بیماران در بانک داده مرگ و میر موجود باشد (با فرض اینکه تمامی مرگ‌ها ثبت می‌شود) در این شرایط تعداد همسان‌های مورد انتظار ۳۰ مورد می‌باشد (اگر کل بیماران در سال ۱۰۰ نفر باشد).

**۲-۲.استراتژی های پیوند داده احتمالاتی**

هنگامی که رکوردهای بانک اطلاعاتی یک منبع با رکوردهای منبع دیگری پیوند داده میشود، تعدادی از رکوردها به‌عنوان همسان صحیح، تعدادی به‌عنوان همسان غلط و تعدادی از رکوردها در وضعیتی قرار می‌گیرند که همسانی و یا عدم همسانی آنها را نمی‌توان با درجه اطمینان بالائی مشخص نمود. با توجه به پیچیدگی‌های آماری و احتمالاتی که در پیوند داده‌ها وجود دارد به منظور کاهش درصد همسان‌های غلط و بالا بردن بهره‌وری بایستی استراتژی و راهکار مناسبی انتخاب نمود. استراتژی و راهکار های ارتباط داده شامل سه مرحله متفاوت ۱) بلوک بندی، ۲) تعیین شاخص همسانی و موافقت بین رکوردها (وزن دهی) و ۳) نحوه‌ی تصمیم‌گیری در مورد همسانی و موافقت رکوردها می‌باشد.

**۱-۲-۲. تکنیک‌های بلوک‌بندی**

در هنگام ارتباط داده‌ها از دو بانک اطلاعاتی روش معمول این است که هر رکوردی از یک بانک با کلیه رکوردهای بانک دیگر مقایسه گردد تا رکورد همسان انتخاب شود. این مسئله در عمل بسیار مشکل و در مجموعه اطلاعات بزرگ، غیر ممکن است. (تصور کنید

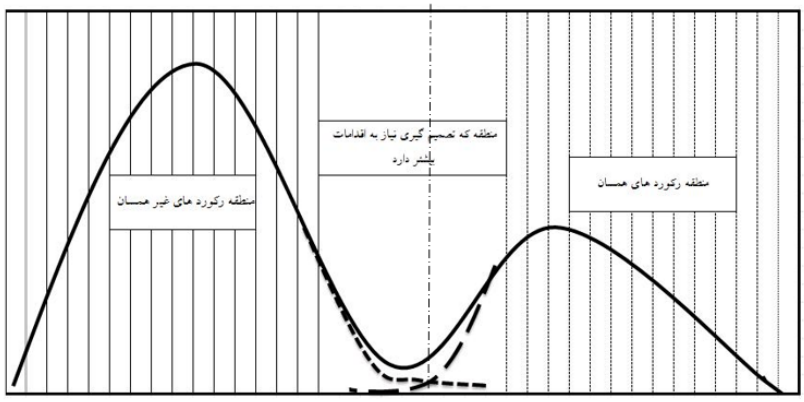
که دو بانک اطلاعاتی هرکدام شامل ۱۰۰۰۰۰ رکورد باشد، در این صورت هر رکورد از بانک اول بایستی با همه رکوردهای بانک دوم مقایسه گردد، یعنی تعداد مقایسه‌ها ۱۰ بیلیون خواهد شد). به‌منظور حل این مشکل در استراتژی ارتباط داده ها پیشنهاد می‌شود که ابتدا رکوردها بر مبنای شاخص‌هایی بلوک‌بندی شده و سپس ارتباط یک جفت رکورد فقط در بلوک‌های مرتبط که شامل تعداد کمتری از رکوردها می‌باشد، بر قرار گردد. در حقیقت با این روش به جای بررسی کلیه رکوردها در هر دو منبع، مقایسه تنها به یک زیر مجموعه کوچک از رکوردها محدود می‌شود و لذا حجم محاسباتی به میزان زیاد کاهش می‌یابد.

متغیرهایی که بلوک‌بندی بر مبنای آنها انجام می‌گردد، متغیرهای شناساگر (identifier variable) می‌باشند. به‌طور مثال اگر در ارتباط داده‌های دو بانک اطلاعات ثبت سرطان و اطلاعات مرگ و میر، بلوک‌بندی بر مبنای نام خانوادگی و به‌صورت نزولی (از الف تا ی) انجام گردد، برای نام خانوادگی ایوبی مقایسه محدود به رکوردهایی می‌شود که نام خانوادگی ایوبی دارند و به این ترتیب تعداد مقایسه‌ها به میزان زیادی کاهش پیدا میکند. بایستی دقت نمود که بلوک کردن در عین اینکه تعداد مقایسات را کاهش می‌دهد، در مواقعی که کیفیت داده‌ها مناسب نیست، ممکن است شانس موافقت واقعی بین رکوردهای موافق را کاهش دهد، مخصوصاً زمانی که یک فیلد نامناسب برای بلوک‌بندی انتخاب می‌گردد (به‌طور مثال اگر در جدول شماره ۲ ، بلوک‌بندی برمبنای " نام و یا نام خانوادگی صورت پذیرد شانس ارتباط و تطبیق بین رکورد شماره ۷٦۵ از بانک داده ثبت سرطان و رکورد شماره ۷۸٦۵۹ از بانک اطلاعات مرگ و میر وجود خواهد داشت ولی اگر تاریخ تولد مبنای بلوک‌بندی قرار گیرد شانس همسانی این دو رکورد از بین می‌رود زیرا روز تولد در دو بانک اطلاعاتی عدم همخوانی دارند، لذا شانس همسانی دو رکورد از بین می‌رود.

استراتژی بلوک‌بندی اگر با دقت و به‌طور مناسب انجام نگیرد ممکن است تعداد همسان‌های احتمالی را بسیار کاهش دهد. بلوک کردن می‌تواند بر مبنای یک فیلد و یا مجموعه‌ای از فیلدها صورت پذیرد. بایستی دقت شود که با افزایش فیلدهای بلوک شده می‌توان تعداد همسان‌های احتمالی را کاهش داد. در مجموع فیلدهای کاندید بلوک شدن فیلدهای شناساگری هستند که می‌بایست از کیفیت بالایی برخوردار باشند. بلوک کردن باید بر اساس ظرفیت سخت‌افزار و نرم‌افزار، نوع و کیفیت داده‌ها انجام گردد. در نهایت در مورد روش‌های بلوک‌بندی باید اشاره کرد که توسعه و گسترش قدرت برنامه‌های کامپیوتری، machine learning، داده کاوی مطالعات آماری بدون شک عملکرد صحت داده کاوی را بهبود بخشیده و در یافتن روش‌های بلوک‌بندی موثر و کارا کمک‌کننده خواهد بود. از جمله روش‌های جدید بلوک‌بندی که معرفی شده‌اند می‌توان به clustering algorithm-high dimensional indexing-stored neighborhood اشاره کرد (۲۰).

### ۲-۲-۲. تعیین شاخص همسانی و موافقت بین رکوردها (محاسبه وزن)

شانس همسانی و یا موافقت رکوردها بستگی مستقیم به تعداد فیلدهای همسان و غیر همسان در دو رکورد دارد. این شانس وقتی بالاست که تعداد فیلدهای همسان در دو رکورد بالا باشد. با توجه به اینکه در تطبیق بین رکوردها لزوماً تمامی فیلد ها همسان نیستند درجه تطبیق هم به فیلدهای همسان و هم غیر همسان بستگی دارد. برای اندازه‌گیری درجه تطبیق به شاخصی نیاز می‌باشد که بر مبنای آن قادر باشیم قدرت تطبیق بین دو رکورد را ارزیابی کنیم. بر این مبنا درجه تطبیق بین فیلدهای مشابه را با استفاده از احتمال m و u محاسبه میکنیم. در واژه‌شناسی ارتباط داده، شاخص تطبیق، وزن نامیده می‌شود که با حرف w نمایش داده می‌شود. فرمول شماره ۱ برای محاسبه وزن یک فیلد در دو رکورد همسان




شکل ۱: توزیع فراوانی وزن های محاسبه شده در ارتباط داده با روش احتمالی

و فرمول شماره ۲ برای محاسبه فیلدهای غیرهمسان استفاده می‌شود. (دقت شود که همسان بودن و نبودن فیلدها در اینجا مبنای عینی دارد).

فرمول شماره ۱ (وزن برای فیلد های همسان) پایه i نشانگر یک فیلد میباشد

$$w_i = log_2(\frac{m_i}{u_i})$$

فرمول شماره ۲ (وزن برای فیلد های غیر همسان )

$$w_i = log_2(\frac{1-m_i}{1-u_i})$$

به‌طور مثال در جدول شماره ۲ اگر رکورد شماره ٦٧٨ از بانک ثبت سرطان و رکورد ٩٨٧٦٤ از بانک مرگ و میر را به عنوان دو رکورد همسان در نظر بگیریم با توجه به احتمال m برای فیلد فامیل (۰/۹۵) و احتمال u برای فامیل "شکوفه" برابر ۰/۰۰۰۱۲ می‌باشد، وزن محاسبه شده برای فیلد همسان فامیل برابر ۸/۹ محاسبه میشود. برای این دو رکورد فیلد تاریخ تولد در دو رکورد همسان نیستند و برای محاسبه وزن این فیلد از فرمول شماره ۲ استفاده می‌کنیم. با توجه به اینکه احتمال m برای فیلد تاریخ تولد برابر است با ۰/۹۸ و احتمال u برای تاریخ تولد "۲۰۰۰/۲۷/۱" برابر است با ۰/۰۰۰۰۲ لذا وزن محاسبه شده برای فیلد غیر همسان تاریخ تولد برابر با ٥/٦٤- محاسبه میشود. بایستی دقت شود که در فیلدهای غیر همسان وزن برمبنای آیتم اطلاعات بانک اطلاعاتی بزرگتر که در اینجا بانک اطلاعات مرگ و میر با تعداد رکوردهای چندین برابر بانک اطلاعات ثبت سرطان محاسبه می‌گردد.

الگوریتم Expectation Maximization (EM) یک رویکرد تکرارشونده برای برآورد احتمالات m و u می‌باشد. البته در مواقعی به خوبی عمل می‌کند که احتمال خطاهای تایپوگرافیکال در متغیرهای شناساگر کم باشد (۲۱). رویکرد دیگر stored-neighborhood نام دارد که منابع داده بر اساس ترکیب‌های مختلفی از شناساگرهای در دسترس مرتب می‌شوند. در هر ترکیب همه رکوردهای درون یک window of n-record با همدیگر مقایسه می‌شوند (۲۲).

در نهایت برای همه فیلدهایی که مبنای مقایسه قرار گرفته‌اند $w_i$ محاسبه شده و بسته به تعداد فیلدهای مورد استفاده (k) از یک تا $w_k$ تغییر می‌کند و برای هر همسانی ممکن، وزن همه فیلدها جمع و یک وزن کل با استفاده از فرمول شماره ۳ محاسبه می‌گردد. نمونه‌ای از این محاسبات در مورد ۸ جفت همسان در جدول ۵ نشان داده شده است.

فرمول شماره ۳ (وزن کل برای رکورد های همسان)، k بیانگر تعداد فیلد های استفاده شده در استراتژی ارتباط میباشد

$$w_i - \sum_{i=1}^{k} w_i$$

### ۳-۲-۲. مبانی تصمیم گیری برای تعیین همسان های صحیح، ناصحیح و نا مشخص:

مقادیر بالاتر وزن کل $w_t$ نشانگر صحیح‌تر بودن همسانی و مقادیر پایین‌تر نشانه ناصحیح بودن آن است. اما با توجه به وابستگی $w_t$ به تعداد و ماهیت فیلدهای شناساگر در ارتباط داده ها، دامنه تغییرات آن بسیار متفاوت بوده و نمی‌توان مرز مشخصی را برای تعیین همسانی یا ناهمسانی ارتباطات تعیین نمود و نیاز به اقدامات دیگری برای این امر می‌باشد. در صورتی که نمودار توزیع $w_i$ ها رسم گردد همانند شکل ۱، مشاهده

می گردد که نوع توزیع، دو نمائی بوده به‌طوری‌که قسمت اول نمودار (خطوط پیوسته در شکل) مربوط به مقادیر پایین $w_t$ و همسان‌های ناصحیح و قسمت دوم مربوط به مقادیر بالای $w_t$ و همسان‌های صحیح (خطوط نقطه چین در شکل) می‌باشد. بر اساس این توزیع لازم است مقداری از $w_t$ مشخص شود که در مقادیر بالاتر از آن نسبت همسان های ناصحیح به صحیح، بسیار ناچیز باشد که این دامنه به‌عنوان «همسان‌های صحیح» نامگذاری می‌شوند. به همین ترتیب دامنه پایین $w_t$ تعیین گردد به طوری‌که نسبت همسان‌های صحیح به ناصحیح بسیار ناچیز باشد که همان دامنه «همسان های ناصحیح» می‌باشد. بدیهی است مقادیری که در محدوده بین دو مقدار فوق قرار می-گیرند نیاز به استفاده از متدهای دیگر مثل تطبیق دستی و مراجعه به تک تک رکوردها و کنکاش بیشتر می‌باشد.

یکی از محدودیت‌های wt این است که دامنه و توزیع تغییرات آن با توجه به تعداد فیلدها، داده‌های مختلف و استراتژی تطبیق متفاوت و متغیر می‌باشد. برای این منظور سعی می‌شود که وزن را تبدیل به احتمال نمود تا تفسیر آن آسان گردد. برای این منظور احتمال همسانی بر اساس فرمول شماره ٤ که شانس مضرب مقادیر Xi میباشد محاسبه می‌گردد. در فرمول شماره ٤ مقدار Xi برای هر فیلد در صورت همسانی با فرمول شماره ٥ و در صورت عدم همسانی با فرمول شماره ٦ محاسبه می‌شود. فرمول شماره ٤ پایه i نشانگر فیلد میباشد شامل فیلد نول (Null) هم می‌شود

$$P=\frac{\prod_{i=0}^{k} x_i}{\prod_{i=0}^{k} x_i+1}$$

فرمول شماره ٥

$$X_i = \frac{m_i}{u_i}$$

فرمول شماره ٦

$$X_i = \frac{1-m_i}{1-u_i}$$

در این فرمول علاوه بر تعداد فیلدهای شناساگر (i=1 to k)، یک فیلد نول که با Xi=0 نشان داده می‌شود اضافه شده که برابر است با شانس اینکه دو رکورد بطور اتفاقی در برنامه ارتباط داده همسانی کامل داشته باشند و این مقدار بر اساس فرمول شماره ٧ محاسبه می‌گردد.

فرمول شماره ٧

$$X_0 = \frac{E}{N_1 N_2 - E}$$

در فرمول شماره ٧، N1 و N2 تعداد رکورد های موجود درهر فایل و E تعداد رکوردهای همسان مورد انتظاردر دو فایل می‌باشد. (به‌طور مثال در ارتباط بانک داده ثبت سرطان و بانک مرگ و میر، N1 تعداد ٤٥٠٠٠رکورد مربوط به بانک مرگ‌و میر و N2 برابر با ٣٠٠٠ رکورد از بانک ثبت سرطان است که در این صورت مقدار E برابر با ٢٤٠٠ می‌باشد زیرا بر اساس دانش قبلی ٨٠ درصد (مثلاً سرطان ریه) موارد بروز سرطان منجر به فوت می‌شود. بایستی دقت شود که معمولاً در برنامه ارتباط داده مخصوصاً در اپیدمیولوژی بیماری‌ها به‌عنوان پیامد نادر بوده و بنابر این مقدار Xi=0 خیلی کوچک می‌باشد). جدول ٦ نحوه محاسبه احتمال همسانی را برای دو بانک اطلاعاتی نشان می‌دهد.

**٣-٢. ارزیابی کیفیت انجام پیوند داده**

هدف از پیوند داده‌ها پیدا کردن همسان‌ها می-باشد. شکل ١ به‌صورت شماتیک یک توزیع دو قله‌ای نمرات وزن کلی همسان و غیر همسان در یک پروژه پیوند داده نشان می‌دهد. در واقعیت این امکان وجود ندارد که تعیین کنیم کدام زوج مقایسه همسان و یا غیر همسان است. ما فقط تعداد ترکیب شده زوج مقایسه ها برای هر وزن کلی خاص مشاهده می‌کنیم. در یک پیوند داده به‌دنبال تعیین نقطه برش‌هایی هستیم که نقاط بالاتر از آن را به‌عنوان پیوند و پایین‌تر از آن را به‌عنوان غیر لینک طبقه‌بندی کنیم. امیدواریم که اکثریت پیوندها همسان باشند (مثبت واقعی) و تعداد ناچیزی از همسان‌ها گم شده باشند (منفی کاذب).

بر اساس جدول ٢ در ٢ زیر می‌توان عملکرد یک پیوند داده را در طبقه‌بندی پیامد محاسبه کرد.

| | همسان | غیر همسان |
|---|---|---|
| پیوند شده | a<br>مثبت واقعی | b<br>مثبت کاذب |
| پیوند نشده | c<br>منفی کاذب | d<br>منفی واقعی |

حساسیت (sensitivity): $\frac{a}{a+c}$

ویژگی (specificity): $\frac{d}{d+b}$

ارزش اخباری مثبت (positive predictive value (PPV)): $\frac{a}{a+b}$

ارزش اخباری منفی (negative predictive value (NPV)) : $\frac{c}{c+d}$

به‌دلیل اینکه تعداد زیادی از همسان‌های بالقوه در طول فاز بلوک‌بندی شناسایی می‌شوند، یک حجم زیادی از فضا به غیرهمسان‌های واقعی اختصاص پیدا می‌کند برای این دلیل نشان داده شده شاخص‌هایی شامل غیرهمسان‌های واقعی مانند ویژگی و ارزش اخباری منفی حالت چوله پیدا می‌کنند و به‌جای آن توصیه شده است که از شاخصی بنام f-measure استفاده شود. این شاخص بیانگر میانگین هارمونیک حساسیت و ارزش اخباری مثبت است که از تعداد زیاد غیر همسان‌های واقعی تأثیر نمی‌پذیرد و به‌صورت زیر محاسبه می‌شود

$$\frac{((\beta^2+1)\times sensitivity \times PPV))}{(\beta^2 \times sensitivity + PPV)}$$

ارزش بتا بیانگر اهمیت حساسیت نسبت به ارزش اخباری مثبت است. اگر وزن برابری دارند پس ارزش بتا برابر یک می‌باشد. و یا اگر احساس می‌شود که حساسیت دو برابر وزن ارزش اخباری مثبت می‌باشد اندازه بتا برابر دو تعیین می‌شود.

در پیوند داده دو خطا وجود دارد: خطای نوع اول که یک غیرهمسان واقعی به‌عنوان همسان طبقه بندی می‌شود و خطای نوع دوم که یک همسان واقعی به‌عنوان غیر همسان طبقه‌بندی می‌شود. این پارامترها بستگی زیادی به وزن نقطه برش دارد. حرکت به سمت چپ در شکل ۱، حساسیت را افزایش می‌دهد اما مثبت کاذب را افزایش می‌دهد. حرکت به سمت راست حساسیت را کاهش می‌دهد اما همچنین تعداد مثبت کاذب کاهش می‌یابد. هنگامی که پیوند داده برای تعیین پیامد در یک مطالعه کوهورت استفاده شود، خطاهایی که در طی پیوند داده اتفاق می‌افتد روی تحلیل‌های همبستگی مواجهه و پیامد تأثیرگذار است. مثبت کاذب اتفاق افتاده در طی پیوند داده باعث تورش در اندازه‌های اثر مانند نسبت خطر و تفاوت خطر شده و آنها را به سمت ارزش نول می‌برد تا زمانی که ویژگی بر حسب مواجهه غیر افتراقی می‌باشد (۲۳). اثر منفی کاذب در طی پیوند داده باعث کمتر از حد نشان داده تفاوت خطر می‌شود و نسبت خطر تا زمانی حساسیت بر حسب مواجهه غیر افتراقی است بدون تغییر می‌ماند (۲٤). بنابراین هنگامی که نیاز است یک حالت تعادل بین تعداد منفی کاذب و مثبت کاذب انجام گیرد یک استراتژی معمول قربانی کردن حساسیت برای داشتن ویژگی بالا می‌باشد. با این استراتژی نسبت خطر در مطالعه کوهورت بدون تورش باقی می‌ماند اما قدرت آماری آن کاهش می‌یابد (۲۵).

استراتژی دیگر نسبت خطر و تفاوت خطر مشاهده شده برای تورش سوء طبقه‌بندی پیامد که در طی پیوند داده انجام می‌گیرد تطبیق داده شود به‌طوری که می‌توان از طریق حساسیت، ویژگی و ارزش اخباری مثبت به‌عنوان پارامترهای تورش طی تحلیل تورش سوء طبقه‌بندی را تصحیح کرد (۲٦). کاهش تعداد پیوندهای مثبت های کاذب نیازمند این است که در ابتدا تعداد آنها از طریق وزن کلی نقطه برش تعیین شده و تعیین این نقطه نیازمند یک تصمیم آگاهانه بر مبنای یک استاندارد طلایی است. برای مثال در مطالعه مربوط به داده‌های مربوط به ایدز برای یک نمونه از افرادی که نام آنها معلوم است به‌عنوان یک منبع معتبر برای بانک اطلاعاتی بزرگ استفاده شده است (۲۷). در غیر این‌صورت و در غیاب یک داده معتبر به عنوان استاندارد طلایی باید از روش‌هایی مانند probabilistic bias analysis استفاده کرد (۲۸). در نهایت باید اشاره کرد که مرورهای سیستماتیک نشان داده‌اند که

خیلی از متغیرها روی اینکه فرآیند پیوند داده می‌تواند با خطا همراه باشد تأثیرگذار است. از جمله این متغیرها می‌توان به سن، جنس، گروه‌های نژادی و قومیتی، منطقه جغرافیایی، وضعیت اقتصادی اجتماعی و وضعیت سلامتی می‌توان اشاره کرد (۲۹).

## نتیجه‌گیری

ارتباط داده ها بر اساس احتمال، ابزاری قوی برای دست اندرکاران بهداشت جامعه و محققین علاقه‌مند به مصورسازی وضع سلامت جامعه بر اساس داده‌های جمعیتی فراهم می‌سازد. علی‌رغم پیچیدگی بیشتر شیوه ارتباط داده‌ها بر مبنای احتمال نسبت به روش قطعی، تطبیق احتمالی به‌علت کاهش تعداد رکوردهای ناهمسان که ناشی از ناهماهنگی در ثبت رکوردها است، می‌تواند روش مفید و مناسبی باشد. جهت ارزیابی دقیق تورش ناشی از خطاهای موجود در پیوند داده‌ها لازم است شاخصی از کیفیت پیوند داده‌ها مثل میزان مثبت کاذب یا منفی کاذب، اندازه‌گیری و گزارش شود و در مراحل پیشرفته تر ارزش اخباری مثبت و منفی هر استراتژی مشخص شده و مد نظر قرار گیرد.

## References


1. Newcombe HB, Kennedy JM, Axford S, James AP. Automatic Linkage of Vital Records Computers can be used to extract" follow-up" statistics of families from files of routine records. Science. 1959;130(3381):954-9.
2. Schouten LJ, Schlangen JT, de Rijke J ,Verbeek AL. Evaluation of the effect of breast cancer screening by record linkage with the cancer registry, the Netherlands. J Med Screen. 1998;5(1):37-41.
3. Goldacre M, Abisgold J, Yeates D, Vessey M. Benign breast disease and subsequentbreast cancer: English record linkage studies. J Public Health. 2010;32(4):565-71.
4. Risch HA, Howe GR. Menopausal hormone usage and breast cancer in Saskatchewan: a record-linkage cohort study. Am J Epidemiol. 1994;139(7):670-83.
5. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked Medicare-tumor registry database. Med Care. 1993; 31(8):732-48.
6. Clark DE, Hahn DR. Comparison of probabilisticand deterministic record linkage in the development of a statewide trauma registry. Proc Annu Symp Comput Appl Med Care. 1995: 397–401.
7. Jamieson E, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. Methods Inf Med. 1995;34(4):371-7.
8. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol. 2002;31(6):1246-52.
9. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. A practical method of linking data from Medicare claims and a comprehensive electronic medical recordssystem. Int J Med Inform. 2003;71(1):57-69.
10. Krewski D, Dewanji A, Wang Y, Bartlett S, Zielinski J, Mallick R. The effect of record linkage errors on risk estimates in cohort mortality studies. Survey Methodology. 2005;31(1): 13-21.
11. Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Services Research. 2006;6(1):48.
12. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J. 2009;157(6):995-1000.
13. Jacobs JP, Edwards FH, Shahian DM, Haan CK, Puskas JD, Morales DL, et al. Successful linking of the Society of Thoracic Surgeons adult cardiac surgery database to Centers for Medicare and Medicaid Services Medicare data. Ann Thorac Surg. 2010;90(4):1150-7.
14. Li Q, Glynn RJ, Dreyer NA, Liu J, Mogun H, Setoguchi S. Validity of claims-based definitions of left ventricular systolic dysfunction in Medicare patients. Pharmacoepidem Dr S. 2011;20(7):700-8.
15. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkageoutperforms deterministic record linkage. J Clin Epidemiol. 2011;64(5):565-72.
16. Howe HL, Lake AJ, Shen T. Method to assess identifiability in electronic data files. Am J Epidemiol. 2007;165(5):597-601.
17. Dusetzina S ,Tyree S, Meyer A, Meyer A, Green L, Carpenter W. Linking Data for Health Services Research: A Framework and Instructional Guide.Agency for Healthcare Research and Quality (US); 2014.

18. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Med Informatics and decision making. 2013;13(1):64.
19. Mason CA, Tu S. Data linkage using probabilistic decision rules: A primer. Birth Defects Research Part A: Clinical and Molecular Teratology. 2008;82(11):812-21.
20. Nicoletta C, Tiziana T. Statistical Perspective on Blocking Methods When Linking Large Data-sets. Studies in Theoretical and Applied Statistics. 2012.
21. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990.
22. Belin TR, Rubin DB . method for calibrating false-match rates in record linkage. Journal of the American Statistical Association. 1995;90(430):694-707.
23. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977;105(5):488-95.
24. Rodgers A, Walker N, Schug S, McKee A, Kehlet H, Van Zundert A, et al. Reduction of postoperative mortality and morbidity with epidural or spinal anaesthesia: results from overview of randomisedtrials. Bmj. 2000;321(7275):1493.
25. Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev. 1998;20(1):112-21.
26. Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. Am J Epidemiol. 1993;138(11):1007-15.
27. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Stat Med. 1995;14(5-7):499-509.
28. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data: Springer Science & Business Media; 2011.
29. Megan A Bohensky DJ, Vijaya Sundararajan, Sue Evans, David V Pilcher, Ian Scott, Caroline A Brand. Data Linkage: A powerful research tool with potential problems. BMC Health Services Research. 2010;10:346.

**Review Article**

# Probabilistic record linkage methodology: a review article

***Erfan Ayubi***
Ph.D Candidate of Epidemiology, School of Medicine, Zabol University of Medical Sciences, Zabol, Iran
Ph.D Candidate of Epidemiology, Department of Epidemiology, School of Public Health, Tehran University of Medical Science,Tehran, Iran

***Kamyar Mansori***
Ph.D Candidate of Epidemiology, School of Medicine, Kurdistan University of Medical Sciences, Sanandaj, Iran
Ph.D Candidate of Epidemiology, Department of Epidemiology, School of Public Health, Iran University of Medical Science, Tehran, Iran

***Mohammad Golmahi***
Cancer Research Center, Tehran University of Medical Sciences

***Ozra Ramezankhani***
Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences

***Alireza Mosavi-Jarrahi***
Department of Social Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences

**Received:**28/08/2015**, Revised:**03/11/2015**, Accepted:**18/12/2015

**Corresponding Author:**
Alireza Mosavi-Jarrahi
Department of Social Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences
E-mail: rmosavi@yahoo.com

## Abstract

Research development and information technology progress lead to generate big dataset with valuable information. In health research, with tracing people from different dataset like registries can provide valuable information about prognosis, prediction, discrimination, detection or etiology for many outcomes without establishing costly studies. Extracting the knowledge from this potential information is applied using advanced methods such as data linkage or record linkage with deterministic or probabilistic algorithm. However, probabilistic linkage is computationally complex and not well understood by many researchers who may wish to apply it in their work. Therefore, the purposes of this review article is to introduce probabilistic record linkage methodology such as quality and standardization of dataset, determining the matching records from different dataset, calculating the matching weights and discrimination matched from unmatched record using a cut point. In follow, with a practical example the probabilistic record linkage methodology is introduced by cancer registry and mortality dataset.

**Keywords:** *Data linkage, Probabilistic Algorithm, Cancer registry, Mortality*