ỨNG DỤNG LÝ THUYẾT NAÏVE BAYES XÂY DỰNG HỆ THỐNG HỖ TRỢ CHẨN ĐOÁN BỆNH VIÊM PHẾ QUẢN

Trình Quốc Việt¹, Phạm Minh Dương², Phạm Thị Trúc Mai³

APPLICATION NAÏVE BAYES THEORY TO BUILDING BRONCHITIS DIAGNOSIS SUPPORT SYSTEM

Trinh Quoc Viet¹, Pham Minh Duong², Pham Thi Truc Mai³

Tóm tắt – Thuật toán phân lớp Bayes áp dụng trong các trường hợp tính xác suất cao nhất để phân lớp cho đối tượng. Tham luận áp dụng thuật toán này để hỗ trợ chẩn đoán bệnh viêm phế quản nhằm giúp các bác sĩ không chuyên khoa, bác sĩ chưa có kinh nghiệm, hoặc trường hợp bác sĩ không có mặt, trong việc khám và chữa bệnh viêm phế quản một cách nhanh chóng, kịp thời và chính xác. Hệ thống hoạt động dựa vào máy suy diễn và tập luật đã được thu thập từ tri thức chuyên gia để đưa ra kết quả chẩn đoán. Tri thức của chuyên gia đã trải qua nhiều năm kinh nghiệm nghiên cứu về bệnh viêm phế quản. Tất cả tri thức thu thập được từ các bệnh án của các bệnh nhân và có sự góp ý của các chuyên gia (bác sĩ chuyên khoa Hô hấp) hoàn toàn chính xác, đáng tin cậy. Tham luận cũng đã tìm giải pháp cho việc tin học hóa các vấn đề về chẩn đoán bệnh trên thực tế, tạo điều kiện thuận lợi cho việc khám chữa bệnh, giúp bảo vệ sức khỏe và nâng cao chất lượng sống của cộng đồng.

Từ khóa: phân lớp Naïve Bayes, chẩn đoán bệnh, viêm phế quản, tin học hóa chẩn đoán bệnh.

Abstract – The Bayesian classifier algorithm is applied in the highest probability cases to classify objects. In the paper, this algorithm is

used to assist non-specialist doctors, inexperienced doctors, or in the absence of doctors in taking the examination and treatment of bronchitis quickly, timely and accurately. The inference system and the rule set were collected from experts' knowledge, who have many years of experience in studying the bronchitis, to provide diagnostic results. All the knowledge collected from the patients' medical records and the experts' opinions (the Respiratory physicians) are completely accurate and reliable. The subject has also found a solution for the computerization of the problems in diagnosing diseases in reality, this facilitates the medical examination and treatment, and also helps to protect as well as to improve the health and the life quality of the community.

Keywords: Naïve Bayes, diagnosis, bronchitis, computerized diagnosis.

I. MỞ ĐẦU

Hàng năm, thế giới có hơn bốn triệu người tử vong do các bệnh liên quan đến đường hô hấp, trong đó phổ biến nhất là viêm phế quản (Bronchitis). Viêm phế quản là bệnh thường gặp ở trẻ em, bất kể lứa tuổi nào và thời tiết nào, đặc biệt trẻ ở thành thị cũng như ở các nơi tập trung dân cư đông đúc thì tỉ lệ bệnh còn cao hơn. Thế giới hiện có khoảng 300 triệu người mắc viêm phế quản và khoảng 250.000 trường hợp tử vong vì viêm phế quản mỗi năm. Tỉ lệ mắc viêm phế quản ước tính khoảng 6-8% ở người lớn và khoảng 10-12% ở trẻ em dưới 15 tuổi. Ở Việt Nam, các nghiên cứu cho thấy, tỉ lệ mắc viêm

^{1,2,3}Khoa Kỹ thuật và Công nghệ, Trường Đại học Trà Vinh

Email: tqvietv@tvu.edu.vn

^{1,2,3}School of Engineering and Technology, Tra Vinh University

phế quản dao động từ 1,1% ở Đà Lạt cho tới cao nhất là 5,5% ở cư dân một số khu vực Hà Nội [1]. Số liệu thống kê năm 2013 của Bộ Y tế về số ca mắc bệnh cao nhất được mô tả trong Bảng 1.

Bảng 1: Thống kê các bệnh mắc cao nhất toàn quốc

Tên Bệnh	Số ca mắc (đơn vị tính: trên 10000 dân)
Viêm họng và viêm amidan	644.18
Viêm phế quản	394.15
Các bệnh viêm phổi	545.33
Viêm đường hô hấp trên	267. 36

Chính vì vậy, hiểu biết một số nguyên nhân và cách đề phòng của căn bệnh này sẽ làm giảm thiểu số ca bệnh, đề phòng được các biến chứng và nâng cao chất lượng cuộc sống. Đồng thời cũng làm giảm nguy cơ dẫn đến các biến chứng do bệnh này gây ra như hen phế quản, ung thư phế quản, viêm phổi, áp xe phổi, suy hô hấp cấp....

Để áp dụng công nghệ thông tin vào giải quyết vấn đề trên, ta cần có một hệ thống thu thập tri thức về các triệu chứng và dấu hiệu nhận biết viêm phế quản, sau đó tổ chức, biểu diễn tri thức này bằng lí thuyết Naïve Bayes trong phân lớp để xây dựng hệ chuyên gia chẩn đoán bệnh viêm phế quản. Chính vì vậy, sau khi tham khảo ý kiến của các chuyên gia công nghệ thông tin và y khoa, chúng tôi đã chọn nghiên cứu hướng: “Ứng dụng Lí thuyết Naïve Bayes trong phân lớp để xây dựng hệ thống hỗ trợ chẩn đoán bệnh viêm phế quản”.

II. TỔNG QUAN NGHIÊN CỨU

Một hệ luật dẫn được mô tả dưới dạng nếu <giả thiết> thì <kết luận>. Đây là dạng biểu diễn tri thức rất phổ biến. Mô hình biểu diễn tri thức dạng này thường bao gồm: tập các kí hiệu mô tả các sự kiện (có cấu trúc đơn giản) và tập luật dẫn. Trong đó, phần giả thiết và kết luận của luật là tập các sự kiện. Mỗi sự kiện được mô tả có cấu trúc đơn giản như (tên đối tượng – thuộc tính – giá trị). Mô hình biểu diễn tri thức dạng này có những ưu điểm là dễ hiểu, dễ giải thích vì các

khái niệm được mô tả có cấu trúc tương đối đơn giản. Có thể xây dựng cơ chế suy luận dễ dàng bằng cách áp dụng phương pháp suy diễn tiến hay lùi. Ngoài ra, các luật được biểu diễn theo mô hình này thường độc lập nhau nên việc cập nhật luật, hiệu chỉnh và bảo trì hệ thống thuận lợi. Tuy nhiên, từ những đặc điểm trong các biểu diễn tri thức dạng này tạo nên ưu điểm thì nó cũng tạo ra các khuyết điểm. Chính vì các sự kiện có cấu trúc đơn giản, trong khi tri thức của một số lĩnh vực lại trừu tượng và phức tạp, các khái niệm của lĩnh vực có quan hệ ràng buộc lẫn nhau nên mô hình biểu diễn này không thể hiện được hết tất cả những yếu tố trừu tượng đó [2].

Phương pháp trên cho thấy cách biểu diễn khá đơn giản và trực quan, có thể suy diễn theo chiến lược khác nhau: suy diễn tiến, suy diễn lùi và suy diễn hỗn hợp, gần gũi về cú pháp, có thể kiểm tra tính mâu thuẫn giữa các luật, tính mô đun cao, có nghĩa là việc thêm, sửa đổi hoặc loại bỏ các luật hoàn toàn không có ảnh hưởng tới các luật khác và cơ chế suy diễn nhưng năng suất xử lí thấp, không sử dụng được các cấu trúc dữ liệu.

Mô hình phân lớp Naïve Bayes.

Mô hình phân lớp Naïve Bayes dựa trên mô hình xác suất.

Xác định xác suất hay dự đoán xác suất có là thành viên của lớp hay không. Nền tảng dựa vào định lí Naïve Bayes:

Cho X , Y là các biến bất kì (rời rạc, số, cấu trúc...)

Dự đoán Y từ X .

Lượng giá các tham số của $P(X|Y)$, $P(Y)$ trực tiếp từ tập dữ liệu đã huấn luyện sau đó sử dụng định lí Naïve Bayes để tính $P(Y|X=x)$.

Thuật toán Naïve Bayes: Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

- D : Tập dữ liệu huấn luyện đã được vectơ hóa $x = (x_1, x_2, \dots, x_n)$

- C_i : Phân lớp i , với $i = 1, 2, \dots, m$. - Các thuộc tính độc lập điều kiện đôi một với nhau. Theo định lí Naïve Bayes: $P(C_i|X) = (P(X|C_i)P(C_i))/(P(X))$

Theo tính chất độc lập điều kiện: $P(X|C_i) = \prod_{k=1}^n P(X_k|C_i)$

Trong đó: $(x_k|C_i)$ xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân lớp i .

Các bước thực hiện thuật toán phân lớp Naïve Bayes:

- Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính $P(C_i)$ và $P(x_k|C_i)$.

- Bước 2: Phân lớp $X_{new} = (x_1, x_2, \dots, x_n)$. X_{new} ta cần tính xác suất thuộc từng phân lớp khi đã biết trước X_{new} . X_{new} được gán vào lớp có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} (P(C_i) \prod_{k=1}^n P(X_k|C_i))$$

III. GIẢI PHÁP ĐỀ XUẤT

A. Giới thiệu hệ thống

Mục đích, đối tượng sử dụng: Ứng dụng những thành quả của công nghệ thông tin vào lĩnh vực y tế, góp phần phục vụ tốt việc tìm hiểu nghiên cứu đánh giá chất lượng y tế tại các phòng khám bệnh của tỉnh Trà Vinh. Hỗ trợ công tác khám và chữa bệnh của bệnh nhân và bác sĩ hoàn toàn chính xác, đáng tin cậy. Triển khai ứng dụng thực tiễn tại các phòng khám bệnh, trạm y tế.

Dữ liệu vào ra của hệ chẩn đoán bao gồm dữ liệu đầu vào là một số triệu chứng lâm sàng và cận lâm sàng; dữ liệu đầu ra là đưa cho người dùng kết luận bệnh hoặc là đợi xác nhận của chuyên gia và có thể xem chi tiết tỉ lệ xác suất bệnh mắc phải từ những triệu chứng đầu vào dựa trên thuật toán Naïve Bayes. Chức năng của hệ thống bao gồm chức năng chẩn đoán bệnh, cập nhật thông tin các luật, cập nhật tên bệnh, cập nhật triệu chứng.

Hệ thống dựa trên tri thức của các bác sĩ chuyên khoa Nội – Hô hấp tại Bệnh viện Trường Đại học Trà Vinh có kiến thức chuyên sâu, có kinh nghiệm lâm sàng về viêm phế quản để phân tích những dấu hiệu, triệu chứng nhằm đưa ra chẩn đoán chính xác và có cách điều trị nhanh chóng và kịp thời. Hệ thống có giao diện thân thiện, dễ sử dụng, thể hiện được sự tương tác với người dùng như trong quy trình chẩn đoán bệnh thực tế tại các bệnh viện hay phòng khám. Đồng thời, chương trình cũng có ý nghĩa quan trọng trong việc giúp các bác sĩ không chuyên nội hô hấp có thể đưa ra những chẩn đoán chính xác và nhanh chóng trong tình huống chưa có bác sĩ chuyên khoa kịp thời.

Hệ thống hỗ trợ chẩn đoán được mô tả tổng thể như Hình 1.

Hình 1: Mô hình tổng thể hệ thống

Đầu tiên, người sử dụng đăng nhập vào hệ thống để biết được đây là chuyên gia hay người sử dụng bình thường. Ngoài ra, nếu không cần đăng nhập thì chỉ cần hủy bỏ đăng nhập, hệ thống sử dụng quyền của người sử dụng.

Người sử dụng thông qua giao diện chẩn đoán nhập tất cả các triệu chứng vào sau đó chọn chức năng chẩn đoán. Máy suy diễn sẽ đưa ra kết luận bệnh và lưu vào bộ nhớ làm việc (xem đây là luật). Trường hợp xác suất không đủ, người sử dụng có thể xem lại tỉ lệ xác suất bệnh có thể mắc phải.

Chuyên gia thông qua giao diện có các quyền cập nhật thông tin: bệnh, triệu chứng, luật.

B. Mô hình giải pháp

Biểu diễn tri thức viêm phế quản: Sau khi thu thập tri thức từ bác sĩ phòng khám Nội – Hô hấp, chúng tôi nhận thấy những tri thức chẩn đoán triệu chứng của căn bệnh viêm phế quản rất phức tạp. Do đó, một mô hình biểu diễn tri thức phức tạp cần phải mô tả được các tri thức thật đầy đủ và chi tiết. Trên thực tế lâm sàng, quy trình chẩn đoán của bác sĩ cho thấy phần lớn họ dựa vào các triệu chứng (triệu chứng cơ năng – triệu chứng bệnh nhân nhận thấy; triệu chứng thực thể – triệu chứng do bác sĩ phát hiện trong quá trình khám) và các kết luận lâm sàng để chẩn đoán. Đồng thời, bác sĩ cũng dùng những triệu chứng đặc hiệu để chẩn đoán phân biệt với các bệnh

lí khác. Ngoài những kiến thức kinh điển trong y học, các bác sĩ cũng sử dụng rất nhiều kinh nghiệm qua thực tế lâm sàng để giúp việc chẩn đoán được chính xác.

Mô hình biểu diễn tri thức: Qua quá trình nghiên cứu, chúng tôi nhận thấy mô hình biểu diễn tri thức trong thuật toán phân lớp của Naïve Bayes rất phù hợp để biểu diễn tri thức này. Mô hình này được xây dựng theo cách tiếp cận hướng đối tượng phù hợp với việc mô tả các triệu chứng của bệnh viêm phế quản. Đồng thời, mô hình này cũng sử dụng các luật suy diễn tương tự như các luật bác sĩ sử dụng để chẩn đoán.

Phân lớp dựa trên mô hình xác suất: Xác định xác suất hay dự đoán xác suất thành viên của lớp dựa trên dữ liệu có sẵn hay dữ liệu đã huấn luyện. Nền tảng dựa trên định lí Naïve Bayes:

Cho X, Y là các biến bất kì (rời rạc, số, cấu trúc....)

Xác định xác suất hay dự đoán Y từ dữ liệu đã huấn luyện X .

Tính giá trị các tham số của xác suất xuất hiện Y và xác suất có điều kiện X khi có Y từ dữ liệu đã huấn luyện ($P(X|Y)$, $P(Y)$).

Sử dụng định lí Naïve Bayes để tính $P(Y|X=x)$

Tập X các thuộc tính (triệu chứng) với $X = \{x_1, x_2, \dots, x_n\}$ trong đó mỗi thuộc tính lấy giá trị trong một miền xác định như integer, boolean, string. Tập X bao gồm các thuộc tính giả thiết (A_1), kết luận (A_2) và trung gian (A_3).

Ta có $Attr = A_1 \cup A_2 \cup A_3$ với A_1 các triệu chứng lâm sàng, A_2 là các kết luận về bệnh và A_3 là các triệu chứng cận lâm sàng:

$$A_1 = \{$$

1. Ngontayduitrang: Boolean;
 2. Daumui: Boolean;
 3. Sot: Boolean;
 4. Ho: Boolean;
 5. Khacdám: Boolean;
 6.
- $$\}$$

Với A_2 là các kết luận về bệnh gồm các tập thuộc tính:

$$A_2 = \{$$

1. Viemphequanman Boolean;
 2. Viemphequancap Boolean;
 3. Henphequan Boolean;
 4. Gianphequan Boolean;
 5. Viempheoi Boolean;
 6.
- $$\}$$

Với A_3 là các triệu chứng cận lâm sàng gồm các tập thuộc tính:

$$A_3 = \{$$

1. Xquangphoi Boolean;
 2. Xetnghiephuyetdo Boolean;
 3. Xetnghiemcaydam Boolean;
 4. Xetnghiemdichmanphoi Boolean;
 5. Soiphequan Boolean;
- $$\}$$

Tập hợp các luật suy diễn trên các loại sự kiện khác nhau liên quan đến các thuộc tính của đối tượng hay bản thân đối tượng. Mỗi luật suy diễn có dạng:

{các sự kiện giả thiết} $\Rightarrow$ {các sự kiện kết luận}

$ski_1, ski_2, ski_3, \dots, ski_n \Rightarrow sko_1, sko_2, sko_3, \dots, sko_n$

Quy trình chẩn đoán lâm sàng, cận lâm sàng

Bước 1: Bác sĩ thu thập thông tin bệnh nhân gồm: Họ tên, năm sinh, địa chỉ, giới tính, thời gian mắc bệnh.

Bước 2: Bác sĩ thăm khám bệnh nhân để lấy triệu chứng lâm sàng mà bệnh nhân đang mắc phải.

Bước 3: Bác sĩ xem xét các triệu chứng lâm sàng của bệnh nhân có đủ kết luận bệnh viêm phế quản không, nếu chưa đủ hoặc có nghi ngờ có liên quan đến bệnh nghiêm trọng hơn như viêm phổi, lao phổi thì qua bước 4, ngược lại qua bước 5.

Bước 4: Bác sĩ yêu cầu bệnh nhân thực hiện cận lâm sàng (X- Quang, xét nghiệm máu, xét nghiệm cấy đàm, xét nghiệm dịch màng phổi, soi phế quản). Sau khi có kết quả cận lâm sàng thì quay lại bước 3.

Bước 5: Kết luận bệnh cho bệnh nhân.

Hình 2: Mô hình chẩn đoán bệnh

Giải pháp suy luận trong chẩn đoán bệnh viêm phế quản đã mô phỏng quy trình chẩn đoán thực tế do bác sĩ chuyên gia cung cấp. Quá trình suy diễn này áp dụng các luật suy diễn phù hợp với

thực tế lâm sàng và cận lâm sàng. Danh sách các triệu chứng được thu thập từ những bệnh nhân cụ thể tại Bệnh viện Trường Đại học Trà Vinh.

Hình 3: Mô hình giải pháp

B1: Chương trình cho người dùng chọn các triệu chứng bệnh.

B2: Kiểm tra triệu chứng có bị trùng hay không. Nếu bị trùng sẽ thông báo chọn lại các triệu chứng. Ngược lại sắp xếp các triệu chứng theo mã và lấy dữ liệu từ kho cơ sở tri thức (CSTT).

B3: Thực hiện so khớp các mã triệu chứng đưa vào và tập hợp các triệu chứng từ kho CSTT, nếu kết quả so khớp bằng 100% thì kết luận bệnh và kết thúc quá trình chẩn đoán, ngược lại thực hiện tiếp B4.

B4: Áp dụng thuật toán Naïve Bayes và tính tỉ lệ các xác suất:

B4.1: Phân lớp các bệnh dựa trên luật của hệ thống.

B4.2: Tính các xác suất bệnh dựa trên các triệu chứng.

B4.3: Tính tổng các xác suất.

B4.4: Tính tỉ lệ các xác suất.

B4.5: Sắp xếp tỉ lệ xác suất theo thứ tự giảm dần.

B5: Lấy phần tử đầu tiên trong tỉ lệ nếu $\geq 60\%$ thì thêm luật vào CSTT, kết luận bệnh kết thúc

quá trình chẩn đoán; ngược lại không kết luận bệnh, đợi chuyên gia xác nhận, cho phép người dùng xem lại các tỉ lệ xác suất và kết thúc quá trình chuẩn đoán.

IV. KẾT QUẢ VÀ THẢO LUẬN

Hệ chẩn đoán này có cơ sở tri thức được thu thập từ các bác sĩ chuyên khoa và các bệnh án thực tế tại Bệnh viện Trường Đại học Trà Vinh. Các tri thức sử dụng trong hệ thống khá đầy đủ, các luật được mô tả trong hệ thống chính xác với tri thức của chuyên gia cung cấp, vì vậy hệ thống chẩn đoán chính xác các trường hợp bệnh.

Kết quả chẩn đoán của hệ thống dựa trên dữ liệu bệnh nhân tại Bệnh viện Trường Đại học Trà Vinh với số mẫu sử dụng là 120 bệnh án, trong đó 90 mẫu dùng để khai thác luật, 30 mẫu để kiểm thử hệ thống.

Bảng 2: Thống kê kết quả thử nghiệm

	Chẩn đoán đúng theo bệnh án	Chẩn đoán sai
Số mẫu thử nghiệm hệ thống: 30	28/30 93.33%	2/30 6.67%

Như vậy kết quả chạy chương trình thử nghiệm chẩn đoán chính xác khoảng 93.33% từ các mẫu bệnh án thực tế. Kết quả này cho thấy hệ hỗ trợ chẩn đoán bệnh viêm phế quản này có hiệu quả thực tế cao.

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đạt Anh, Ngô Quý Châu, Nguyễn Lâm Việt, Phạm Quang Vinh (2012), *Bệnh học nội khoa tập 1*, Nhà Xuất bản Y học.
- [2] Hoàng Kiếm, Đỗ Văn Nhơn, Đỗ Phúc (2005), *Giáo trình Các hệ cơ sở tri thức*, Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM.