روشهای مبتنی بر گراف در کاوش الگو های ترتیبی دارای فاصله زمانی

هدی معمارزاده[1]، دکتر محمد رضا خیام باشی[2]، دکتر محمد حسین سرایی[3]

چکیده:

کشف الگوهای ترتیبی پر بسامد از مهمترین روشهای داده کاوی محسوب می شود. الگوی استخراج شده، معرف زنجیره ای از رویدادهای ثبت شده در زمانهای مختلف است و برای پیش بینی عملکرد سایر موجودیت هایی که رفتار مشابهی دارند استفاده شود. الگوریتم هایی که در این رابطه ارائه شده است در دو گروه طبقه بندی می شود: در گروه اول روشهای مبتنی بر کاوش قوانین وابستگی قرار دارند و گروه دوم شامل روشهایی است که با استفاده از ساختارهای منظم، روابط بین الگوها را نشان می دهند. از جمله کاراترین ساختارهای مورد بررسی در گروه دوم گراف است که در آن نودها معرف رویدادها بوده و یالها بیان کننده ارتباط بین آنها هستند. این ارتباط که شامل ترتیب اتفاق افتادن رویدادهاست می تواند در بردارنده اطلاعاتی درخصوص فاصله نسبی بین رویداد های مختلف نیز باشد. در این مطالعه دو روش تولید گراف برای نمایش الگوهای ترتیبی پر بسامد مورد بررسی قرار گرفته است. تفاوت این دو، در توصیف قوانین رابطه ای زمانی بین رویداد هاست. در روش اول، مفاهیم ارائه شده در جبر Allen در توصیف رابطه زمانی بین دو رویداد به کارگرفته شده و از داده های زماندار برای تولید این قوانین استفاده می شود. در روش دوم که نوع توسعه یافته ای از الگوریتم Apriori است رابطه زمانی بین دو رویداد با استفاده از محدوده های زمانی از پیش تعیین شده توسط کاربر ارائه و الگوهای موجود از طریق جستجوی گراف استخراج می شوند. برای هر روش از مثال کاربردی استفاده شده است. هدف اصلی ، استخراج پربسامدترین دنباله های علائم بالینی مشاهده شده ای است که در نهایت به تشخیص یک بیماری خاص منجر شده اند. در این تحقیق نشان داده شده که گراف تولید شده از طریق الگوریتم توسعه یافته Apriori به دلیل داشتن اطلاعات بیشتر درباره فاصله زمانی از قابلیت پیش بینی بالاتری برخوردار است.

کلمات کلیدی:

الگوهای ترتیبی پر بسامد، قوانین رابطه ای زمانی، فاصله زمانی، روشهای جستجوی گراف

Graph base approaches in mining Time interval sequence patterns

Hoda Meamarzadeh, Mohammad Reza Khayyambashi, Mohammad Hussein Saraee

Abstract:

Frequent Sequence Pattern is one of the most important data mining domains. Each extracted pattern is a chain of events that observed in deferent times and can be used in predicting the manner of similar entities. Generally, two principal approaches are utilized to extract sequential patterns: methods based on association rule mining and techniques based on using regular structures. One of the most efficient structures in second group is Directed Graph. In denoting frequent sequence pattern, events are presented by vertexes and edges are expressing the relations between them. This relation is about ordering the events and can include more detail about relational time space between two events. In this paper two methods for generating graph are studied. The difference between these two approaches is in expressing time interval between events and temporal relational rules. First method extract temporal interval relation rules from temporal interval data by using Allen's theory. Second one is a generation of Apriori algorithm and uses of predefined ranges for presenting time interval between events. For clarify the characteristics of both methods we use of example. In this example we try to discover frequent chains of medical symptoms that finally lead to detection of risk full maladies. This study shows that second approach can represents more information about time interval between events, therefore is more useful to prediction.

Key Words:

Frequent sequence pattern, Time interval, temporal relational rules, Graph search technique

[1] دانشجوی کارشناسی ارشد نرم افزار ، دانشگاه آزاد اسلامی واحد نجف آباد-اصفهان

[2] عضو هیئت علمی گروه کامپیوتر- دانشکده فنی مهندسی – دانشگاه اصفهان

[3] عضو هیئت علمی گروه کامپیوتر- دانشکده برق و کامپیوتر – دانشگاه صنعتی اصفهان

1. مقدمه

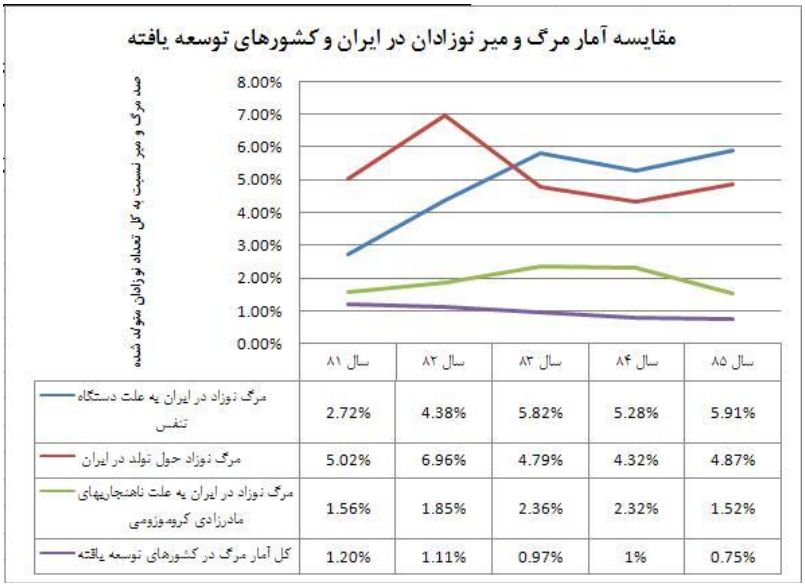
داده کاوی زمانی یکی از مهمترین روشهای استخراج دانش از میان مجموعه های داده ی زماندار بوده و شامل استخراج دنباله های ترتیبی[1]، دنباله های مشابه از نظر زمان[2] چرخه ها و قوانین وابستگی[3] می باشد. استخراج دنباله های ترتیبی که نوع خاصی از کشف قوانین وابستگی محسوب می شود، روشی است برای یافتن دنباله هایی از آیتم ها ، که درون مجموعه تراکنشها به دفعات تکرار شده اند. مساله اصلی در یافتن دنباله های ترتیبی، استخراج طولانی ترین دنباله هایی است که تعداد دفعات تکرار آنها از میزان خاصی که توسط کاربر تعیین می شود بزرگتر یا مساوی باشد. تا کنون الگوریتم های بسیاری در زمینه کشف دنباله های ترتیبی پر بسامد ارائه شده است که از آن جمله می توان به الگوریتم Apriori اشاره کرد. این الگوریتم که به صورت اول سطح به جستجوی زیر دنباله های پر بسامد می پردازد خود دارای انشعاباتی است که عملکرد آن را از نظر مدت زمان اجرا و یا حافظه مورد نیاز بهبود می بخشند. با وجود اهمیت کشف دنباله هایی از رویدادها که در صورت وقوع متوالی می توانند به نتایج از پیش تعیین شده منجر شوند، دانستن فاصله زمانی بین رخداد این رویدادها نیز می تواند در بسیاری از موارد از اهمیت بسیاری برخوردار باشد. بررسی فاصله زمانی بین رویدادهای موجود در دنباله های متواتر از جمله مواردی است که به تازگی مورد بررسی و تحقیق قرار گرفته است و نتایج حاصل از آن می تواند در بالابردن کیفیت قوانین تولید شده موثر باشد. در این مطالعه دو روش کشف دنباله های پر بسامد ارائه شده است که هر کدام با داشتن تعاریف متفاوتی از فاصله زمانی به تولید قوانینی پرداخته اند که در آنها زمان نه تنها به عنوان عاملی برای تعیین ترتیب رویدادها بکار رفته بلکه خود به عنوان بخشی از قانون تولید شده و نیز یکی از خروجی های فرایند داده کاوی محسوب می شود. نتایج حاصل از هر دو روش به صورت گراف بیان شده است. در قسمت دوم این مقاله به ارائه توضیحاتی در خصوص سیستم داده ای مورد مطالعه پرداخته شده و در قسمت های سوم و چهارم به ترتیب روش مبتنی بر تئوری Allen و سپس روش مبتنی بر الگوریتم Apriori مورد بحث قرار گرفته است . در انتها نتایج بدست آمده از هر دو روش مورد مقایسه قرارا گرفته اند.

2. معرفی مورد مطالعه

کشف دنباله های پربسامد در پیش بینی رفتار بعدی موجودیت های مورد مطالعه مفید است. مدلهایی که تا کنون این گونه الگوریتمها در آن به کار رفته اند، مدیریت سبد خرید مشتریان و بررسی صفحات پیمایش شده توسط کاربران اینترنتی است. در این مقاله به بررسی استفاده از الگوریتمهای استخراج دنباله های پر بسامد در یک پایگاه داده پزشکی پرداخته ایم. این پایگاه داده شامل اطلاعات استخراج شده از پرونده- های پزشکی مادران بارداری است که در طول دوران بارداری به طور متناوب مورد معاینه پزشکی قرار گرفته اند و علائم متعددی از آنها در هر بار ویزیت پزشکی ثبت شده است. در انتخاب پرونده ها بروز ناهنجاری های حین تولد در مادر یا نوزاد مورد نظر قرار گرفته است.

طبق آمار ارائه شده از طرف معاونت بهداشت استان اصفهان اصلی ترین عوامل مرگ و میر نوزادان زیر یکسال را می توان در سه گروه نشان داده شده در شکل(1) دسته بندی کرد. همانطور که نشان اده شده است مرگ و میر حین تولد یکی از دلایل اصلی محسوب شده و متاسفانه رقم مربوط به شاخص MMR (Maternal mortality ratio) نیز در ایران در مقایسه با آمار رسمی منتشر شده کشورهای در حال توسعه بسیار بالاست. این آمار رسمی که توسط سازمان جهانی بهداشت WHO و به تفکیک عوامل مختلف منجر به مرگ به صورت سالانه برای کشورهای مختلف منتشر می شود [1] نشان می دهد که این شاخص در کشورهای توسعه یافته زیر 1% درصد است.

در ویرایش دهم طبقه بندی بین المللی بیماریها (ICD-10) [2] مرگ مادر (Maternal death) به مرگی که ناشی از بارداری بوده ویا در اثر بارداری و یا مراقبت ها و درمانهای انجام شده برای آن ایجاد شده ولی ناشی از تصادف وسانحه نباشد اطلاق شده است. با توجه به تعاریف مرگ مادری به دوگروه مستقیم وغیر مستقیم تقسیم می شود: مرگ مادری مستقیم و مرگ مادری غیر مستقیم. در گروه اول، مرگ مادری ناشی از عوارض بارداری، زایمان وپس از آن بوده وبه علت مداخلات، بی توجهی، درمان نامناسب ویا مجموعه ای‌از عوامل فوق رخ می دهد. ولی در گروه دوم مرگ مادری ناشی از وجود بیماری قبلی زمینه ای و یا بیماری است که در دوران بارداری به علت تاثیرات فیزیولوژیک بارداری‌تشدید گردیده است. هدف این مطالعه بررسی عوارض دسته بندی شده در گروه اول است و سعی دارد با کشف دنباله هایی از علائم بالینی که در نهایت منجر به بروز ناهنجاری های گروه اول شده اند راه حلی برای تشخیص زودرس افرادی که در معرض خطرند ارائه شود[1].



	سال ۸۱	سال ۸۲	سال ۸۳	سال ۸۴	سال ۸۵
مرگ نوزاد در ایران به علت دستگاه تنفس	2.72%	4.38%	5.82%	5.28%	5.91%
مرگ نوزاد حول تولد در ایران	5.02%	6.96%	4.79%	4.32%	4.87%
مرگ نوزاد در ایران به علت ناهنجاریهای مادرزادی کروموزومی	1.56%	1.85%	2.36%	2.32%	1.52%
کل آمار مرگ در کشورهای توسعه یافته	1.20%	1.11%	0.97%	1%	0.75%

شکل(2-1) بررسی میزان فراوانی عوامل منجر به مرگ نوزادان درفاصله سالهای 81 تا 85

3. استفاده از تئوری آلن در کاوش قوانین نسبی دارای فاصله زمانی به فرم گراف

در این بخش به بررسی تکنیکی برای کاوش قوانین نسبی دارای فاصله زمانی می پردازیم. این تکنیک از دو بخش عمده تشکیل شده است. قسمت اول شامل تولید رویدادهایی با فاصله زمانی با استفاده از داده های اولیه می باشد. پایگاه تولید شده در قسمت اول به عنوان ورودی قسمت دوم مورد استفاده قرار گرفته و قوانین نسبی دارای فاصله زمانی با استناد به مفاهیم ارائه شده در تئوری Allen استخراج می شوند در ادامه این بخش به توضیح هر یک از مراحل خواهیم پرداخت[3,5,7].

3.1 تعاریف اولیه

در این قسمت به بررسی تعاریف اولیه مورد استفاده در الگوریتم پرداخته می شود. مثالهایی که برای تبیین بیشتر موضوع بکاررفته در رابطه با موضوع مورد مطالعه است.

مجموعه رویدادها:

به مجموعه مواردی اتلاق می شود که امکان ثبت و مشاهده زیر مجموعه ای از آنها در هر تراکنش وجود دارد. به عنوان مثال مجموعه همه کالاهای درون فروشگاه. در مثال مورد مطالعه مجموعه مورد نظر شامل 25 علامت بالینی مورد بررسی در هنگام معاینه بیمار است. با توجه به اینکه فرایند داده کاوی یک فرایند وابسته به کاربرد است در این قسمت توضیحی در خصوص انواع مختلف این علائم داده می شود. در حوزه پزشکی علائم بالینی را می توان به دو گروه گسسته و پیوسته تقسم کرد. علائم پیوسته به آن دسته از ویژگی هایی گفته می شود که ثبت بروز یا عدم بروز آنها به تنهایی کافی است و نیازی به اطلاعات اضافه تر مثل مقدار و شدت ندارند. به عنوان مثال احساس تشنگی یا وجود سرگیجه. اما در مورد برخی دیگر از علائم باید شدت و مقادیر مربوطه هم ثبت شوند. که از این جمله می توان به وزن و فشار خون بیمار اشاره نمود. در ثبت علائم گسسته به مواردی که بروز یک حالت غیر نرمال را نشان می دهد پرداخته شده است و در ثبت علائم پیوسته بازه مقادیر قابل انتخاب به زیر بازه هایی تقسیم شده و هر زیر بازه به عنوان یک علامت گسسته در نظر گرفته می شود. به عنوان مثال اگر فشارخون سیستولیک یک فرد سالم را 80 میلیگرم جیوه در نظر بگیریم چنانچه فشار خون بیماری بین 90 تا 110 میلیگرم جیوه اندازه گیری شود علامت فشار خون بالا و چناچه از 110 بیشتر اندازه گیری شود علامت فشار خون خیلی بالا برای او ثبت می شود.

تراکنش:

هر تراکنش به مجموعه رویدادهایی گفته می شود که در رابطه با یک موجودیت در یک زمان خاص ثبت شده اند. یک رویداد فقط یکبار در هر تراکنش ظاهر می شود اما می تواند در چند تراکنش که در زمانهای مختلف و در رابطه با یک موجودیت واحد ثبت شده اند حضور داشته باشد. در

مثال ما رویدادها مجموعه علائم بالینی مشاهده شده در هربار ویزیت بیمار هستند. با در نظر گرفتن اینکه برای هر ویزیت پزشکی یک زمان منحصر بفرد در نظر گرفته می شود لذا هر تراکنش را می توان به صورت TR= (PID, t, S) نشان داد. در این تعریف PID شماره شناسایی بیمار مورد نظر بوده، t زمان ثبت تراکنش و S مجموعه در بردارنده علائم ثبت شده بیماری در آن نوبت ویزیت محسوب می شود. به عنوان مثال TR1= {P10, 01/06/09, {S1, S2, S3} به این معنا است که بیمار شماره 10 در تاریخ اول ژوئن سال 2009 ویزیت شده و در نتیجه این معاینه بروز علائم S2،S1 و S3 مشاهده و ثبت شده اند.

زمان تراکنش:

در رابطه با زمان تراکنش با مفهوم واحد زمان مواجهیم. واحد زمان می تواند بسته به نوع کاربرد یک ثانیه ، یک دقیقه ، یک روز و مواردی از این دست باشد. با توجه به اینک این مطالعه دررابه با مادران باردار انجام گرفته زمان تراکنش را سن بارداری با واحد یک هفته در نظر گرفته ایم. در نتیجه دو تراکنش برای دو بیمار مختلف که اولی در زمان 21 هفته و 3 روز و دومی در زمان 21 هفته و 5 روز صورت گرفته باشند با زمان 21 ثبت خواهند شد. شایان ذکر است به صورت معمول معاینات مورد بحث با فواصل یک الی دو هفته انجام شده و همین مساله باعث می شود که احتمال وجود دو تراکنش برای یک بیمار با زمان یکسان وجود نداشته باشد.

دنباله رویدادها:

همانطور که اشاره شد با فرض اینکه هر تراکنش در یک زمان منحصر بفرد رخ داده است، یک رویداد می تواند در چند تراکنش ظاهر شود. به عنوان مثال در ویزیت های دوم، سوم و پنجم برای یک بیمار علامت فشار خون بالا ثبت شده است. برای هر علامت نوعی اولین زمان مشاهده آن که در واقع زمان ثبت اولین تراکنش دربردارنده آن علامت می باشد با $s_i.vs$ نشان می دهیم زمانی که $s_i \in S$ باشد. و آخرین زمان مشاهده علامت را که مربوز به زمان ثبت آخرین تراکنش دربردارنده آن علامت است با $s_i.ve$ نمایش می دهیم. به عنوان مثال اگر پایگاه داده اولیه شامل تراکنش های DB = {T1 = (101,{D,B},1), T2 = (101, {D,A,E},2), T3 = (101,{A,E,B}, 3), T4 = (101, {C},5), T5 = (101,{E,C},7)} باشد دنباله های مختلفی که برای علائم موجود در این تراکنشها و در رابطه با بیمار شماره 101 می توان تصور کرد عبارتند از: SS(PID=101) = {[(D,1)(D,2)], [(B,1)(B,3)], [(A,2)(A,3)], [(E,2)(E,3)(E,7)], [(C,5)(C,7)]}. زمانی که SS به مهنای مجموعه دنباله ها باشد.

رویداد با فاصله زمانی:

هر رویداد یا علامت بالینی که در مجموعه دنباله ها عنوان شده است به صورت (s, [vs, ve]) نشان داده می شود. به عنوان مثال دنباله مربوط به رویداد E که به صورت [(E,2)(E,3)(E,7)] می باشد قابل تبدیل به رویداد E با فاصله زمانی [2,7] است که آن را به صورت (E,[2,7]) نشان می دهیم.

روابط حاکم بر فواصل زمانی:

روابط حاکم بر فواصل زمانی قابل تعریف بین دو رویداد x و y را می توان به صورت زیر تعریف نمود: $R(x, y) = \{P(x, y) | (x, y) \in \Omega, P \in IO\}$. زمانی که IO = {before, equals, meets, overlaps, during} . برای تعریف این واژگان که در توصیف فاصله زمانی بین دو رویداد به کار گرفته می شوند از تئوری Allen استفاده شده است. مفاهیم مرتبط در جدول(1) ارائه شده است.

Relation	Definition	Abbreviation
Before (x; y)	x. ve < y. vs	B
Equals (x; y)	(x. vs=y. vs) ^ (x. ve=y. ve)	E
Meets (x; y)	x. ve = y. vs	M
Overlaps (x; y)	(x. vs < y. vs) ^ (x. ve > y. vs)	O
During (x; y)	(x. vs > y. vs) ^ (x. ve < y. ve)	D

جدول (1) روابط فاصله زمانی ممکن بین دو رویداد x و y

3.2 مرحله اول: تولید رویدادهایی با فاصله زمانی با استفاده از داده های اولیه

بخش اول شامل توضیحاتی در خصوص نوع پایگاه داده ورودی و مراحل تبدیل آن به پایگاه رویدادهای داری فاصله زمانی می باشد:

3.2.1 پایگاه داده اولیه

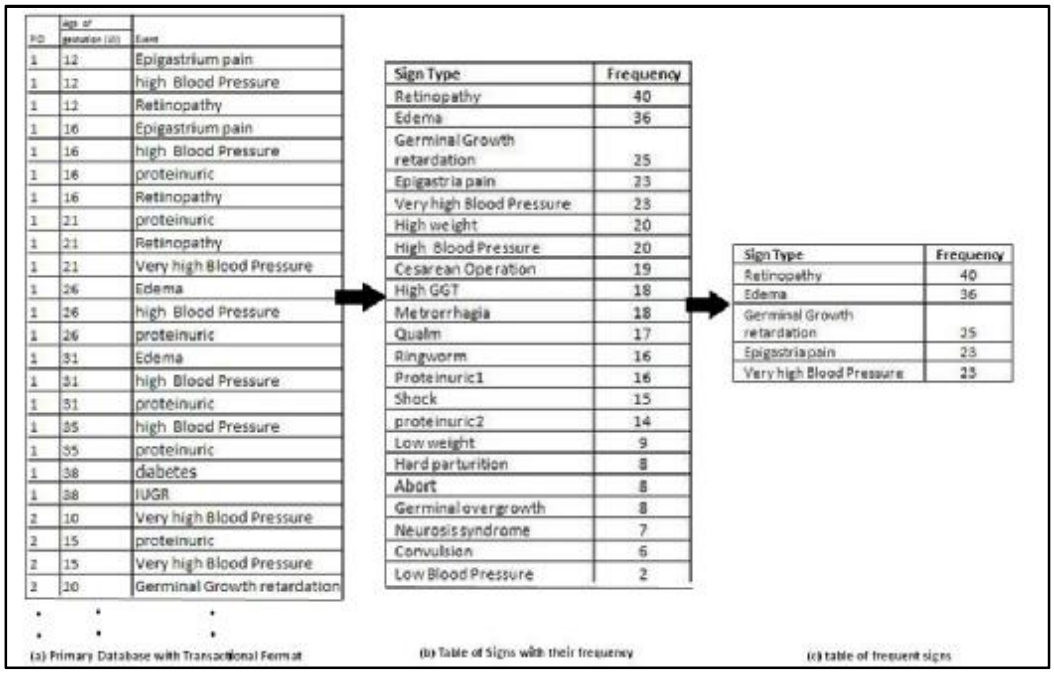
اطلاعات ورودی این قسمت پایگاه داده ای از تراکنش هاست. فرم ذخیره سازی پایگاه داده اولیه به صورت تراکنشی است در این فرم هر ردیف از اطلاعات در بردارنده سه فیلد اطلاعاتی خواهد بود:(شماره شناسایی بیمار، زمان ثبت تراکنش، علامت بالینی مشاهده شده) به تعداد علائم بالینی ثبت شده برای هر بار ویزیت بیمار ردیف هایی با شماره شناسایی بیمار و زمان ویزیت یکسان در پایگاه ایجاد می کنیم و سپس این پایگاه را براسا شماره شناسایی بیمار و سپس زمان ثبت ویزیت مرتب می کنیم.

3.2.2 استخراج رویدادهای پر بسامد

در مرحله بعد با توجه به کمترین مقدار Support (min-sup) که از طرف کاربر تعیین شده است به اسخراج رویدادهای پر بسامد می پردازیم. برای هر علامت باین مورد بررسی مقدار support برابر خواهد بود با تعداد بیمارانی که این علامت بالینی را در طول دوران مشاهده از خود بروز داده اند . نکته حائز اهمیت این است تکرار یک علامت بالین در دفعات مختلف ویزیت یک بیمار تاثیری در میزان support آن علامت نخواهد داشت. پس از تعیین support برای تمام علائم بالنی مورد بررسی با توجه به min-sup علائمی که غیر متواتر را حذف می کنیم

3.2.3 تولید رویدادهایی دارای فاصله زمانی

پس از انتخاب رویدادهای پر بسامد برای تولید رویدادهای دارای فاصله زمانی به این صورت عمل می شود که برای هر رویداد در حوزه تراکنش های ثبت شده برای هر بیمار مورد بررسی قرار گرفته و اولین زمان و نیز آخرین زمان مشاهده این رویداد استخراج می شود. اولین زمان را با vs و آخرین زمان را با ve نشان می دهیم. شکل (2) بیانگر مراحل صورت گرفته در قسمت اول الگوریتم است. در قسمت a از شکل (2) یک پایگاه داده اولیه از ویزیت هایی که به فرم تراکنشی ذخیره شده اند نشان داده شده است. در قسمت b بسامد هر علامت بالینی بر اساس تعداد بیمارانی که آن علامت را بروز داده اند محاسبه شده و در قسمت c با فرض اینکه میزان min-sup برابر 40% موارد غیر متواتر حذف شده اند. در پایان قسمت اول الگوریتم با پایگاه داده ای از علائم بالینی پر بسامد مواجهیم که اولین و آخرین زمان مشاهده هر کدام از این علائم بالینی برای هر بیمار تعیین شده است.



PID	age of gestation (d)	Event
1	12	Epigastrium pain
1	12	high Blood Pressure
1	12	Retinopathy
1	16	Epigastrium pain
1	16	high Blood Pressure
1	16	proteinuric
1	16	Retinopathy
1	21	proteinuric
1	21	Retinopathy
1	21	Very high Blood Pressure
1	26	Edema
1	26	high Blood Pressure
1	26	proteinuric
1	31	Edema
1	31	high Blood Pressure
1	31	proteinuric
1	35	high Blood Pressure
1	35	proteinuric
1	38	diabetes
1	38	IUGR
2	10	Very high Blood Pressure
2	15	proteinuric
2	15	Very high Blood Pressure
2	10	Germinal Growth retardation

(a) Primary Database with Transactional Format

Sign Type	Frequency
Retinopathy	40
Edema	36
Germinal Growth retardation	25
Epigastria pain	23
Very high Blood Pressure	23
High weight	20
High Blood Pressure	20
Cesarean Operation	19
High GGT	18
Metrorrhagia	18
Qualm	17
Ringworm	16
Proteinuric1	16
Shock	15
proteinuric2	14
Low weight	9
Hard parturition	8
Abort	8
Germinal overgrowth	8
Neurosis syndrome	7
Convulsion	6
Low Blood Pressure	2

(b) Table of Signs with their frequency

Sign Type	Frequency
Retinopathy	40
Edema	36
Germinal Growth retardation	25
Epigastria pain	23
Very high Blood Pressure	23

(c) table of frequent signs

شکل 2 تبدیل پایگاه داده اولیه به پایگاه رویدادهای دارای فاصله زمانی

3.3 مرحله دوم: استخراج قوانین دارای فاصله زمانی

پس از ایجاد پایگاه داده رویدادهای دارای فاصله زمانی نوبت به تولید قوانین دارای فاصله زمانی می رسد. این قوانین با استفاده از گراف جهتدار نمایش داده می شوند. روش کار به این صورت است که با استفاده از مفاهیم ارائه شده در جدول(1) که بر اساس تئوری Allen می باشد برای هر بیمار روابط نسبی موجود بین رویدادهای پر بسامد مشاهده شده برای وی استخراج می شود. این کار با استفاده از مقایسه بین زمان شروع و پایان هر رویداد نسبت به زمانهای شروع و پایان سایر رویدادهایی که برای همان بیمار مشاهده شدهاند انجام می پذیرد. شکل (4) بیانگر این مرحله برای بیمار P33 است.

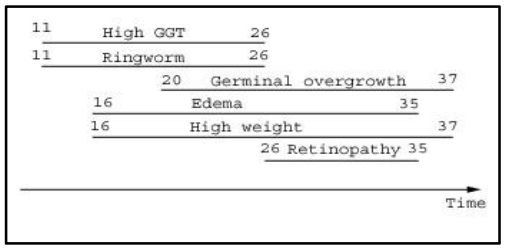
پس از تولید روابط نسبی برای تمام بیماران در مرحله بعد متواتر ترین قوانین نسبی دارای فاصله زمانی با استفاده از ساختار درخت استخراج می شود. همانطور که در شکل (5) نشان داده شده گره های این درخت علائم بالینی محسوب شده و یالها روابط نسبی تشخیص داده شده بین آن

علائم هستند بنابراین با حرکت روی جدول شامل روابط نسبی مواردی که در درخت نیستند به آن اضافه می شوند. چنانچه رابطه ای قبلا در درخت نشان داه شده است با مشاهده موارد جدید با عدد مربوط به support آن رابطه که توسط نودهای انتخابی نگهداری می شود یک واحد اضافه می شود.

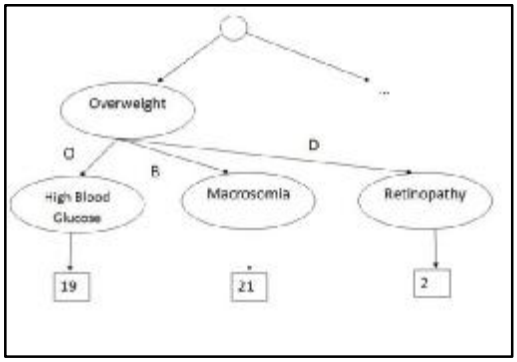
قوانین استخراج شده به صورت گراف جهتدار قابل نمایش هستند. میزان اهمیت هر رابطه با تقسیم عدد support به تعداد کل روابط استخراج شده به دست آمده و عددی بین صفر تا 1 خواهد بود که آن را با درصد نمایش می دهیم. بسته به حداقل میزان support که از طرف کاربر تعین شده است برخی از یالهای این گراف به دلیل احتمال کم وقوع حذف می شوند. ولی در مثال مورد نظر و با توجه به اهمیت درنظرگرفتن تمام شرایط ممکن برای بروز یک حالت پرخطر تمام روابط بدست آمده مورد ارزیابی قرار گرفت. در شکل 6 و 7 کد SQL مربوط به تعیین نوع رابطه زمانی و نیز دسته بندی هر گروه از روابط آوره شده است. شکل 8 نمایش دهنده گراف مورد بحث بوده و در شکل 9 بخشی از جدول استفاده شده در رسم گراف نمایش داده شده است.

PID	Start Point	End Point	Sign
33	11	26	High GGT
33	11	26	ringworm
33	16	35	Edema
33	16	37	High weight
33	20	37	Germinal overgrowth
33	26	35	Retinopathy
34	11	32	High GGT
34	11	25	Retinopathy
34	11	25	ringworm
34	20	38	Germinal overgrowth
34	20	38	high weight
34	32	38	Edema
36	11	31	high weight
36	11	20	Retinopathy
36	20	38	Edema
36	20	35	high Blood Pressure
36	26	38	Germinal overgrowth
36	26	35	High GGT
41	7	27	ringworm
41	33	33	shock
41	22	33	high Blood Pressure
41	22	33	Low weight
41	27	33	High GGT

شکل 3 پایگاه داده رویدادهای دارای فاصله زمانی



شکل 4 رابطه نسبی رویدادهای دارای فاصله زمانی



شکل 5 درخت روابط نسبی دارای فاصله زمانی

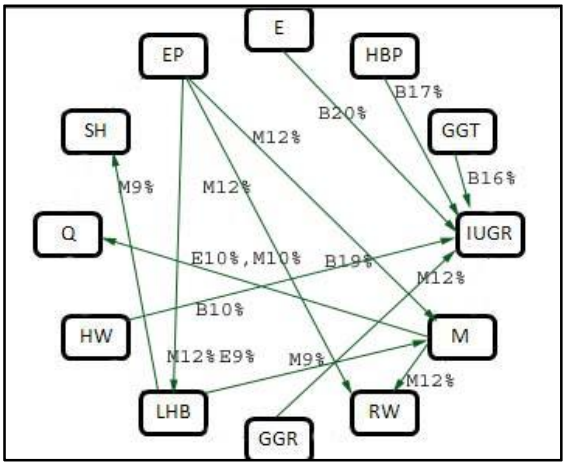
```
SELECT SecoundAllen.S, Count(SecoundAllen.S)
AS CountOfS
FROM SecoundAllen
GROUP BY SecoundAllen.S;
```

شکل 6 کد QL مربوط به دسته بندی روابط زمانی

```
INSERT INTO SecoundAllen ( PID, S )
SELECT P.PID, (P.Sign+' DURING  '+Q.Sign)
FROM PrimeryAllen AS P, PrimeryAllen AS Q
WHERE P.PID=Q.PID And P.Sign<Q.Sign And

P.START>Q.START
AND
P.END <Q.END
```

شکل 7 کد SQL مربوط به تعیین روابط زمانی بین رویداد ها



شکل 8 گراف روابط نسبی دارای فاصله زمانی

S	Count
Edema BEFORE IUGR	20
diabetes EQUALS IUGR	20
diabetes MEETS IUGR	19
high Blood Pressure BEFORE IUGR	17
High GGT BEFORE IUGR	16
Metrorrhagia MEETS ringworm	12
Germinal Growth retardation MEETS IUGR	12
Epigastrium pain MEETS ringworm	12
Epigastrium pain EQUALS Metrorrhagia	12
Epigastrium pain MEETS Low Heart-Beat in a Chrysalis	12
high weight BEFORE IUGR	10
Metrorrhagia EQUALS qualm	10
Metrorrhagia MEETS qualm	9
Epigastrium pain EQUALS Low Heart-Beat in a Chrysalis	9
Metrorrhagia EQUALS ringworm	9
Low Heart-Beat in a Chrysalis MEETS shock	9
Low Heart-Beat in a Chrysalis MEETS Metrorrhagia	9

شکل 9 جدول مربوط به روابط زمانی استخراج شده به همراه درصد مشاهده آنها

4. استفاده از مدل توسعه یافته الگوریتم Apriori برای تولید گراف

در ادامه این تحقیق از نوع توسعه یافته ای از الگوریتم Apriori برای کاوش دنباله های پر بسامد دارای فاصله زمانی استفاده می کنیم ، نتیجه این بخش به صورت گرافی ارائه می شود که نودهای آن به منزله رویدادها و یالها بیانگر ارتباطات بین این رویدادها هستند. در این مطالعه فاصله های زمانی مختلف به صورت محدوده های از پیش تعیین شده توسط کاربر مشخص می شود. در ابتدای این بخش به توضیح مختصری دربـاره خاصـیت Apriori پرداخته شده و در ادامه چگونگی توسعه آن جهت کاوش دنبایه های پر بسامد دارای فاصله زمانی توضیح داده می شود[5,6,7].

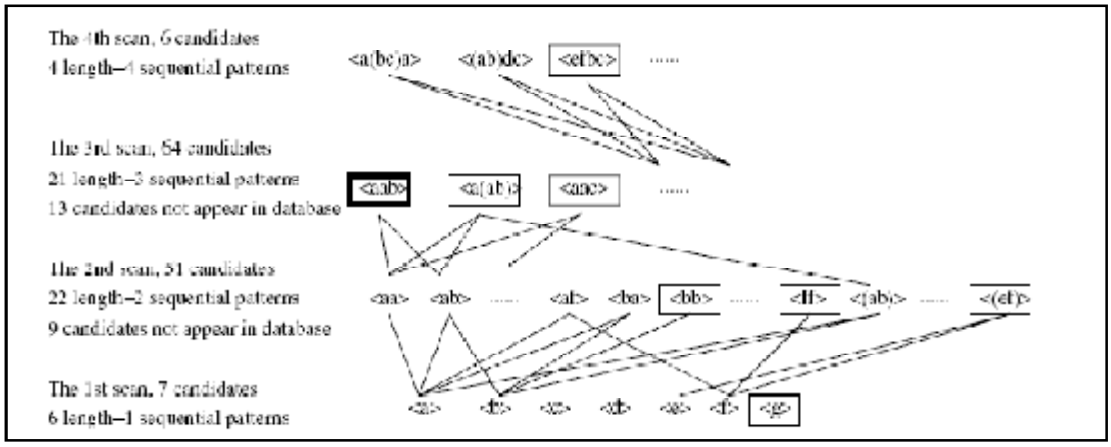
4.1 خاصیت Apriori

الگوهای دنباله ای از یک خاصیت یکنواختی برخوردارند. یکنواختی به این معنا است که اگر (ab)dc یک الگو محسوب شود تمام زیر مجموعه های آن هم که عبارتند از a, b, d, c, (ab), ad, ac, bd, bc, dc, (ab)d, (ab)c, adc, , bdc, نیز الگو محسوب خواهند شد. در واقع خاصیت Apriori یـا استقرایی به این معناست که اگر s یک دنباله و s' زیر دنباله ای از آن باشد می توان ادعا نمود که support(s)≤support(s') است. ساده ترین مدل الگوریتم Apriori که به دلیل استفاده از خاصیت استقرایی به این نام شناخته شده با استفاده از یک جستجوی اول سطح و نیز با در نظـر گـرفتن ایـن اصل عمل می کند:" اگر دنباله ای به عنوان الگو شناخته نشده و یا شرایط شناخته شدن به عنوان الگو را ندارد نیازی به بررسی دنبالـه هـای بزرگتـری که دربردارنده آن هستند نخواهد بود". با داشتن جدولی از تراکنشها مثل جدول (2) و نیز مقدار تعیین شده توسـط کـاربر (min-sup)، الگـوریتم بـه صورت زیر عمل می کند. در هر مرحله از دو مجموعه با نام های L_i وC_i استفاده می شود. مجموعه C_i در بردارنده تمام دنباله هایی به طـول i خواهـد بود که قرار است بر اساس میزان support آنها تعیین شود که آیا قابلیت تعیین شدن به عنوان الگو را دارند یا نه؟ و مجموعه L_i دربردارنده آن دسـته از دنباله هایی به طول i است که بر اساس مقدار Support به عنوان الگو شناخته شده اند.

در ابتدا دنباله های به طول 1 (length-1 subsequences) استخراج شده و میزان support هر یک محاسبه می شود. : a : 4, b : 4, c : 3, d : 3, e : 3, f : 3, g : 1. و مواردی که support آنها از min-sup کمتر است حذف می شوند.مجموعه باقیمانده عبارتسـت از L1 = {a, b, c, d, e, f}, موارد باقیمانده مبنای دنباله هایی به طول 2 (length-2 sequential)خواهند بود. بدین ترتیب ابتدا تمام ترکیبات مختلف به طـول 2 ایجـاد شـده و سپس موارد غیر متواتر و یا مواردی که اصلا دیده نشده اند حذف می گردد. در این مثال , C2 = {aa, ab, . . . , af, ba, bb, . . . , ff, (ab), (ac), . . . , (ef)}. خواهد بود که از میان این 51 عضو تنها 22 دنباله به عنوان الگوهایی به طول 2 شناخته می شوند (9 دنباله C2 در جدول تراکنشها دیده نشده اند و 20 دنباله هم support لازم را برای گذر ازاین مرحده ندارند.) . استقرای بکارگرفته شده در این الگوریتم به مراحل بعدی هم قابل تعمیم است. به این صورت که دنباله ای به طول k زمانی به عنوان الگو شناخته می شود که تمام زیر دنباله های آن به طول k-1 در مرحله قبل به عنوان الگـو شـناخته شده باشند.

جدول (2) جدول تراکنش. در این جدول فقط تقدم و تاخر رویدادها نشان داده شده است.

P۱	a(abc)(ac)d(cf)
P۲	(ad)c(bc)(ae)
P۳	(ef)(ab)(df)cb
P٤	eg(af)cbc



شکل 10 مراحل اولیه تولید دنباله ها با استفاده از استقرا

4.2 تولید گراف بر اساس خاصیت Apriori

برای تولید گراف جهتدار نیاز به برقراری یال هایی است که رابطه بین گره ها را نشان دهد. هر یال بین دو گره رسم می شود. بنابراین با در دست داشتن رابطه بین هر دو گره می توان گرافی تولید کرد که رابطه بین تعداد بیشتری از گره های آن از طریق پیمایش گراف قابل تعیین است. با این فرض، خاصیت Apriori می تواند در تولید چنین گرافی بکاربرده شود. به این ترتیب که ابتدا به ایجاد مجموعه L_1 و سپس با استناد به آن به تولید مجموعه L_2 پرداخته خواهد شد و سپس با استفاده از الگوهای دوتایی استخراج شده از مرحله دوم روابط ممکن بین گره ها تشخیص داده شده و گراف ایجاد می شود. برای یافتن دنباله هایی که تعداد عناصر آنها سه تا یا بیشتر باشد می توان با پیمایش گراف، تمام مسیر های ممکن به طول مورد نظر را استخراج کرده و با توجه به میزان support الگوهایی به طول مطلوب را تعیین کرد. در ادامه مطلب به توضیح این روش پرداخته خواهد شد:

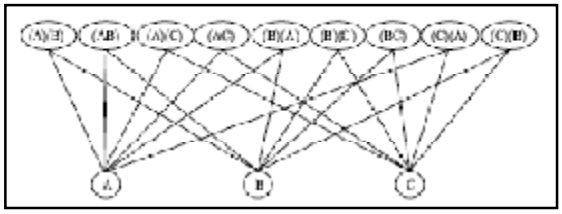
4.2.1 جستجوی جدول تراکنشها برای یافتن الگو هایی به طول 1

در گام اول به جستجوی جدول تراکنشها با هدف یافتن الگوهایی به طول 1 خواهیم پرداخت. نکته حائز اهمیت این است که ساختار جدول دربردارنده تراکنشها مثل الگوریتم قبلی به صورت تراکنشی TR= (PID, t, S) می باشد. این ساختار اجازه می دهد که برای یافتن الگوهای به طول 1 فقط نیاز به یکبار جستجوی جدول وجود داشته باشد. (جستجوی جدول به خصوص زمانی که تعداد رکوردهای افزایش یافته باشد زمانگیر بوده و اگر الگوریتم طوری باشد که نیاز به چند بار جستجو در آن مشاهده شود، کارایی بالایی نخواهد داشت). میزان support هر $s_i \in S$ را تعداد PID های تشکیل میدهد که درتمام تراکنشهای خود حداقل یکبار شامل s_i بوده باشند. این قسمت از الگوریتم مشابه قسمت اول از الگوریتم Allen است در نتیجه با این حجم از داده های جمع آوری شده و نیز با در نظر گرفتن مقدار support برابر با 25% مجموعه علائمی که بسامد آنها بیش از 18 بار بوده است این علام در شکل (2) قسمت دوم قابل مشاهده اند. به عنوان متواتر یک عضوی استخراج می شوند.

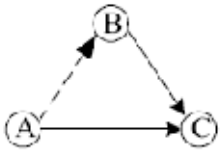
4.2.2 تولید الگوهای به طول 2

از طریق اتصال الگوهای پربسامد یک عضوی موجود در L_1 میتوان به مجموعه دنباله های دوعضوی C_2 دست یافت. با داشتن دو الگو به طول 1 مثل A,B دونوع الگوی به طول 2 می توان تولید نمود. نوع اول به صورت(AB) نوشته شده و به آن هیبرید یا ترکیبی گفته می شود. به این معناست که A و B هر دو متعلق به تراکنش مربوز به یک PID بوده و در یک زمان رخ داده اند . نوع دوم به صورت (B) , (A) نشان داده می شود و به این معناست که رویدادهای A و B هر دو متعلق به تراکنش های یک PID هستند اما در دو زمان متفاوت ثبت شده اند . رویداد A در این مورد قبل از رویداد B ثبت

شده است. به عنوان مثال اگر سه عضو A, B و C را داشته باشیم تمام زیر مجموعه های دو عضوی به صورت (A)(B), (AB), (A)(C),(AC), (B)(A), (B)(C), (BC), (C)(A)و(C)(B) خواهند بود. شکل (10) . برای نشان دادن <(A),(B)> از ⟵ و برای <(A,B)> از --> استفاده می شود. شکل (۱۱).



شکل 10 نحوه تولیددنباله های دو عضوی



شکل 11 روش نمایش انواعدو نوع رابطه بین دو گره گراف

با توجه به اینکه پایگاه داده اولیه به صورت تراکنشی ذخیره شده و هر رکورد آن به صورت {PID,t,S} است برای تولید L2 از الگوریتم زیر می توان استفاده نمود.شکل (12) . با ذکر این نکته که بعد از گذر از مرحله اول پایگاه داده اولیه فقط شامل علائم پر بسامد بوده و رکورد های در بردارنده علائم دیگر حذف شده اند. در قسمت اول از کد نشان داده شده در شکل (12) مجموعه هایی به صورت (AB) و در قسمت دوم مجموعه هایی به صورت (A),(B) استخراج می شود maxgap حداکثر فاصله زمانی معتبر بین دو رویداد است که از طرف کاربر تعیین می شود. بعد از تولید تمام زیر مجموعه های دوعضوی از عناصر L1 میزان support هر زیر مجموعه را با محاسبه تعداد دفعاتی که آن زیر مجموع در جدول تراکنشها دیده شده محاسبه خواهیم کرد. در شکل (13) کد مربوط به حذف موارد غیر متواتر نشان داده شده است. حاصل اجرای این کد تولید مجموعه C2 خواهد بود.

```
INSERT INTO dbo.L2 (START,PID, STime,ETime)
SELECT (P.Sign +','+ Q.Sign),P.PID,P.Time,Q.Time
FROM PrimaryTable P,PrimaryTable Q
WHERE P.PID=Q.PID
AND P.TIME=Q.TIME
AND P.Sign < Q.Sign
UNION
SELECT (P.Sign +','+ Q.Sign),P.PID,P.Time,Q.Time
FROM PrimaryTable P,PrimaryTable Q
WHERE P.PID=Q.PID
AND P.TIME<Q.TIME
AND  (Q.TIME-P.TIME) <6
AND P.Sign < Q.Sign
```

شکل 12 . کد SQL تولید مجموعه L2

```
ETE FROM L2
RE START NOT IN
LECT START FROM L2
UP BY START
ING COUNT(*) >20)
```

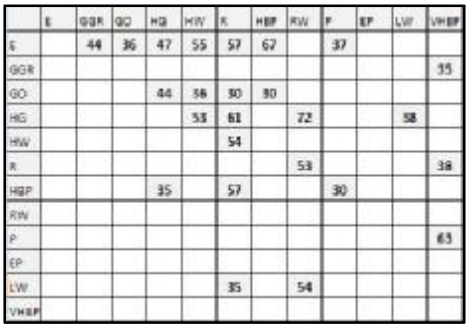
شکل 13 . کد SQL تولید مجموعه C2

ساختار مجموعه L2 به صورت {Sequence,PID,St,Et} خواهد بود که در آن Sequence مجموعه دو عضوی تولید شده در مرحله دوم است، PID به شماره بیمار اشاره کرده و St زمان رخداد اولین عضو از مجموعه Sequence و Et زمان رخداد دومین عنصر خواهد بود..

در این تحقیق با استناد به اطلاعات 120 پرونده پزشکی مربوط به مادران باردار که به صورت غیر نرمال دوران بارداری خود را به اتمام رسانده اند، جدول اولیه (Primary Table) دربردارنده 1000 تراکنش به صورت {PID,t,S} می باشد. با استفاده از کد SQL نشان داده شده در شکل(12) ، در ابتدا 2221 رکورد در جدول L2 وارد شده و سپس با استفاده از مقدار 20 برای min – sup تعداد 850 تراکنش که از میزان بسامد مورد نظر برخوردار نبوده اند از جدول L2 حذف می شوند . در شکل 11 بخشی از مجموعه های دوتایی باقیمانده به همراه میزان بسامد آنها نشان داده شده است. در کل در این مرحله 36 مجموعه 2 تایی تولید شد که میزان بسامد آنها بین 21 تا 72 مورد متغیر است. در این مجموعه های 2 تایی در مجموع 12 نوع مختلف از علائم بالینی قابل مشاهده است . که بر این اساس می توان ماتریس مجاورت را برای این گروه از علائم به صورت شکل (14) تولید کرد. بر اساس ماتریس مجاورت گراف شکل (15) بدست می آید.

Edema,Germinal Growth retardation	٤٤
Germinal overgrowth,High GGT	٤٤
Edema,High GGT	٤٧
High GGT,high weight	٥٣
Retinopathy,ringworm	٥٣
high weight,Retinopathy	٥٤
Low weight,ringworm	٥٤
Edema,high weight	٥٥
Edema,Retinopathy	٥٧
high Blood Pressure,Retinopathy	٥٧
High GGT,Retinopathy	٦١
Edema,high Blood Pressure	٦٧
High GGT,ringworm	٧٢

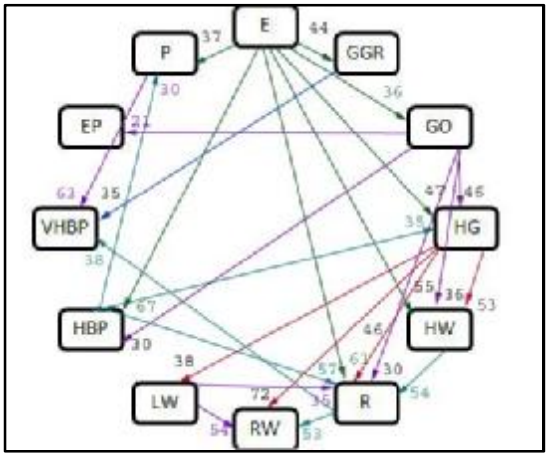
جدول (3) مجموعه های دوتایی به همراه بسامد



شکل14 ماتریس مجاورت

Edema	E
Germinal Growth retardation	GGR
Germinal overgrowth	GO
High GGT	HG
High weight	HW
Retinopathy	R
Ringworm	RW
Low weight	LW
High Blood Pressure	HBP
Very high Blood Pressure	VHBP
Epigastrium pain	EP
Proteinuric	P

جدول(4) علائم اختصاری بکاررفته در گراف



شکل 15 گراف علائم بالینی

در گراف علائم بالینی نشان داده شده در شکل (15) اطلاعات قابل مشاهده به تقدم و تاخر رویدادها و نیز بسامد رخداد هر زیر مجموعه دوتایی محدود است. و به عبارتی چنانچه مجموعه قوانین از روی این گراف تولید شود عباراتی نظیر « اگر Edema (ورم دست و پا) رخ دهد آنگاه با احتمال 20% می توان انتظار داشت که بیمار علائم Retinopathy (اختلال بینایی) را نیز مشاهده می کند.» تولید خواهد شد. در واقع این نوع قوانین اطلاع اضافه تری در خصوص زمان نسبی مشاهده Retinopathy بعد از مشاهده Edema گزارش نمی کند. برای حل این مشکل به روش زیر عمل می شود.

4.2.3 مفهوم فاصله زمانی و نحوه استفاده از آن در تولید گراف

بررسی مفهوم فاصله زمانی در افزایش ارزش اطلاعات تولید شده توسط گراف موثر است. در الگوریتم قبلی فاصله زمانی بین بروز دو رخداد با استفاده از مفاهیم مبتنی بر تئوری Allen و با واژگانی مثل before, equals, meets, overlaps, during بیان می شد. روش دیگری که برای بیان فاصله زمانی بکار می رود استفاده از محدوده های زمانی از پیش تعیین شده است. اگر t متغیر تصادفی مربوط به فاصله زمانی، r تعداد بازه های فاصله زمانی و k طول بازه باشد، مجموعه $T_k = \{I_0; I_1; I_2;...; I_r\}$ به صورت شکل(16) بیان خواهد شد. ستون سوم مثالی است که در آن طول بازه 3 واحد زمانی در نظر گرفته شده و تعداد بازه ها 4 بازه است[9,10,11].

نام	تعریف	مثال (k=3,r=4)
I_0	t=0	t = 0
I_1	$0 < t \le T_1$	0 < t ≤ 3
I_j	$T_{j-1} < t \le T_j$	3 < t ≤ 6
I_c	$Tr-1 < t \le \infty$	6 <t ≤ ∞

شکل 16 تعریف فاصله های زمانی

Sid	sequence
10	⟨(a, 1), (e, 3) (a, 4), (b, 4), (a, 6), (c, 6), (c, 10)⟩
20	⟨(d, 5), (a, 7) (b, 7), (e, 7), (d, 9), (e, 9), (c, 14), (d, 14)⟩
30	⟨(a, 8), (b, 8) (e, 11), (d, 13), (b, 16), (e, 16), (c, 20)⟩
40	⟨(b, 15), (f, 17), (e, 18), (b, 22), (c, 22)⟩

شکل 17 جدول تراکش که در آن زمان رخداد هر رویداد مشخص شده است.

اگر شکل (17) را به عنوان تراکنشهای اولیه در نظر بگیریم. پس مرحله تولید L1 و تعیین الگوهای پر بسامد با طول 1 برای تولید C2 از قانون L1 × Tk × L1; استفاده می شود. به عنوان مثال با داشتن دو عضو b وc و نیز با تعریف مجموعه T به صورت {I0; I1; I2} مجموعه C2 عبارت است از:

(b; I0; b); (b; I1; b); (b; I2; b); (b; I0; c); (b; I1; c); (b; I2; c); (c; I0; b); (c; I1; b); (c; I2; b); (c; I0; c); (c; I1; c) و(c; I2; c):

فرم جدید علاوه بر اینکه وجود رابطه را بین دو عضو نشان می دهد حاوی اطلاع اضافه تری در خصوص فاصله زمانی موجود بین آنها نیز می باشد. بعد از تعیین support هر کدام از سه تایی ها مواردی که شرط min-sup را دارند برای تولید گراف استفاده می شوند. برای افزایش دقت گراف نهایی محدوده ها را به صورت فاصله های دو هفته ای تعیین نمودیم . بنابراین مجموعه T6 به صورت زیر تعریف می شود:

T6={I0(t=0),I1(0<t≤2),I2(2<t≤4),I3(4<t≤6),I4(6<t≤8),I5(t>8)}

برای استفاده از این 6 بازه زمانی ابتدا تغییر کوچکی در ساختار جدول L2 می دهیم به این صورت که هر رکورد این جدول علاوه بر فیلد های قبلی دربردارنده فیلدی به نام Interval است که در این فیلد تفاضل Etime و STime محاسبه و نگهداری می شود. پس از اعمال این تغییر از اعداد موجود در فیلد Interval می تواند برای تفکیک رکوردها در محدوده های تعریف شده استفاده کرد. این تفکیک با استفاده از جدول سومی به نام I2 صورت می گیرد. هر رکورد I2 به صورت {Set,I0,I1,I2,I3,I4,I5} تعریف می شود. فیلد Set دربردارنده دوتایی های پر بسامد بوده و هرکدام از I_i ها دربردارنده دفعات مشاهده آن مجموعه با فاصله زمانی مورد نظر هستند. دستور SQL نمایش داده شده در شکل (18) برای مقدار دهی هر کدام از فیلد های جدول I2 بکار می رود. با توجه به اینکه این دستور در هر بار اضافه کردن یک رکورد فقط محتوای یکی از Ii ها را مقدار دهی می کند از دستور شکل (19) برای دسته بندی نهایی استفاده می شود. در نهایت جدول ویرایش شده به صورت شکل (20) و شکل (21) خواهد بود. لازم به ذکر است هر رکورد جدول I2 پس از ویرایش نه تنها به زیر مجموعه های پر بسامدی اشاره می کند که به صورت متوالی مشاهده شده اند بلکه اطلاع اضافه تری نیز درخصوص فاصله زمانی مشاهده علامت دوم بعد از مشاهده علامت اول بدست می دهد. به عنوان مثال رکورد {High GGT,Retinopathy,41,,6,24,,76} به این صورت تفسیر می شود که تعداد افرادی که High GGT(افزایش قند خون) و Retinopathy(اختلال بینایی) را به طور همزمان دیده اند 41 نفر، تعداد افرادی که این دو علامت را در فاصله 4 الی 6 هفته مشاهده کرده اند 6 نفر بوده است و .... . علاوه بر تعداد افرادی که علائم مذکور در آنها مشاهده شده است درصد مشاهده این دو علامت بالینی هم اهمیت دارد، بنابراین می توان تعداد این افراد را به تعداد کل دنباله های پر بسامد تقسیم کرد. بدون هیچ ویرایش قبلی در این مورد 385 دنباله متواتر تشخیص داده شد که با حذف مواردی که بسامد آنها از 10% تعداد کل دنباله ها کمتر بوده است 121 دنباله باقیماند (در اینجا عدد 10% یک مقدار انتخابی برای min-supp است) و پس از دسته بندی 64 دنباله برای تولید گراف مشخص می شود. هر چند تمام 121 قانون بدست آمده می توانند ارزش مدیکال داشته باشند ولی برای اجتناب از تراکم بیش از حد یالها در گراف نهایی می توان در مرحله قبل عدد بزرگتری را به عنوان min-supp در نظر گرفت و تعداد رکوردهایی که به مرحله بعدی می رسند را کاهش داد. بخشی از گراف نهایی تولید شده به صورت شکل (22) خواهد بود. در رسم این گراف مقدار min-sup برابر با 20% فرض شده است.

```
INSERT INTO I2 ( S, I2 )
SELECT START, Count(Interval)
FROM L2
WHERE Interval>2 AND Interval<=4
GROUP BY START;
```

شکل 18 مقدار دهی فیلدهای جدول I2

```
SELECT Set, sum(I0) AS a, sum(I1) AS b,
 sum(2) AS c, sum(I3) AS d, sum(I4) AS e,
 sum(I5) AS f
FROM I2
GROUP BY Set;
```

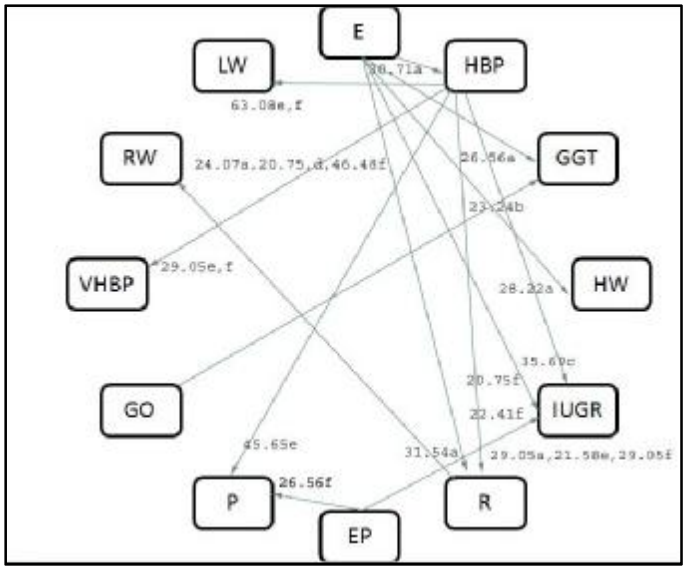
شکل 19 دسته بندی نهایی جدول I2

S	a	b	c	d	e	f
high Blood Pressure,Very hig			4	14		35
High GGT,high weight	29		6	20		70
High GGT,IUGR			6		11	45
High GGT,Low weight	25		6	12		17
High GGT,Retinopathy	41		6	24		76
High GGT,ringworm	44		6	23		37
high weight,IUGR			2			21
high weight,Retinopathy	29		3	15		24
Low Blood Pressure,Low weig			2			14
Low Heart-Beat in a Chrysalis	12		2			
Low weight,Retinopathy	22		4	17		

شکل 20 جدول I2 پس از ویرایش (اعداد نمایش دهنده تعداد)

S	a	b	c	d	e	f
high Blood Pressure,Very hi			3.32	11.62		29.05
High GGT,high weight	24.07		4.98	16.6		58.1
High GGT,IUGR			4.98		9.13	37.33
High GGT,Low weight	20.75		4.98	9.96		14.11
High GGT,Retinopathy	34.03		4.98	19.92		63.08
High GGT,ringworm	36.52		4.98	19.09		30.71
high weight,IUGR			1.66			17.43
high weight,Retinopathy	24.07		6.64	12.45		19.92
Low Blood Pressure,Low we			1.66			11.62
Low Heart-Beat in a Chrysali	9.96		1.66			
Low weight,Retinopathy	18.26		3.32	14.11		

شکل 21 جدول I2 پس از ویرایش (اعداد نمایش دهنده درصد)

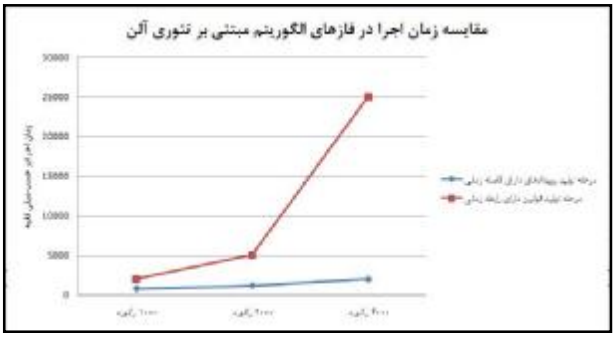


شکل 22 گراف نهایی حاصل از الگوریتم استقرایی توسعه یافته

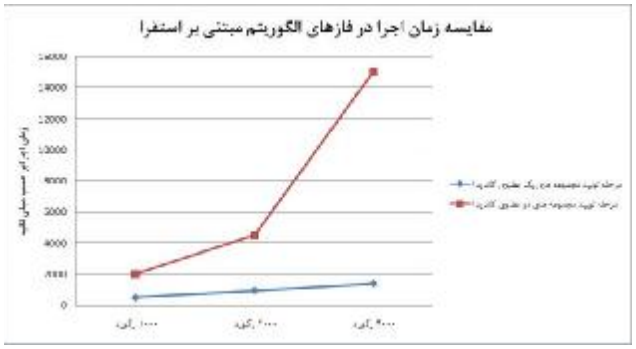
5. نتیجه گیری و اقدامات آینده

در این مطالعه صرف نظر از اعتبار قوانین مجموعه ی تراکنشهای ثبت شده دوبار و هربار به تعداد متفاوت تکرار شده است و زمان اجـرای فازهـای مختلف الگوریتم روی مجموعه هایی با تراکنشهای 1000 و 2000 و 4000 تراکنش و بر حسب میلی ثانیه محاسبه شده است. بررسی زمان اجرا نشان می دهد که زمان اجرای فاز دوم که مربوط به تعیین روابط زمانی از رویدادهای تولید شده است بیشتر از فاز اول بوده است و با اختلاف زیادی نسبت به روند خطی زمان اجرا در فاز اول تغییر پیدا می کند. این موضوع مشخص می کند که یکی از جنبه های تحقیقاتی قابل تعریف در زمینه افزایش کارایی الگوریتم بررسی روشهای کاهش زمان اجرا در فاز دوم است. همین مطالعه روی فازهای الگوریتم مبتنی بر استقرا صورت پذیرفتـه اسـت. فازهـای ایـن الگوریتم عبارتند از یافتن کاندیداهای یک عضوی و مجموعه های کاندیدای دو عضوی. با سه مجموعه استفاده شده در تحلیل قبلی نتایج بدسـت آمـده حاکی از آن است که در کل رشد پیچیدگی زمانی در فاز زمانگیر الگوریتم که تولید مجموعه های دو عضوی است نسبت به الگوریتم آلـن کمتـر اسـت.

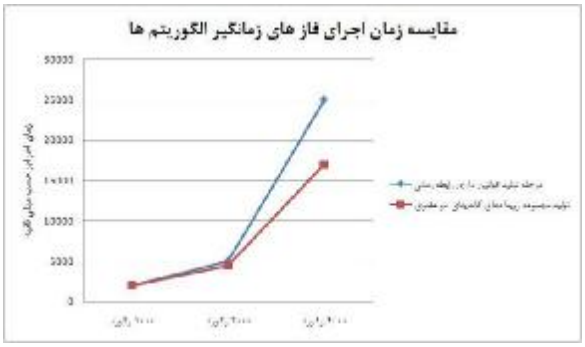
شکل های (23)، (24)و (25) به ترتیب به مقایسه فازهای الگوریتم آلن و مقایسه فازهای الگوریتم مبتنی بر استقرا و سپس فازهای زمانگیر هر دو الگوریتم پرداخته اند.



شکل 23 مقایسه زمان اجرا در فاز های الگوریتم مبتنی بر تئوری آلن



شکل 24 مقایسه زمان اجرا در فاز های الگوریتم مبتنی بر استقرا



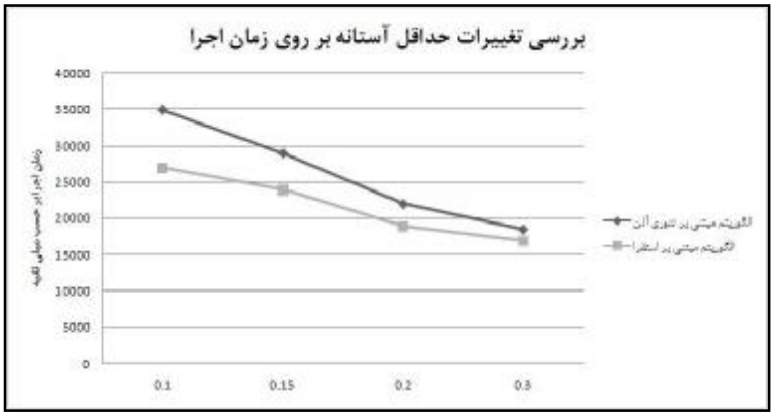
شکل 25 مقایسه زمان اجرای فاز های زمانگیر الگوریتم ها

5.1 بررسی نوسانات حداقل آستانه بر تعداد قوانین

دیدگاه دوم بررسی تاثیر نوسانات حداقل آستانه روی تعداد قوانین تولید شده است. همانطور که انتظار می رود با کاهش مقدار حداقل آستانه تعداد قوانینی که اعتبار لازم را بدستمی آورند افزایش می یابد و با افزایش حداقل آستانه تعداد قوانین کمتری تولید می شوند.مقایسه الگوریتم ها نشان داد که این روند در بررسی رفتار هر دو الگوریتم دیده شده است.شکلهای (26) و (27) نشان دهنده این موضوع است که در هرکدام رروند تغییرات حداقل آستانه روی داده های یک گروه از دادههای اولیه در الگوریتم آلن نشان داده شده است.

شکل 26 بررسی تاثیر میزان حداقل آستانه در تعداد قوانین تولید شده در داده های هایپرتانسیون

شکل27 بررسی تاثیر میزان حداقل آستانه در تعداد قوانین تولید شده در داده های دیابت



شکل 28بررسی تغییرات حداقل آستانه بر روی زمان اجرا

5.2 بررسی تاثیر میزان حداقل آستانه در زمان اجرا

دیدگاه سوم به بررسی تاثیر مقدار حداقل آستانه در زمان اجرا می پردازد به طور کلی مشاهده شد که با افزایش مقدار حداقل آستانه و کاهش تعداد موارد شرکت کننده در تولید قوانین زمان اجرا نیز کاهش می یابد. این تحلیل روی مجموعه های داده های 4000 تایی به صورت شکل (28) است.

6. پیشنهادات آتی

پیشنهادات زیر در رابطه با افزایش کارایی سیستم ایجاد شده به شرح زیر ارائه می گردد:

- با توجه به اینکه داده کاوی یک دانش وابسته به کاربرد است و هرچه حجم و دقت اولیه داده ای که برای کشف الگو مورد استفاده قرار می گیرد افزایش یابد صحت و اعتبار قوانین تولید شده نیز افزایش خواهد یافت، پیشنهاد می شود که تولید قوانین با استفاده از اطلاعات تعداد بیشتری از پرونده های بالینی انجام بگیرد و بدین ترتیب اعتبار الگوریتم و قوانین تولید شده توسط آن به صورت دقیق تر با نتایج آماری مدیکال ارائه شده مقایسه گردد. بکارگیری روشهایی که بتواند روند جمع آوری اطلاعات را با دقت و سرعت بیشتری همراه کند و خطاهای معمول را کاهش دهد می تواند در مسیر ایجاد یک سیستم داده کاوی چند منظوره با اهداف پژوهشی و درمانی موثر باشد.
- در این مطالعه با استفاده از الگوریتم مبتنی بر استقرا قوانین زمانی در شش بازه دو هفته ای تولید شدند. پیشنهاد می شود که در مطالعات آتی تعداد این بازه ها همزمان با کاهش طول بازه افزایش یابد . این امر به منظور بررسی دقیق تر نوسانات احتمالی در علائم بالینی و تولید قوانین دقیق تر می تواند موثر باشد.

- همانطور که مشاهده شد قوانین تولید شده توسط الگوریتم مبتنی بر استقرا بازه های زمانی را به صورت دقیق تعریف و استفاده کرده اند. به عنوان مثال می توان از این قوانین اینگونه استنباط کرد که فاصله بین دو رویداد می تواند بزرگتر یا مساوی دو هفته و یا کوچکتر از چهار هفته باشد. این درحالیست که معمولا در بیان فاصله های زمانی بین دو رویداد در اصطلاح عامیانه، از الفاظ تطبیقی دقیقی مثل این عبارت ها استفاده نمی شود. علاوه بر اینکه صرف تعلق به یک بازه زمانی نمی تواند نفی کننده تعلق به بازه های دیگر باشد. این موارد که نشان دهنده نارسایی نسبی بازه های منطقی برای نشان دادن مفهوم پیوسته زمان است، ایده استفاده از منطق فازی را برای بیان فاصله بین دو رویداد تقویت می کند. در این راستا پیشنهاد می شود که بر اساس بازه های زمانی تعیین شده توابع عضویت فازی با استناد به تعاریفی و اصطلاحاتی که برای بیان فاصله زمانی کاربرد دارد مورد استفاده قرار بگیرد.

7. منابع

[۱] G.Young, *"Safer childbirth: a critical history of maternity care. Third edition."* , Oxford University Press Family Practice Vol. ۱٦, No. ۳,pp. ۳۲۱-۳۲۲, ۱۹۹۹

[2] http://www.who.int/healthinfo/statistics/regions/en/index.html

[3] Guozhu Dong, Jian Pei "*Sequence Data Mining*";Springer Link;2007

[٤] Y.Lee, J.Lee, D.Chai, B.Hwang, KH.Ryu,"*Mining temporal interval relational rules from temporal data* ", The Journal of Systems and Software Vol. ۸۲,pp. ۱٥٥–۱٦۷,۲۰۰۹.

[٥] Y. Huang , Sh.Lin,"*Mining Sequential Patterns Using Graph Search Techniques* ",IEEE,Proceedings of the ۲۷th Annual International Computer Software and Applications Conference (COMPSAC'۰۳),۲۰۰۳.

[٦] Y.Chen, T.Cheng, K.Huang,"*Discovering Fuzzy Time-Interval Sequential Patterns in Sequence Databases* ", IEEE Transaction on systems—Part B: , Vol.۳٥, No. ٥, ۲۰۰٥.

[۷] J. Jbilou, N.Amara, R.Landry,"*Research-Based-Decision-Making in Canadian Health Organizations: A Behavioural Approach*", Springer Science, J Med Syst, Vol. ۳۱, pp. ۱۸٥–۱۹٦, ۲۰۰۷.

[۸] M.Lin, S-Ch.Hsue, Ch-W.Chang,"*Fast discovery of sequential patterns in large databases using effective time-indexing*", Elsevier, Information Sciences, Vol. ۱۷۸ ,pp. ٤۲۲۸–٤۲٤٥,۲۰۰۸.

[۹] S. Laxman. P. S.Sastry," A *survey of temporal data mining* ",ACM,S¯adhan Vol. ۳۱, Part ۲, pp. ۱۷۳–۱۹۸. April ۲۰۰٦.

[۱۰] M.Lin, S-Ch.Hsue, Ch-W.Chang,"*Fast discovery of sequential patterns in large databases using effective time-indexing*", Elsevier, Information Sciences, Vol. ۱۷۸ ,pp. ٤۲۲۸–٤۲٤٥,۲۰۰۸.

[۱۱] Ch,Chu, V S. Tseng, T.Liang,"*E.cient mining of temporal emerging itemsets from data streams* ",Expert Systems with Applications Vol. ۳٦,pp. ۸۸٥–۸۹۳,۲۰۰۹.

[1] Frequent Sequence Patterns (FSP)

[2] Similar time sequences

[3] Cyclic and temporal association rules