**BÁO CÁO GIAI ĐOẠN 1**

*Thời gian dự kiến 09/09/2012 – 09/10/2012*

**Nội dung công việc**

[**1.** **ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU VÀO Y HỌC:** 2](#_Toc337061582)

[1.1. Mục tiêu: 2](#_Toc337061583)

[1.2. Các nghiên cứu: 2](#_Toc337061584)

[1.2.1. Khai pha dữ liệu trong y học: 2](#_Toc337061585)

[1.2.2. Tầm quan trọng và lợi ích của khai phá dữ liệu mang lại: 4](#_Toc337061586)

[1.3. Khó khăn và thử thách: 4](#_Toc337061587)

[**2.** **BỆNH TIỂU ĐƯỜNG:** 4](#_Toc337061588)

[**3.** **CÁC THUẬT TOÁN ĐÃ ĐƯỢC ỨNG DỤNG VÀO KHAI PHÁ DỮ LIỆU CỦA BỆNH NHÂN MẮC BỆNH TIỂU ĐƯỜNG:** 4](#_Toc337061589)

1. **ỨNG DỤNG CỦA KHAI PHÁ DỮ LIỆU VÀO Y HỌC:**

Với những thành công của việc kết hợp khai phá dữ liệu vào xây dựng các ứng dụng trong thương mại điện tử, marketing và bán lẻ đã cho thấy một tiềm năng lớn của việc trích xuất tri thức từ cơ sở dữ liệu ( Knowledge Discovery in Database – KDD) trong tất cả các lĩnh vực hiện nay.

Trong chương này, nhóm nghiên cứu sẽ nêu ra những kỹ thuật KDD hiện tại, các ứng dụng được tích hợp khai phá dữ liệu trong y tế và sức khỏe cộng đồng, các vấn đề quan trọng và thử thách trong việc kết hợp khai phá dữ liệu với y học. Ngoài ra còn nêu lên sự phát triển của Khai phá dữ liệu tại Việt Nam.

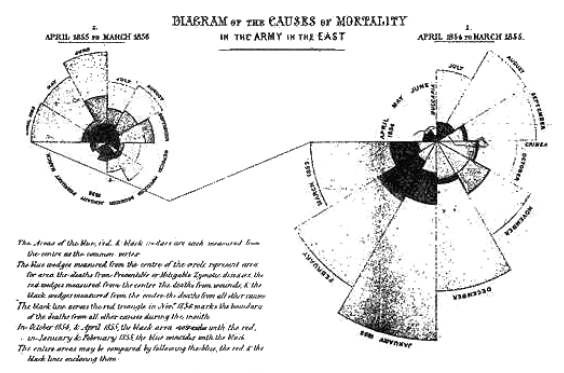
Phương pháp luận chính là dựa trên những bài báo đã được xuất bản từ năm 2000 trở lại đây trong nhiều lĩnh vực như y khoa, khoa học máy tính, kỹ thuật máy tính.

* 1. Mục tiêu:
* Liệt kệ các lợi ích và tầm quan trọng của khai phá dữ liệu trong y học.
* Tìm kiếm các kỹ thuật khai phá dữ liệu đã được sử dụng trong những lĩnh vực khác nhưng có thể áp dụng vào y học.
* Nếu lên sự phát triển của một lượng lớn các ứng dụng được kết hợp với khai phá dữ liệu.
* Xác định các vấn đề và thử thách trong ứng dụng khai phá dữ liệu vào y học.
* Đưa ra những khuyến nghị cho việc thu thập trị thức từ cơ sở dữ liệu thông qua khai phá dữ liệu.
  1. Các nghiên cứu:
     1. Khai pha dữ liệu trong y học:
* Có thể nói, việc sử dụng dữ liệu thật để hỗ trợ cho việc ra quyết định trong y học ( Evidence Bases Medicine – EBM) đã tồn tại trong nhiều thế kỷ.
* Năm 1854, John Snow, cha để của dịch tễ học hiện đại đã sử dụng các bản đồ đồ thị để phát hiện ra mầm bệnh thổ tả và đã chứng minh rằng bệnh này lây lan qua hệ thống bơm nước. Snow đã đếm số lượng bệnh nhân và vẽ sơ đồ địa chỉ của các bệnh nhân trên bản đồ bằng các thanh mà đen. Lúc đó ông đã phát hiện rằng hầu hết các ca tử vong đều nằm xung quanh một điểm bơm nước cố định tại London.



*Bản đồ mà John Snow dùng để phát hiện mầm bệnh*

* Năm 1855, Florence Nightingale đã phát minh ra biển đồ phân cực để tìm ra nhiều ca tử vong của binh lính do mất vệ sinh trong khám lâm sàn và có thể ngăn ngừa được. Bà đã sử dụng biểu đồ này để thuyết phục chính phủ cải cách các chính sách và việc này đã làm giảm đi đáng kể số lượng các ca tử vong.



*Biểu đồ phân cực do Nightinale phát minh*

* Snow và Nightingale đã tự mình thực hiện đầy đủ các bước: thu thập dữ liệu, sàng lọc và phân tích thông qua các dữ liệu về tỉ lệ tử vong trong suốt thời gian nghiên cứu vì số lượng dữ liệu có thể quản lý được. Nhưng hiện nay, việc bùng nổ dân số, sự toàn cầu hóa và tốc độ phát bệnh của bệnh dịch làm cho việc thao tác dữ liệu gần như không thể thực hiện.
* Đó cũng là nguyên nhân tại sao khai phá dữ liệu trở nên hữu ích trong y học. Tuy còn phát triển khá châm những vẫn được áp dụng để giải quyết các vấn đề khác nhau trong việc trích xuất tri thức. Có thể nói việc ứng dụng khai phá dữ liệu trong y học là một lĩnh vực còn rất non trẻ.
* Năm 2003, Wilson đã đưa ra một kết luận hết sức bất ngờ sau khi đã thực hiện một nghiên cứu trong các trường hợp mà KDD và các kỹ thuật khai phá dữ liệu đã được áp dụng trong y học. Cho đến thời điểm này, vẫn có nhiều tác giả cho rằng khai phá dữ liệu chỉ là một quá trình thu thập tri trức, còn một số khác thì cho rằng khai phá dữ liệu sử dụng kỹ thuật thống kê quá trình thu thập tri thức.
* Những khái niệm sai lầm này vẫn tồn tại trong cộng động y khoa. Trong đó có một định nghĩa tạm chấp nhận về khai phá dữ liệu ngày nay đó là một quy trình và kỹ thuật trong khai phá mô hình và hướng của dữ liệu (Witten and Frank 2005).
  + 1. Tầm quan trọng và lợi ích của khai phá dữ liệu:
* Tuy có nhiều sự khác biệt trong hướng tiếp cận nhưng lĩnh vực y học vẫn rất cần đề khai phá dữ liệu. Không chỉ trong y tế cộng đồng mà còn có y tế tư nhân.
* *Tràn dữ liệu*: Tin học hóa dữ liệu y tế đã thu được nhiều tri có giá trị nhưng hiện nay có một lượng lớn đỡ liệu làm cho việc khai phá trở nên khó khăn và gần như không thể nếu phải sàng lọc và khai phá tri thức bằng tay. Trong thực tế, một số chuyên gia cho rằng những đột phá trong lĩnh vực y học đang có dấu hiệu chậm lại do độ phức tạp của dữ liệu ngày càng tăng. Nhưng cũng chính điều này đã làm cho khai phá dữ liệu trở thành một ứng cử viên sáng giá.
* *Ngăn chặn các sai sót trong bệnh viện*: Khi các tổ chức y tế ứng dụng khai pha dữ liệu, họ có thể tìm được những tri thức và có khả năng cứu sống nhiều bệnh nhân. Một nghiên cứu đang được thực hiện trong các bệnh viện tại Mỹ cho thấy rằng 87% các ca tử vong có thể được phát hiện và ngăn chặn nếu tất cả các nhân viên ( kể cả bác sĩ) cẩn thận hơn. Thông qua việc khi phá dữ liệu, các nhà quản lý bệnh viện và chính phủ có thể tìm và giải quyết các vấn đề an toàn một cách dễ dàng hơn.
* *Hoạch định chính sách trong y tế cộng đồng*: Lavrac đã kết hợp GIS và khai dữ liệu, Weka để phân tích điểm giống nhau giữa các trung tâm y tế tại Slovenia. Sử dụng khai phá dữ liệu để tìm ra các mô hình trong trung tâm y tế nhằm đưa ra các chính sách khuyến nghị đến Viện Y Tế Cộng Đồng. Tháng 10 năm 2006 đã xảy ra một sự cố đáng tiếc tại trung tâm y tế Rizal thuộc thành phố Pasig của Indonexia. Bệnh viện đã thất bại trong việc thực hiện vệ sinh nghiêm ngặt và các biện pháp khử trùng đã dẫn đến các ca tử vong của trẻ sơ sinh do bị nhiểm khuẩn. Các bác sĩ chỉ biết được chính xác nguyên nhân cho đến khi các ca tử vong trở nên thường xuyên. Sau khi kiểm tra dữ liệu của bệnh viện, Sở Y Tế đã tìm thấy 12 trên 28 trẻ được sinh vào ngày 4 tháng 10 đã tử vong vì nhiểm khuẩn. Với khai phá dữ liệu, Sở Y Tế đã có thể phát hiện các biến cố không bình thường và hạn chế các diễn biến trở nên tệ hơn
* *Tiết kiệm tiền và chi phí*: Các tổ chức đã dùng khai phá dữ liệu nhằm rút trích tri thức từ cơ sở dữ liệu với chi phí thấp nhất. Thậm chí, có nhiều tổ chức đã áp dụng KDD và khai phá dữ liệu để tìm ra các gian lận trong thẻ tín dụng và bảo hiểm. Một hệ thống bảo hiểm thuộc quốc gia Philippin đã áp dụng khai phá dữ liệu vào việc tìm ra các bệnh nhân đặc biệt trong bảo hiểm y tế.
* *Phát hiện và ngăn chặn các bệnh nguy hiểm*: Cheng đã áp dụng thuật toán phân lớp dữ liệu để phát hiện ra các trường hợp bệnh nhân có khả năng mắc bệnh tim. Năm 2008, Cao đã sử dụng khai phá dữ liệu như một công cụ để kiểm soát các thí nghiệm vacxin lâm sàn. Sử dụng khai phá dữ liệu đã giúp cho y học có thể phát hiện ra nhiều bệnh nhân có biểu hiện khác thường hơn là nhìn vào từng tập dữ liệu riêng biệt.
* *Phát hiện, quản lý dịch bệnh và đưa ra các chính sách trong y tế cộng đồng*: Trong những năm gần đây, các chuyên gia y tế đã bắt đầu tìm cách ứng dụng khai phá dữ liệu vào việc phát hiện và quản lý dịch bệnh. Năm 2006, Kellogg đã đưa ra một kỹ thuật kế hợp mô hình không gian, mô phỏng và khai phá dữ liệu không gian để tìm ra các điểm bùng phát dịch bệnh. Kết quả phân tích được dùng để đưa ra nhằm quản lý dịch bệnh bùng phát. Năm 2005, Wong đã giới thiệu một thuật toán phát hiện dịch biện khi vừa ở giai đoạn đầu là WSARE (What’s Strange About Recent Events). Thuật toán được tạo nên dựa trên các luật kết hợp và mạng Bayes.
* *Chuẩn đoán không làm tổn thương bệnh nhân và hệ thống ra quyết định*: Một số chuẩn đoán và thí nghiệm có thể làm tổn thương bệnh nhân như sinh thiết ở phụ nữ nhằm phát hiện bệnh ung thư cổ tử cung. Năm 2006, Thangavel đã sử dụng thuật toán gom cụm K-means để phân tích các bệnh nhân ung thư cổ tử cung và đã cho thấy rằng việc gom cụm dữ liệu có thể đưa ra những kết quả chuẩn đoán tốt hơn là các phương pháp hiện có. Ông cũng đã tìm thấy một số dữ liệu có thể giúp các bác sĩ trong việc ra quyết định đưa bệnh nhân đi thiến hành thiết sinh hay không.
* *Tác hại tiềm ẩn của các loại thuốc*: Một số loại thuốc và hóa chất đã được đánh giá rằng sẽ không gây hại cho con người nhưng sau một thời gian dài sử dụng thì câu trả lời lại là có. Năm 2003, Wilson đã tiết lộ rằng tổ chức US Food và Drug Administration đã sử dụng khai phá dữ liệu để tìm ra các loại thuốc có hại khi sử dụng lâu dài trong cơ sở dữ liệu của họ. Thuật toán được áp dụng có tên là MGPS (Multi-item Gamma Poisson Shrinker) đã tìm 67% thuốc có hại và nếu dùng cách bình thường thì phải mất thêm 5 năm nữa.
* Ngoài những ứng dụng ở trên, khai phá dữ liệu vẫn còn nhiều cách khác nhau để áp dụng trong lĩnh vực y tế. Những tri thức có ích vẫn còn nằm đâu đó trong cơ sở dữ liệu và đang chờ được khai phá.
  1. Khó khăn và thử thách:

1. **BỆNH TIỂU ĐƯỜNG:**
2. **CÁC THUẬT TOÁN ĐÃ ĐƯỢC ỨNG DỤNG VÀO KHAI PHÁ DỮ LIỆU CỦA BỆNH NHÂN MẮC BỆNH TIỂU ĐƯỜNG:**