[**I.** **DATA MINING IN HEALCARE: CURRENT APPLICATIONS AND ISSUES** (By Ruben D. Canlas Jr., MSIT, MBA 5 August 2009) 2](#_Toc336771247)

[1. Trích dẫn: 2](#_Toc336771248)

[2. Giới thiệu và các nhân tố cơ bản: 2](#_Toc336771249)

[3. Phương pháp luận: 2](#_Toc336771250)

[4. Các nghiên cứu: 2](#_Toc336771251)

[5. Các vấn đề và thử thách: 6](#_Toc336771252)

[6. Kết luận và khuyến nghị: 7](#_Toc336771253)

[**II.** **Exploration Mining in Diabetic Patients Databases: Finding and Conclusions** 8](#_Toc336771254)

[1. Trích dẫn: 8](#_Toc336771255)

[2. Giới thiệu: 8](#_Toc336771256)

[3. Công việc liên quan: 9](#_Toc336771257)

[4. Làm sạch dữ liệu bán tự động: 9](#_Toc336771258)

[5. Khai phá dữ liệu của bệnh nhân mắc bệnh tiểu đường: 13](#_Toc336771259)

[6. Kết luận 13](#_Toc336771260)

[**III.** **Nect** 13](#_Toc336771261)

1. **DATA MINING IN HEALCARE: CURRENT APPLICATIONS AND ISSUES** (By Ruben D. Canlas Jr., MSIT, MBA 5 August 2009)
2. Trích dẫn:

* Với sự thành công của các ứng dụng được tích hợp Data mining trong các lĩnh vực rõ rệt như thương mại điện tử, marketing và bán lẻ đã cho ta thấy tính phổ biến của việc sử dụng những tri thức được lấy từ cơ sở dữ liệu ( Knowledge Discovery in Database – KDD) trong tất cả các lĩnh vực. Một trong các lĩnh vực mới nhất vừa được áp dụng Data mining đó chính là y học và sức khỏe cộng đồng.
* Bài báo ngày cho thấy một cuộc khảo sát các kĩ thuật hiện tại của KDD, sử dụng các công cụ đã tích hợp Data mining trong y tế và sức khỏa công đồng. Ngoài ra còn bao gồm các vấn đề quan trọng và thử thách trong việc kết hợp Data mining với y tế nói chung.
* Bài nghiên cứu còn cho thấy sự phát triển của một lượng lớn các ứng dụng có Data mining, bao gồm các ứng dụng như phân tích dữ liệu tại các bệnh viện để đưa ra các chính sách hoạch định sức khỏe tốt hơn, phát hiện dịch bệnh, ngăn ngừa các trường hợp tử vong trong bệnh viện và phát hiện các gian lận trong quyền bản hiểm.

1. Giới thiệu và các nhân tố cơ bản:

* Mục tiêu:
* Liệt kê các lợi ích và nổi bật tầm quan trọng của Data mining trong y học vào sức khỏe cộng đồng.
* Tìm kiếm các kỹ thuật Data mining đã được sử dụng trong các lĩnh vực khác có thể áp dụng cho sức khỏe cộng đồng.
* Xác định các vấn đề và thử thách trong ứng dụng Data mining vào y khoa.
* Vạch ra những khuyến nghị cho việc thu thập các tri thức trong các cơ sở dữ liệu điện tử thông qua Data mining

1. Phương pháp luận:

* Dựa theo các phương pháp bị hạn chế, phương pháp luận chính được sử dụng trong bài nguyên cứu này thông qua các bài đã được xuất bản trong nhiều lĩnh vực của y học, khoa học và kỹ thuật máy tính. Bài nghiên cứu dựa trên những xuất bản gần đây nhất, từ năm 2000 trở về đây.

1. Các nghiên cứu:
   1. Data mining trong phát hiện dịch bệnh:

* Sử dụng dữ liệu thiết thực để hỗ trợ cho việc ra quyết định trong y học ( còn được biết là EBM – Evidence Based Medicine) đã tồn tại trong nhiều thế kỷ. John Snow, đã được biết đến là cha đẻ của dịch tễ học hiện đại, đã sử dụng các bản đồ đồ thị vào năm 1854 để phát hiện ra nguồn bệnh thổ tả và đã chứng minh rằng bệnh này lây lan qua các hệ thống cấp nước.
* Snow đếm số lượng các bệnh nhân và vẽ sơ đồ địa chỉ của các bệnh nhân trên bản đồng là các thanh màu đen. Ông đã phát hiện rằng hầu hết các ca tử vong đều nằm xung quanh một điểm bơm nước nhất định tại London.
* Florence Nightngale đã phát minh ra biểu đồ vùng phân cực vào năm 1855 để chi ra nhiều ca tử vong của binh lính là do sự mất vệ sinh trong khám lâm sàn và có thể ngăn ngừa được. Bà đã sử dụng những biểu đồ để thuyết phục các chinh sách thực hiện cải cách và đã giảm được số lượng các ca tử vong.
* Snow và Nightingale đã chính mình thực hiện thu thập dữ liệu, sang lọc và phân tích thông qua các dữ liệu về tỉ lệ tử vong trong suốt thời gian nghiên cứu vì số lượng bộ thông tin có thể quản lý được. Ngày nay, số lượng dân số, dữ liệu điện tử thu thập được cùng với sự toàn cầu hóa và tốc độ phát bệnh của các dịch bệnh làm cho việc thao tác dữ liệu gần như hoàn toàn không thể hoàn thành.
* Đó là nguyên nhân tại sao Data mining trở nên hữu ích trong việc chăm sóc sức khỏe. Tuy phát triển khác chậm nhưng vẫn được áp dụng để giải quyết các vấn đề khác nhau của việc thu thập tri thức trong lĩnh vực này.
* Data mining và các ứng dụng của nó trong y học và sức khỏe cộng đồng là một lĩnh vực học tập còn non trẻ. Vào năm 2003, Wilson et al bắt đầu thăm dò các trường hợp mà KDD và kỹ thuật Data mining đã được áp dụng trong cơ sở dữ liệu y tế. Người ta cũng đã phát hiện nhiều sự nhầm lẫn trong các lĩnh vực được tạo thành từ Data mining. “Một vài tác giả cho rằng Data mining chỉ là một quá trình thu thập, trong khi các tác giả khác cho rằng Data mining sử dụng kỹ thuật thống kế trong quá trình thu thập trị thức.” (Wilson et al 2003)
* Vì những quan niệm sai lầm vẫn tiếp tục trong cộng đồng y khoa về những gì mà Data mining chứa đựng, chúng ta sẽ định nghĩa lại những quan niệm đó. Một định nghĩa được chấp nhận nhiều nhất về Data mining ngày nay là một bộ các quy trình và kỹ thuật trong khai phá và mô tả kiểu và hướng trong dữ liệu (Witten and Frank 2005). Chúng ta sẽ sử dụng định nghĩa này xuyên suốt trong bài cáo
  1. Tầm quan trọng và lợi ích của Data mining trong y học và sức khỏe cộng đồng:
* Mặc dù sự khác biệt và xung đột trong hướng tiếp cận, lĩnh vực y tế hiện nay cần nhiều Data mining hơn. Có 1 vài đối số có thể được nâng cao để hỗ trợ trong việc sử dụng Data mining trong lĩnh việc y tế. Không chỉ trong y tế cộng đồ mà còn trong lĩnh vực y tế tư nhân.
* Tràn dữ liệu (*Data overload*). Nhiều tri thức có giá trị được thu bởi việc tin họa hóa các dữ liệu y tế. Tuy nhiên, một lượng lớn dữ liệu được lưu trữ sẽ làm cho việc này trở nên khó khăn và gần như là không thể nếu người ta phải sàng lọc và khai phá tri thức (Cheng, et al 2006).
* Trong thực tế, một số chuyên gia tin rằng những đột phá trong lĩnh vực y tế đã chậm lại, thêm vào đó là độ phức tạp của các thông tin y tế hiện nay. Máy tính và Data mining là thích hợp nhất cho mục đích trên. (Shillabeer and Roddick 2007).
* Y học dựa trên thông tin và ngăn chặn các sai sót trong bệnh viên. Khi các tổ chức y tế ứng dụng Data mining vào dữ liệu, họ có thể khai phá được các tri thức hữu ích và khả năng cứu sống nhiều mạng sống vẫn còn nằm trong cơ sở dữ liệu của họ. Ví dụ, một nghiên cứu đang được thực hiện trong các bệnh viên đã cho thấy rằng 87% ca tử vong trong các bệnh viên tại Mỹ có thể ngăn ngừa nếu các nhân viên bệnh viện (kể cả bác sĩ) cẩn thận hơn trong ngăn ngừa sai sót ( HeathGrades Hospitals Study 2007). Bằng việc khai phá các dữ liệu bệnh viên, các vấn đề về an toàn có thể được phát hiện và giải quyết bởi các quản lý bệnh viện và chính phủ.
* Hoạch định chính sách trong y tế cộng đồng (*Policy-making in public heath*). Lavrac et al. (2007)đã kết hợp GIS và Data mining với nhau, Weka với J48, để phân tích điểm giống nhau giữa các trung tâm y tế tại Slovenia. Sự dụng Data mining, họ có thể khai phá được các mô hình trong trung tâm y tế để có thể đưa ra các chính sách khuyến nghị đến Viện Y Tế Cộng Đồng. Họ đã kết luận rằng “Data mining và hệ thống hỗ trợ ra quyết định, kể cả các phương pháp trực quan mới, có thể nâng cao hệ thống ra quyết định.
* Các yếu tố trên đã nhắc chúng ta nhớ tới một sự cố tại trung tâm y tế Rizal của Pasig City của Indonexia vào tháng 10 năm 2006. Thất bại trong việc thực hiện vệ sinh nghiêm ngặt và các biện pháp khử trùng trong bệnh viên đã góp phần dẫn đến các ca tử vong của trẻ so sinh do bị nhiểm khuẩn. Không ai có thể biết được chính xác chuyện gì đang xảy ra cho tới khi các ca tư vong trở nên thường xuyên. Sau khi kiểm tra dữ liệu của bệnh viên, bộ phận y tế ( Department of Health – DOH) đã tìm thấy 12 trong 28 trẻ được sinh ra vào ngày 4 tháng 10 đã tử vong vì nhiểm khuẩn (Tandoc 2006). Với cơ sở dữ liệu đã được xử lý và một ứng dụng Data mining, DOH có thể các hiện các sự kiện không bình thường và hạn chế chúng diễn biến tệ hơn.
* Tiết kiệm tiền và chi phí. Các tổ chức dùng Data mining để rút trích nhiều cơ sở dữ liệu với chi phí thấp nhất. KDD và Data mining đã được áp dụng vào việc tìm ra các gian lận trong thẻ tín dụng và bảo hiểm (Kou et al. 2004). Thêm vào đó, những kỹ thuật này có thể được sử dụng để tìm ra các bệnh nhân khác biệt trong bảo hiểm y tế - một hệ thống bảo hiểm y tế thuộc quốc gia Philippin đã từng thực hiện điều này.
* Phát hiện sớm và ngăn chặn bệnh dịch. Cheng, el al đã đưa ra dẫn chứng trong việc sử dụng các thuật toán phân lớp để phát hiện các trường hợp mắc bệnh tim, một loại bệnh được quan tâm nhất trên thế giới. Cao et al (2008) đã mô tả việc sử dụng Datamining như một công cụ hỗ việc kiểm soát các thí nghiệm vac xin lâm sàn. Sử dụng Data mining đã giúp cho y học có thề phát hiện ra nhiều bệnh nhân có biểu hiện khác thường hơn là chỉ nhìn vào một tập dữ liệu.
* Ngăn ngừa, quản lý bệnh dịch và đưa ra các chính sách y tế cộng đồng. Các chuyên gia y tế đã bắt đầu tìm cách ứng dụng Data mining vào việc phát hiện và quản lý dịch bệnh. Kellogg et at (2006) đã nêu ra một kỹ thuật kết hợp mô hình không gian, mô phỏng và khai phá dữ liệu không gian để tìm ra các được điểm của dịch bệnh. Phân tích kết quả của việc khai phá dữ liệu để đưa ra các chính sách nhằm phát hiện và quản lí dịch bệnh bùng phát.
* Wong et al. (2005) đã giới thiệu WSARE, một thuật toán phát hiện dịch bệnh khi vừa ở giai đoạn đầu. WSARE – “What’s Strange About Recent Events” được tạo nên dựa trên các luật kết hợp và mạng Bayesian. Áp dụng WSARE trên các mô hình giả lập có thể cho ra những kết quả tương đối chính xác các dịch bệnh đã được mô phỏng. Và nó cũng đi kèm với một lời cảnh báo khi áp dụng thật toán này trong đời sống thật.
* Chuẩn đoán và các hệ thống ra quyết định không gây hại. Một số chuẩn đoán và thí nghiệm có thể gây hại cho bệnh nhân. Lấy một ví dụ là việc tiến hành sinh thiết ở phụ nữ nhằm phát hiện bệnh ung thư cổ tử cung. Thangavel et al (2006) đã sử dụng thuật toán gom cụm K-means để phân tích các bệnh nhân ung thư cổ tử cung và đã nhận thấy việc gom cụm dữ liệu có để đưa ra những kết quản dự đoán tốt hơn các phương pháp hiện có. Ông cũng đã tìm thấy một số bộ thuộc tính có thể được bác sĩ sử dụng như một sự hỗ trợ trong việc ra quyết định đưa các bệnh nhân có khả năng bị bệnh ung thư đi tiến hành sinh thiết hay không.
* Gorunescu (2009) đã mô tả cách sử dụng hệ hỗ trợ chuẩn đoán bằng máy tính (CAD) và siêu âm nọi soi Elastography (EUSE) được ứng dụng Data mining để đưa ra những quyết định ngăn ngừa bệnh ung thư mà không gây hại cho bệnh nhân. Trong khi đó cách truyền thống là bác sĩ sẽ nhìn vào các bộ phim siêu âm và thiết bị để đề nghị bệnh nhân là sinh thiết hay không.
* Quyết định của bác sĩ là hoàn toàn chủ quan, phụ thuộc hoàn toàn vào các Video siêu âm. Gorunescu đã giải quyết việc này bằng một cách hoàn toàn khác, sử dụng Data mining. Ông đã không học về nhân khẩu học của bệnh nhân, thay vì tập trung vào các Video siêu âm, trước tiên họ huấn luyện một thuật toán phân lớp sử dụng MLP (multi-layer percetron) trên các trường hợp khối u lành tính và ác tính.
* Mô hình của họ phân tích các điểm ảnh và RGB của chúng để tìm ra các bệnh nhân nằm phân biệt khối u lành tính và ác tính. Sau đó, họ áp dụng kết quả của mô hình vào các trường hợp khác. Và kết quả của mô hình đã đạt độ chính xác cao trong chuẩn đoán với độ lịch chuẩn rất thấp.
* Phân loại các thuốc có hại (ADEs). Một số thuốc và hóa chất đã được đánh giá rằng sẽ không gây hại cho con người nhưng sau một thời gian dài sự dụng thì kết quả là ngược lại. Wilson et al (2003) đã tiết lộ rằng tổ chức US Food và Drug Administration đã sử dụng data mining để tìm ra các loại thức có hại khi sử dụng lâu dài trong cơ sở dữ liệu của họ. Thuật toán này có tên là MGPS – Multi-item Gamma Poisson Shrinker đã tìm ra 67% loại thuốc có hại sớm hơn 5 năm nếu dùng cách thông thường.
* Chúng ta đã thấy được những ứng dụng của data mining trong việc phát hiện bệnh dịch, ngăn ngừa tử vọng, cải thiện chuẩn đoán và thậm chí phát hiện các gian lận trong bảo hiểm y tế. Dù sao, có nhiều cách sử dụng Data mining khác nhau trong lĩnh vực y tế.

1. Các vấn đề và thử thách:

* Việc áp dụng Data mining vào lĩnh vực y tế gặp rất nhiều khó khăn do chính đặc trưng riêng của lĩnh vực này. Shillabeer và Roddick (2007) đã nêu ra một số mâu thuẫn giữa các phương pháp khai phá dữ liệu truyền thống và y học.
* Trong nghiên cứu, khai phá dữ liệu bắt đầu với một giả thuyết và sau đó đưa ra một kết quả đã được điều chỉnh phù hợp với giả thuyết đó. Trong khi đó, phương pháp chuẩn của khai phá dữ liệu trước đây là bắt đầu với các bộ dữ liệu.
* Ngoài ra, trong khi khai phá dữ liệu truyền thống quan tâm đến mô hình và hướng của các bộ dữ liệu thì trong y học lại ít quan tâm hơn. Khác biệt lớn nhất đó là phương pháp tiếp cận của khai phá dữ liệu chuẩn chỉ quan tâm đến việc mô tả chứ không phải giải thích mô hình và hướng của dữ liệu. Y học thì lại quan tâm đến từ những giải thích vì một khác biệt nhỏ nhặt nhất có thể là ranh giới giữa sống và chết.
* Lấy một ví dụ, bệnh than và bệnh cúm có cùng những triệu chứng về đường hô hấp. Nếu loại bỏ đi một số đặc điểm nhỏ trong khi thực hiện khai phá dữ liệu thì người ta có thể cho đó chỉ một đợt dịch cúm nhưng thật sự đó lại là dịch bệnh than (Wong et al 2005). Không phải ngẫu nhiên mà các bài báo về khai phá dữ liệu trong bệnh dịch và chữa trị luôn đưa ra những kết luận mơ hồ. Những thất bài này cho ta thấy được độ thiếu tin cậy của khai phá dữ liệu trong y tế.
* Việc nhầm lẫn về định nghĩa của khai phá dữ liệu cũng làm vấn đề trở nên phức tạp. Ví dụ, ta có thể thấy một số bài báo với từ khóa là “data mining” nhưng thực sử chúng chỉ là những biểu đồ bình thường. Shillabeer (2009) cho biết việc hiểu làm này là rất thường xuyên đối với những nơi mới bắt đầu áp dụng data mining vào y học.
* Cho dù các kết quả thu được từ khai phá dữ liệu là đáng tin cậy nhưng việc thay đổi thói quen của các bác sỹ, y tá không phải là một điều dễ dàng. Ayres (2008) đã cho biết đã có một số trường hợp các bác sỹ từ chối thay đổi các chính sách của bệnh viện. Đã có một trường hợp một bác sỹ đã quên rửa tay sau khi khám nghiệm tử thi và gây ra một nhiều ca tư vong cho nhiều bệnh nhân được khám ngay sau đó. Đến lúc này, các bác sỹ mới thay đổi các thói quen của họ.
* Shillabeer (2009) cũng cho biết rằng hầu hết các bác sỹ (ở Úc) thường nghe lời khuyên từ các chuyên gia y tế hơn là xem kết quả thu được từ khai phá dữ liệu.
* Ngoài ra những dữ liệu riêng tư của bệnh nhân cũng là một vật cản lớn trong việc ứng dụng khai phá dữ liệu vào y học vì để đưa ra một kết quả chính xác nhất thì cần một lượng lớn những dữ liệu cần thiết. Nhưng chỉ có những dữ liệu này mới có thể giúp con người ta tránh được những bệnh chết người.

1. Kết luận và khuyến nghị:

* Bài nghiên cứu về ứng dụng của khai phá dữ liệu trong y học và y tế cộng đồng đã cho ta thấy sơ qua những thành công và thử thách của khai phá dữ liệu. Các tổ chức sức khỏe có thể ứng dụng chúng để tìm kiếm những tri thức cần thiết nằm trong dữ liệu của họ.
* DOH có thể tìm ra được các trường hợp tử vọng có thể xảy ra để có thể đưa ra các chính sách tốt hơn về kế hoạch tim phòng, xác định vật truyền bệnh sốt rét, ngăn chặn các lỗi trong bệnh viện, dự đoán nơi dịch bệnh sẽ bùng phát và ngăn chặn các gian lận trong bảo hiểm.
* Trước khi bắt tay vào áp dụng data mining, các tổ chức phải thực thi các chính sách về quyền riêng tư vả bảo mật dữ liệu của các bệnh nhân đến các đối tượng liên quan, các chi nhánh và cơ quan của tổ chức.
* Chất lượng dịch vụ của y tế công đồng hoàn toàn phụ thuộc vào việc thao tác và ứng dụng khai phá dữ liệu lên các bộ dữ liệu.

1. **Exploration Mining in Diabetic Patients Databases: Finding and Conclusions**
2. Trích dẫn:

* Bài báo đề cập đến các bước tiền và hậu xử lý dữ liệu của các bệnh nhân mắc bệnh tiểu đường trong khai phá tri thức đã bị lãng quên. Bài báo mô tả một phương pháp làm sạch dữ liệu của bệnh nhân mắc bệnh tiểu đường một cách bán tự động và giới thiệu cách tiếp cận từng bước một.

1. Giới thiệu:

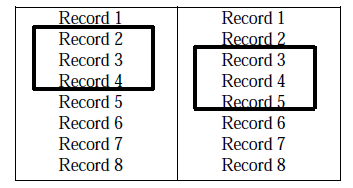
* Singapore có khoảng 10% dân số mắc bệnh tiểu đường, bệnh đã gây nên nhiều nguy cơ mắc các bệnh về mắt, suy thận và các biến chứng nguy hiểm khác. Tuy nhiên nếu được chuẩn đoán và điều trị riêng thì sẽ làm giảm được tỷ lệ trên. Để chuẩn đoán bệnh tiểu đường, Singapore đã giới thiệu một chương trình kiểm tra sức khỏe thường xuyên dành cho bệnh nhân tiểu đường vào năm 1992. Thông tin bệnh nhân, triệu chứng lâm sàn, chuẩn đoán các bệnh về mắt và cách đều trị đều được gộp chung vào một cơ sở dữ liệu. Sau 8 năm thu thập một lượng lớn dữ liệu có giá trị, nước này đã bắt đầu áp dụng khai phá dữ liệu để tìm ra những bệnh nhân đặc biệt. Mục tiêu là tìm ra các tập luật dành cho các bác sĩ hiểu rỏ thêm các đặc tính của bệnh tiểu đường và các triệu chứng của nó. Mặc dù đã gặt hái được nhiều thành công trong lĩnh vực kinh tế nhưng khi được ứng dụng vào y học thì vẫn còn khó khăn đối với khai phá dữ liệu. Điều đó xuất phát từ 2 lý do sau:
* Lượng dữ liệu thu được rất hỗn tạp. Nhiều dữ liệu bệnh nhân mắc phải các lỗi đánh máy, thiếu giá trị, sai thông tin về địa chỉ, ngày sinh và nhiều lúc có nhiều dữ liệu bị lập lại. Để làm sạch sẽ những dữ liệu này cần một lượng lớn công sức và thời gian. Thêm vào đó là một số dữ liệu thu được không thích hợp để thực hiện khai phá dữ liệu. Vì thế họ phải thực hiển chuyển đổi các thuộc tính cần thiết trước khi sử dụng.
* Việc các bác sĩ quá bận với công việc của mình nên hầu như họ không có đủ thời gian và sức lực để sàng lọc các luật đã được khai phái. Vì thế việc tổng hợp nên các tập luật dễ hiểu là điều rất cần thiết.
* Bài báo sẽ chỉ ra cách giải quyết các dữ liệu hỗn tạp bằng hệ thống làm sạch dữ bán tự động. Hệ thống cho phép các bác sĩ chỉ định việc ánh xa giữa các thuộc tính ở một định dạng khác và các lược đồ đã được sử dụng để điều chỉnh các định dạng dữ liệu khác nhau. Một khi một định dạng mới được thêm vào thì việc nhận dạng và loại bỏ các dữ liệu trùng lặp trở nên dễ dàng hơn. Để giải quyết vấn đề có quá nhiều tập luật được khai phá, nhóm nghiên cứu đã áp dụng cách tiếp cận hướng người dùng bằng cách từng bước khai phá dữ liệu nhằm làm cho những tri thức khai phá được dễ hiểu hơn

1. Công việc liên quan:

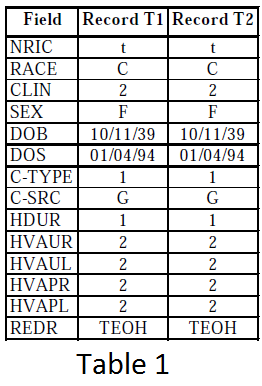
* Y tế là một lĩnh vực màu mỡ cho việc phát triển các ứng dụng khai phá dữ liệu. Tuy nhiên hiện nay số lượng ứng dụng được phát triển là rất ít so với các lĩnh việc khác. Với những tập luật khai phá được thì lại có một lượng lớn các luật kết hợp trở nên dư thừa. Các nghiên cứu trước đó đã giải quyết vấn đề này theo những cách tiếp cận sau:
* Cho phép người sử dụng truy vấn những tập luật đã được khai phá trước đó. Người dùng có thể điều chỉnh các tập luật là họ chú ý. Sau đó hệ thống sử dụng các mẫu đó để thu các tập luật phù hợp với bộ tập luật khai phá được.
* Sử dụng các ràng buộc để hạn chế việc khai phá dữ liệu sao cho chỉ tạo ra các tập luật có liên quan. Phương pháp này nhiều lúc lại không thật sự tốt vì nhiều bác sĩ lại không biết cách định dạnh các tập luật để khai phá.
* Tìm ra các tập luật không mong đợi.
* Cách tiếp cận thứ nhất và hai đã gặp nhiều lỗi trong việc sử dụng vì thường các bác sĩ không biết học đang tìm gì và vì thế không thể đưa ra được các mẫu chuẩn. Cách thứ 3 thì lại bị hạn chế vì bác sĩ lại không có khả năng định dạng các mẫu có sẵn. Ngoài ra việc định dạng các tri trức cũng là một gánh nặng đối với các bác sĩ.
* Các nghiên cứu trước đây đều tập trung vào việc tạo ra các tập luật là chính. Theo bài báo cao, quá trình tiền và hậu xử lý dữ liệu còn quan trọng hơn quá trình tạo ra các tập luật.

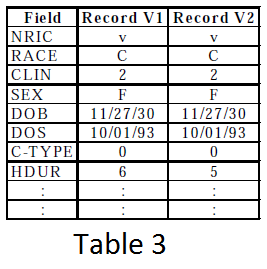
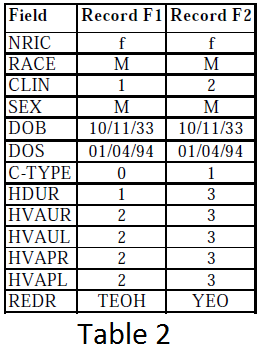
1. Làm sạch dữ liệu bán tự động:

* Chúng tôi bắt đầu thực hiện dự án khái phá dữ liệu bệnh nhân mắc bệnh tiểu đường từ những năm 1999. Và đã nhận được hơn 200 000 bộ dữ liệu từ năm 1992-1996. Một bộ có hơn 60 thuộc tính. Một phân tích sơ bộ cho thấy thông tin của bệnh nhân bao gồm số chứng minh, quốc gia, giới tính, ngày sinh, thời gian mắc bệnh tiểu đường và ngày nhận dữ liệu. Một bộ dữ liệu đi kèm với một kết quả kiểm tra mắt nhằm phát hiện các trường hợp bệnh tiểu đường gây biến chứng sang mắt như các bệnh về võng mạc, đục thủy tinh thể, tăng nhãn áp. Kết quả kiểm tra mắt cũng được ghi vào các bộ dữ liệu.
* Nhóm nguyên cứu đã áp dụng nguyên tắc “Rác vào, rác ra” để làm sạch dữ liệu của bệnh nhân trước khi tiến hành khai phá. Đã có nhiều vấn đề trong quá trình này:
* Trong thực tế, với tốc độ cập nâng cấp phần mềm với phần cứng hiện nay. Nhiều cơ sở dữ liệu đã trải qua nhiều lần nâng cấp trong một thời gian ngắn đã làm thay đổi định dạng một cách đáng kể. Cơ sở dữ liệu của chúng tôi là một ví dụ, từ năm 1992 – 1996 chúng tôi đã gặp 4 định dạng dữ liệu khác nhau.
* Mỗi khi thay đổi định luôn đi kèm theo sự không nhất quá dữ liệu như dd/mm/yy hoặc mm/dd/yy và cách viết tắt như “ONE” với “1”.
* Có khá nhiều thuộc tính bị bỏ trống như tiền sử bệnh án, cách điều trị…
* Để giải quyết các vấn đề trên, chúng tôi đã sử dụng hệ thống làm sạch dữ liệu. Hệ thống này cho phép người sử dụng định nghĩa các thuộc tính thuộc định dạng khác, các lược đồ được mã hóa và các thuộc tính đã được làm sạch trong cơ sở dữ liệu để tạo ra một chuẩn lược đồ định dáng cuối cùng. Và được dùng để làm chuẩn để chuyển đổi cơ sở dữ liệu. Nhưng có một vấn đề vẫn còn tồn tại sau khi thực hiện bước chuẩn hóa dữ liệu đó là việc trùng lập dữ liệu. Trùng lập dữ liệu làm gia tăng lượng dữ liệu mà mất đi tính chính xác của tri thức được khai phá khi mỗi bệnh nhân có thể có một bộ dữ khác ngày thu thập.
* Phương pháp SNM đã được áp dụng để loại bỏ đi các dữ liệu trùng lặp. Sau đây là các bước của thuật toán:
* 1: Chọn 1 hoặc nhiều các thuộc tính là khóa chính.
* 2: Sắp xếp cơ sở dữ liệu theo các thuộc tính đã được chọn.
* 3: So sánh các thuộc tính được chọn trong cửa sổ trượt. Các dữ liệu trùng lặp được xuất thành 1 file.
* 4: Người dùng kiểm tra các dữ liệu trùng lặp bị xóa.
* SNM được áp dụng trước quá trình tiền xử lý dữ liệu trước khi dữ liệu được sắp xếp để làm tăng khả năng trùng lặp. Kỹ thuật này bao gồm việc loại bỏ các dữ liệu bị lỗi đánh máy hoặc viết tắt, mã hóa và sắp xếp các mã thông báo đó. Giữ sử dữ liệu 1 và 2 được đánh dấu là trùng lặp trừ thuộc tính Name, “Tan Lay-Hoon” và “Lay-Hoon Tan”. Nếu thuộc tính Name được dùng để sắp xếp thì không xảy ra trường hợp 2 dữ liệu này bị trùng lặp. Nhưng khi sắp xếp các mã thông báo sẽ tìm ra 2 bộ dữ liệu này bị trùng lặp. Chúng ta phải chọn một khóa cụ thể khi sử dụng kỹ thuật này để sắp xếp dữ liệu trước khi loại bỏ các dữ liệu trùng lặp. Khóa cụ thể đó là một thuộc tính con hoặc một chuỗi con trong thuộc tính. Hiện không có một luật cụ thế nào để thiết kế các khóa những việc chọn một thuộc tính đại diện cho bộ dữ liệu là rất quan trọng khi thực hiện sắp xếp cơ sở dữ liệu.
* Trong cơ sở dữ liệu, chúng tôi đã thận trọng chọn ra một bộ thuộc tính độc nhất dùng để xác định bộ dữ liệu của bệnh nhân. Bộ dữ liệu này gồm ID của bệnh nhân (NRIC) và ngày sàng lọc (DOS). NRIC là thuộc tính duy nhất đại diện cho người Singapore và nó cũng giống như số ID của nước Mỹ. Chúng tôi đã thực hiện tiền xử lý các NRIC này với những tập tin chứa ID và tên có thuật trong cơ sở dữ liệu của nhà nước nhằm loại bỏ lỗi đánh máy đồng thời sắp xếp cơ sở dữ liệu theo khóa này.
* Bước tiếp theo, so sách các bộ dữ liệu ở gần nhau. Hình dưới cho thấy một cửa sổ trượt có thể tùy chỉnh kích thước được thực thi trong khi thực hiện so sánh dữ liệu. Giả sử kích thước của cửa sổ là *w* bộ dữ liệu thì khi có một bộ dữ liệu mới được thêm vào, nó sẽ được so sánh với bộ dữ liệu *w-1* để tìm ra các bộ dữ liệu giống nhau.



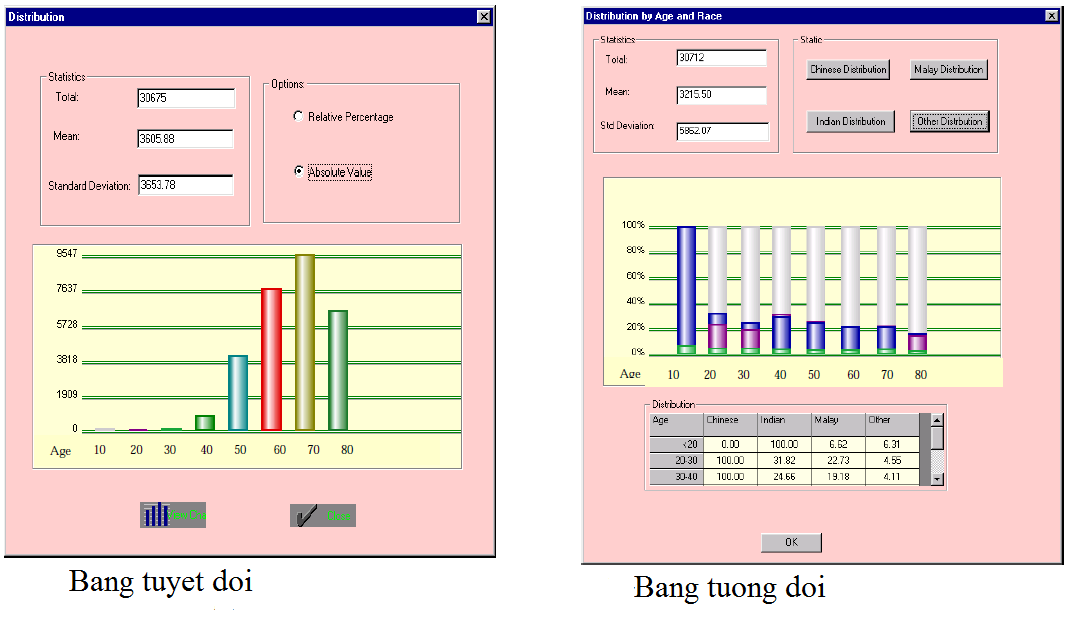
* Có 3 loại trùng lặp dữ liệu: trùng lặp đúng, trùng lặp sai và trùng lặp cần được kiểm tra sâu hơn. Tên và giá trị của thuộc tính đã được mã hóa để bảo đảm tính bảo mật của dữ liệu. Nên khóa chính được sử dụng trong so sánh là NRIC và DOS. Bảng 1 cho thấy T1 và T2 mang cùng một giá trị trong tất cả các thuộc tính
* Bảng thứ hai minh họa cho trường hợp trùng lặp sai:
* Bộ dữ liệu F1 & F2 là của cùng một bệnh nhau khi mang cùng giá trị của các thuộc tính như NRIC, DOB, DOS, RAC và SEX
* Các thuộc tính như CLIN, C-TYPE và kết quả sàng lọc (HDUR, HVAUR, HVAUL, HVAPR, HVAPL) đều khác nhau
* Thuộc tính cuối cùng - REDR là tên của bác sĩ đã khám cho bệnh nhân – đã cho thấy F1 và F2 là 2 bộ dữ liệu của cùng một bệnh nhân nhưng được sàng lọc riêng.
* Việc trùng lặp sai này hoàn toàn là do NRIC và DOS không phải là cặp khóa tốt nhất. Giải pháp được đưa ra là NRIC, DOS, CLIN nhưng hiếm khi một bệnh nhân lại đến 2 phòng khám trong cùng một ngày. Do đó, chúng tôi vẫn chọn NRIC và DOS là cặp khóa chính.
* Cuối cùng, bảng 3 cho thấy V1 và V2 giống nhau ở tất cả các thuộc tính chỉ trừ HDUR. Trong trường hợp này, các chuyên gia cần nghiên cứu thêm về giá trị của HDUR là 6 hay 5
* 70% dữ liệu trùng lặp đã được loại bỏ trong bước tiền xử lý. Nhưng trong thực thế, việc loại bỏ tất cả các dữ liệu trùng lặp rất khó.



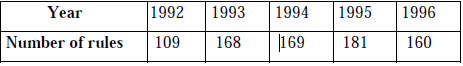


1. Khai phá dữ liệu của bệnh nhân mắc bệnh tiểu đường:

* Một khi dữ liệu đã được làm sạch thì quá trình khai phá bắt đầu. Hiện nay đã có rất nhiều kỹ thuật khai phá dữ liệu được áp dụng để khai phá các mô hình dữ liệu. Nhìn chung, các mô hình khai phá được nằm trong những dạng sau: mô hình phân lớp, mô hình kết hợp, mô hình tuần tự, mô hình không gian – thời gian. Trong nghiên cứu này, chúng tôi chủ yếu tập trung vào 2 mô hình: phân lớp và kết hợp. Mô hình phân lớp mô tả các đặc điểm của bệnh nhân mắc các bệnh về mắt do biến chứng của bệnh tiểu đường và dùng chúng để dựng đoán khả năng mắc bệnh về mắt của bệnh nhân. Mô hình kết hợp đưa ra danh sách các triệu chứng và cách điều trị đi kèm với nhau.
* Nhóm nghiên cứu đã sử dụng độ hỗ trợ tối thiểu là 1% và độ tin cậy tối thiểu là 50% để khai phá các luật kết hợp để tìm tổng số khoảng 700 luật và đã khiến cho các bác sỹ cảm thấy quá tải với số lượng này. Nhưng đã có quá trình hậu xử lý giúp các bác sỹ hiểu hết tất cả các luật này. Phương pháp khai phá của nhóm hướng tới việc đưa ra những tri thức dễ hiểu nhất và khai phá các mô hình nhằm giúp họ không phải làm việc với lượng lớn thông tin của người bệnh.
* Bước đầu, thông tin cơ bản về nhân khẩu học của bệnh nhân được đưa lên cho bác sỹ. Mục đích là để bác sỹ đưa ra những chuẩn đoán cơ bản dựa trên khác thuộc tính như là tuổi, nước, số lượng bệnh nhân mắc bệnh. Có 2 tỷ lệ được đưa ra là: tỉ lệ tuyết đối và tương đối.



* Bảng tương đối thể hiện tỷ lệ mắc bệnh về mắt của các bệnh nhân ở mỗi độ tuổi. Mỗi thanh thể hiện một độ tuổi khác nhau và các màu trong thanh biểu thụ một quốc gia. Qua đó bác sĩ có thể biết được các nhân tố chiếm ưu thế về số lượng bệnh nhân mắc các bệnh về mắt và độ lịch của các nhóm trong những năm qua.
* Khi đã nắm rõ các thông tin cơ bản về sự phân phối nhân khẩu, các bác sĩ có thể bắt đầu tìm hiểu các mối tương quan tồn tại trong dữ liệu. Ngay lúc này đây, giai đoạn 2 của việc khai phá dữ liệu bắt đầu. Trong suốt giai đoạn tiếp, bác sĩ sẽ khai phá các triệu chứng nào thường xảy ra trong các trường hợp mắc bệnh mắt. Nhóm nghiên cứu đã sử dụng phương pháp khai phá luật kết hợp để tìm ra các mối tương quan đó: Cho *I={i­1,…,in­}* là tập các thuộc tính và *D* là tập các bộ dữ liệu. Mỗi bộ dữ liệu chứa một tập các thuộc tính *I*. Vấn đề của việc khai phá luật kết hợp là phải tìm ra được các tập luật có độ hỗ trợ và tin cậy lớn hơn độ hỗ trợ và tin cậy tối thiểu đã đưa ra.



* Bảng trên cho thấy số lượng các luật kết hợp khai phá được. Nó cũng cho thấy sự gia tăng các luật một cách chóng mặt các luật kết hợp qua từng năm khi sử dụng công cụ CBA. Đồng thời số lượng này cũng làm cho các bác sỹ cảm thấy quá tải và mất đi sự quan tâm với chúng. Giai đoạn thứ 3 là bước đưa ra các phân tích để quyết định các nhân tố gây nên các bệnh về mắt. Trong giai đoạn này, nhóm nghiên cứu đã sử dụng thuật toán LCD. Thuật toán dựa trên việc kiểm tra tính phụ thuộc, tính độc lập và điều kiện độc lập để hạn chế khả năng ngẫu nhiên tạo nên mối quan hệ giữa các biến. Về cơ bản thì kỹ thuật này là một điều kiện Markov. Định nghĩa điều kiện Markov: *Cho A là một nút ngẫu nhiên trong mạng ngẫu nhiên Bayes và B là một nút bất kì nhưng không phải là con cháu của A trong cùng một mạng ngẫu nhiên. Điều kiện Markov xảy ra khi A và B độc lập.*
* Khi một điều kiện Markov xảy ra, thì kéo theo các trường hợp sau. Giả sự A không phụ thuộc vào ai. Nếu B phụ thuộc vào A thì B phải được A tạo thành, mặc dù có tểh là gián tiếp. Sau đó có một biến thứ 3 là C phụ và cả A lẫn B thì cả 3 biến đều nằm trên một chuỗi ngẫu nhiên. A không phụ thuộc biến nào nên nó đúng đầu chuỗi nhưng ta lại không biết C có phụ thuộc B hay không và ngược lại. Tuy nhiên, nếu A và C trở nên độc lập trên B thì ta có thể kết luận rằng B tạo nên C. Giả sử nếu B và C độc lập nhưng vẫn có tương quan với A thì B và C không phải là một đường ngẫu nhiên nữa nhưng A lại nằm trên đường ngẫu nhiên với chúng. Nếu B và C phụ thuộc C thì theo điều kiện Markov, chúng không phải là con cháu của A. Nên ta có thể kết luận B và C tạo ra A.
* Trong ứng dụng này, nhóm nghiên cứu dùng thuật toán LCD để xác định các mối quan hệ ngẫu nhiên đa nhân tố. Thuật toán giả định điều kiện phụ thuộc và độc lập sử dụng thống kê X2.

|  |
| --- |
| Input: Một tập các luật M có dạng (A→B), A chứa 2 hoặc nhiều hơn 2 thuộc tính điều kiện. Một tập các luật có dạng (C→B) nhưng C lại chứa 1 thuộc tính điều kiện.  Output: Danh sách các mối quan hệ ngẫu nhiên  For each rule (A→B) in M  For each rule (C→B) in S  If C   Perform CI(A,B,C): X2 (AB|C=0) + X2 (AB|C=1)  Endfor  If all chi-square threshold X^2  Output ‘A causes B’  Endfor |

* Việc xác định điều kiện phụ thuộc và độc lập được thực hiện như sau:
* Kiểm tra tính phụ thuộc: Cho s(0,1) là một ngưỡng hỗ trợ và c(0,1)

1. Kết luận
2. **Nect**