



Trường Đại học Công Nghệ- Đại học Quốc Gia Hà Nội

Khoa Công Nghệ Thông Tin

XÂY DỰNG MÔ HÌNH VÀ HỆ THỐNG DỰ BÁO DỊCH TẢ KHU VỰC HÀ NỘI DỰA TRÊN KỸ THUẬT KHAI PHÁ DỮ LIỆU

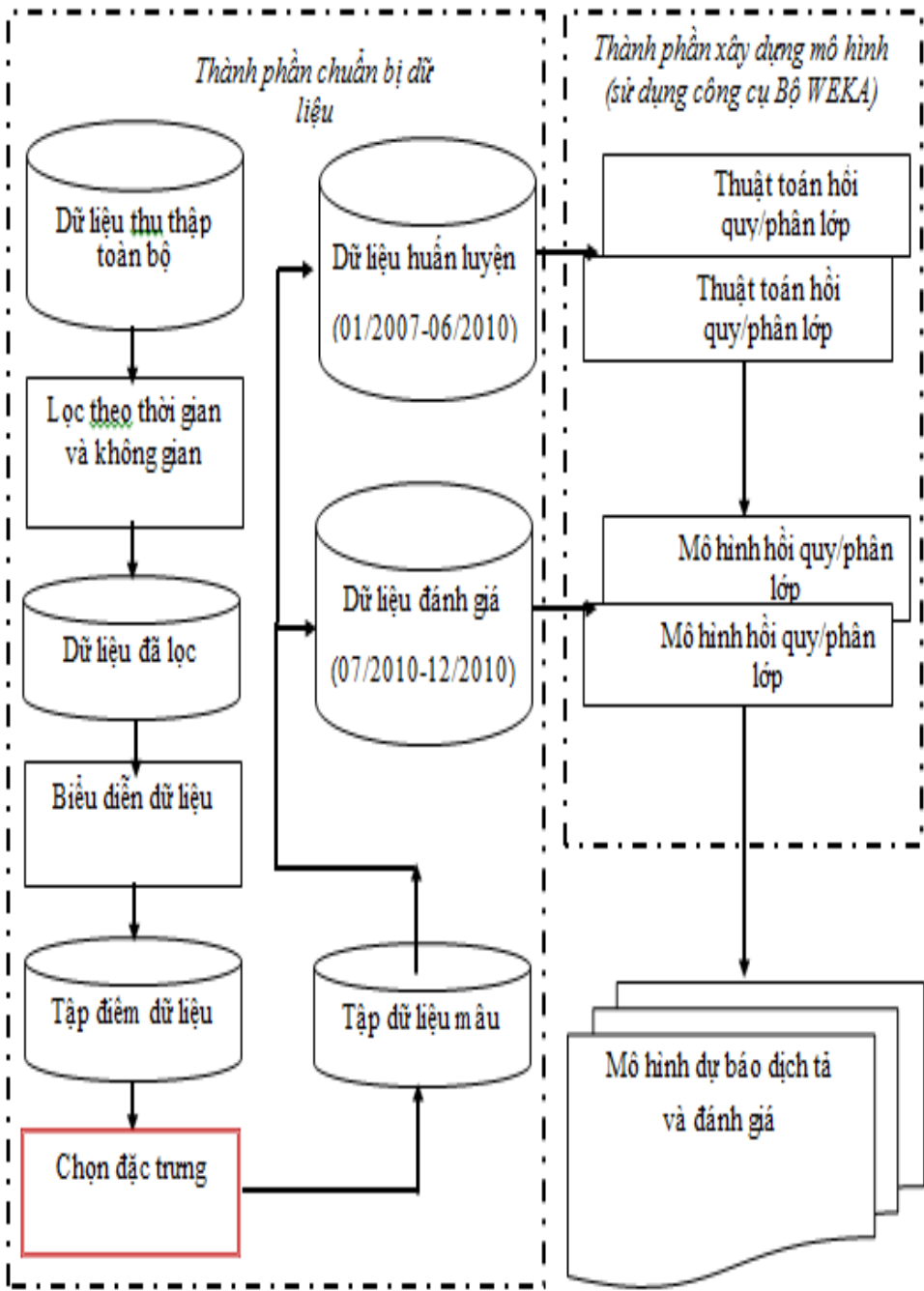
Sinh viên: Ngô Thị Oanh, Chu Chí Quang
Lại Thị Huyền Trang, Nguyễn Thị Dung

Cán bộ hướng dẫn: PGS.TS Hà Quang Thụy
Th.S Lê Hoàng Quỳnh

1. Giới thiệu

- Dịch tả:**
 - Là một trong những dịch bệnh nghiêm trọng nhất trên thế giới
 - Là loại dịch bệnh nhạy cảm với những thay đổi khí hậu
 - Dự báo tình trạng tả là chủ đề nghiên cứu được quan tâm đặc biệt.
- Ba loại phương pháp mô hình hóa dịch tả:**
 - Phương pháp dựa trên sử dụng dịch tễ toán học
 - Phương pháp dựa trên sử dụng kỹ thuật khai phá dữ liệu
 - Phương pháp dựa trên sử dụng hệ thống đa tác nhân
- Nội dung nghiên cứu:**
 - Xây dựng một mô hình dự báo dịch tả cho Hà Nội dựa trên kỹ thuật khai phá dữ liệu
 - Xác định và so sánh hai phiên bản biểu diễn dữ liệu (cục bộ và toàn cục)
 - Đánh giá mối quan hệ giữa biến mục tiêu(giá trị dịch tả thời điểm t) với các biến điều kiện khí hậu (giá trị khí hậu thời điểm t-1, t-2).

2. Mô hình hệ thống



3. Phương pháp tiếp cận mô hình

- Biểu diễn dữ liệu:**
 - Mô hình cục bộ: ($KH_{i,t-2}$, $KH_{i,t-1}$, $KHLC_{i,t-1}$, $KHLC_{i,t-2}$, $DT_{i,t-1}$, $DT_{i,t-2}$, $DTLC_{i,t-1}$, $DTLC_{i,t-2}$, $\underline{DT_{i,t}}$)
 - Mô hình toàn cục (KH_{t-2} , KH_{t-1} , DT_{t-1} , DT_{t-2} , $\underline{DT_t}$)
- Xây dựng mô hình:**
 - Thuật toán áp dụng: Linear Regression, Naïve Bayes, SVM, và Random Tree.
 - Tập dữ liệu học: Tập dữ liệu sớm hơn; Tập dữ liệu kiểm thử: Tập dữ liệu muộn hơn.
 - Kiểm tra và so sánh hai phiên bản biểu diễn dữ liệu: toàn cục và cục bộ
 - Kiểm tra trường hợp biến điều kiện chỉ chứa các đặc trưng khí hậu
 - Đánh giá cả hai trường hợp giá trị dịch tả {0, 1} và {0, 1, 2}

3. Thực nghiệm và kết quả

- Tập dữ liệu:**
 - Số ca tả theo ngày (2007-2010) → số ca tả theo tháng → các giá trị {0, 1} hoặc {0, 1, 2},
 - Dữ liệu khí hậu: giá trị trung bình, giá trị lớn nhất, giá trị nhỏ nhất của nhiệt độ không khí, lượng mưa, độ ẩm, số giờ nắng, tốc độ gió (2001-2012)
 - Dữ liệu mực nước sông tại ba trạm đo Hà Nội (1960-2012), Sơn Tây và Thượng Cát (1960-2013)

Biểu diễn dữ liệu cục bộ:
Bảng 1: Kết quả mô hình cục bộ cho hai quận Đống Đa và Hai Bà Trưng

Quận/huyện	Các độ đo	Linear Regression	NaiveBayes	LibSVM	Random Tree
Đống Đa	CC	-0.0713			
	MAE	22.8332	0.2504	0.2222	0.3333
	RMSE	26.5469	0.4741	0.4714	0.5774
	Precision		0.5830	0.4440	0.7220
	Recall		0.6670	0.6670	0.5000
	F-Measure		0.6110	0.5330	0.5280
Hai Bà Trưng	CC	0.2739			
	MAE	13.9127	0.364	0.3333	0.3333
	RMSE	15.2025	0.5782	0.5774	0.5774
	Precision		0.3330	0.2500	0.4170
	Recall		0.5	0.5	0.5
	F-Measure		0.3970	0.3330	0.4520

- Biểu diễn dữ liệu toàn cục:**
Bảng 2: Kết quả mô hình sử dụng tất cả các biến: Naïve Bayes, LibSVM, RandomTree

Số lớp	Các độ đo	NaiveBayes		LibSVM		RandomTree	
		t-12	t-1	t-12	t-1	t-12	t-1
Ba lớp	MAE	0.1406	0.0958	0.0958	0.0958	0.1145	0.1860
	RMSE	0.2913	0.3095	0.3095	0.3095	0.3082	0.3935
	Precision	0.7190	0.7330	0.7330	0.7330	0.7330	0.7190
	Recall	0.7010	0.8560	0.8560	0.8560	0.8560	0.7010
	F-Measure	0.7100	0.7900	0.7900	0.7900	0.7900	0.7100
Hai lớp	MAE	0.1558	0.1437	0.1437	0.1437	0.1718	0.2354
	RMSE	0.3791	0.3790	0.3790	0.3790	0.3775	0.4223
	Precision	0.7330	0.7330	0.7330	0.7330	0.7330	0.8890
	Recall	0.8560	0.8560	0.8560	0.8560	0.8560	0.7760
	F-Measure	0.7900	0.7900	0.7900	0.7900	0.7900	0.8060

Bảng 3: Kết quả mô hình sử dụng tất cả các biến: Linear Regression

	Độ dài thời gian	t-12	t-1
Correlation coefficient		0.4857	0.5197
Mean absolute error		2.8307	2.6200
Root mean squared error		5.1425	5.0157

Bảng 4: Kết quả mô hình chỉ sử dụng biến khí hậu: Naïve Bayes, LibSVM, RandomTree

Số lớp	Các độ đo	NaiveBayes		LibSVM		RandomTree	
		t-12	t-1	t-12	t-1	t-12	t-1
Ba lớp	MAE	0.0958	0.0975	0.0958	0.0958	0.1315	0.1900
	RMSE	0.3095	0.3093	0.3095	0.3095	0.3261	0.3973
	Precision	0.7330	0.7330	0.7330	0.7330	0.7330	0.7190
	Recall	0.8560	0.8560	0.8560	0.8560	0.8560	0.7010
	F-Measure	0.7900	0.7900	0.7900	0.7900	0.7900	0.7100
Hai lớp	MAE	0.1437	0.1437	0.1437	0.1437	0.3363	0.5660
	RMSE	0.3790	0.3790	0.3790	0.3790	0.5322	0.7178
	Precision	0.7330	0.7330	0.7330	0.7330	0.7200	0.6330
	Recall	0.8560	0.8560	0.8560	0.8560	0.7010	0.3790
	F-Measure	0.7900	0.7900	0.7900	0.7900	0.7100	0.4700

Bảng 5: Kết quả mô hình chỉ sử dụng biến khí hậu: Linear Regression

	Độ dài thời gian	t-12	t-1
Correlation coefficient		-0.0179	0.4699
Mean absolute error		8.2019	12.3981
Root mean squared error		10.5614	14.4287

Bảng 6: So sánh kết quả thực nghiệm giữa 2 mô hình toàn cục và cục bộ cho 5 quận-huyện có nhiều ca nhiễm tả

Quận/huyện	Độ tương quan CC		Độ đo F1 tốt nhất	
	Cục bộ	Toàn cục	Cục bộ	Toàn cục
Đống Đa	-0.0713	0.9178	0.6110	0.7780
Hai Bà Trưng	0.2739	0.7374	0.4520	0.3970
Thanh Xuân	0.1829	0.9568	0.7580	0.7410
Hoàng Mai	0.5317	0.0067	0.6110	0.6110
Ba Đình	-0.0233	0.9862	0.6670	1.0000
Trung bình năm quận-huyện			0.6198	0.7054

Nhận xét: Mô hình với biểu diễn dữ liệu toàn cục cho kết quả tốt hơn biểu diễn dữ liệu cục bộ

4. Một số nghiên cứu liên quan

- Tại Việt Nam:**
 - Bình Minh Nguyen và cộng sự. [1]: phát hiện vi khuẩn tả gây bởi *V.cholerae* O1 Ogawa serotype bằng việc phân tích 70 *V.cholerae* phân lập
 - Nghiên cứu dịch tả tại Bắc Ninh [2]: đánh giá sự liên quan giữa yếu tố môi trường-kinh tế-xã hội tới tỉ lệ mắc, phân bố, lây truyền dịch tả ở Bắc Ninh
 - Michael Emch và cộng sự. [3]: công bố kết quả nghiên cứu về tác động của các yếu tố môi trường địa phương tới dịch tả tại Bangladesh và Việt Nam.

-> **Ba nghiên cứu trên chưa đề cập đến việc xây dựng mô hình dự báo dịch tả ở Việt Nam.**

- Công trình tham chiếu:** Yujuan Yue và cộng sự. [4]
 - Mục tiêu: phân tích mối quan hệ giữa tỉ lệ dương tính *V.cholerae* và yếu tố khí hậu giai đoạn 2008- 2009, cửa sông Châu Giang, Quảng Châu, Trung Quốc.
 - Phương pháp: phân tích mối quan hệ giữa tỉ lệ dương tính *V.cholerae* tại ô quan sát (i, j) với các ô lân cận theo tám hướng tại mỗi nhịp thời gian và yếu tố khí hậu

-> **Do tương đồng về địa lý, đồng thời mô hình của Yue[4] có đề cập tới tác động của yếu tố khí hậu. Như vậy, mô hình dịch tả được tham chiếu như mô hình dịch tả cục bộ ở Hà Nội.**

[1] Nguyen, B.M. et al.: Cholera outbreaks caused by an altered *Vibrio cholerae* O1 El Tor biotype strain producing classical cholera toxin B in Vietnam in 2007 to 2008. J. Clin. Microbiol. 47(5), 1568–1571 (2009)

[2] Nguyễn Phương Thuộc (2012). Báo cáo Đề tài KCBN-(17)11 “Đặc điểm dịch tễ học và một số yếu tố nguy cơ gây mắc bệnh tả tại Bắc Ninh. Ứng dụng hệ thống thông tin địa lý (GIS) trong kiểm soát và dự báo dịch tả trên địa bàn tỉnh”, tháng 3/2011-tháng 5/2012.

[3] Emch, M. et al. : Local environmental predictors of cholera in Bangladesh and Vietnam. Am. J. Trop. Med. Hyg. 78(5), 823–832 (2008)

[4] Yue, Y. et al. : Influence of climate factors on *Vibrio cholerae* dynamics in the Pearl River

5. Kết luận

- Kết quả và đóng góp**
 - Nghiên cứu thực nghiệm đầu tiên xây dựng mô hình dự báo dịch tả khu vực Hà Nội sử dụng kỹ thuật khai phá dữ liệu
 - Biểu diễn dữ liệu toàn cục tốt hơn biểu diễn dữ liệu cục bộ
 - Mô hình kết quả cho độ đo F1 khoảng 0.79 - 0.86 trong những trường hợp tốt nhất.
 - Mối quan hệ giữa biến mục tiêu và biến điều kiện khí hậu cũng được chỉ ra với hệ số tương quan xấp xỉ 0.50.
- Hạn chế và định hướng nghiên cứu tiếp theo:**
 - Xây dựng mô hình dự báo theo đơn vị tuần nên được thực hiện.
 - Xây dựng mô hình dự báo theo đơn vị ngày nên được xem xét trong mỗi đợt bùng phát dịch

- Công bố khoa học:**
- Ngoc-Anh Le Thi, Thi-Oanh Ngo, Huyen-Trang Lai Thi, Hoang-Quynh Le, Hai-Chau Nguyen and Quang-Thuy Ha (2015). An Experimental Study on Cholera modeling in Hanoi. 8th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2016):March, 2016, Da Nang, Viet Nam