USING DATA MINING FOR PREDICTION AND TIME SERIES FORECASTING: A LITERATURE REVIEW

1. Data mining trong bài toán dự đoán

Khám phá tri thức (Knowledge discovery) là một quá trình tự động tìm ra những kiến thức tiềm ẩn hữu ích trong các cơ sở dữ liệu lớn [7]. Quá trình này gồm các bước sau: xác định vấn đề và không gian dữ liệu của vấn đề (Problem understanding and data understanding), chuẩn bị dữ liệu (Data Preparation), khai thác dữ liệu (Data mining), đánh giá (Evaluation), triển khai (Deployment). Trong đó, khai thác dữ liệu là một bước rất quan trọng của quá trình trên. Một trong những nhiệm vụ chính của khai thác dữ liệu là dự đoán (predictive). Bài toán dự đoán sẽ sử dụng một vài thông tin đã biết để dự báo thông tin chưa biết hoặc thông tin trong tương lai. Để giải quyết bài toán này, các nhà nghiên cứu đã áp dụng nhiều kỹ thuật của data mining như phân lớp (classify), hồi quy (regression), phát hiện sự thay đổi/lạc hướng (discriminant analysis). Phần sau sẽ trình bày một số công trình nghiên cứu về bài toán dự đoán (prediction).

1.1. Các công trình nghiên cứu đã có

Đã có rất nhiều công trình nghiên cứu về bài toán dự đoán được công bố. Các công trình này chủ yếu tập trung vào xây dựng hệ thống dự báo cho một lĩnh vực cụ thể như y khoa, tài chính,… Một vài trong số đó tiến hành so sánh các phương pháp dự đoán và đưa ra các nhận xét cho từng phương pháp này.

IC Yeh et al. (2009) so sánh 6 kỹ thuật data mining (discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, classification trees) trong dự báo nguy cơ vỡ nợ thẻ tín dụng. Khi cho vay tín dụng, người ta muốn dự báo: **1) Nên hoặc không nên cho vay tín dụng? 2) Ước tính xác suất vỡ nợ khi cho vay**. IC Yeh đã đề xuất phương pháp "Sorting Smoothing Method" để dự báo và đưa ra các giải pháp cho lĩnh vực trên với 2 câu hỏi sau:

1) Độ chính xác phân lớp khi áp dụng 6 kỹ thuật data mining (trả lời cho câu hỏi 1)?

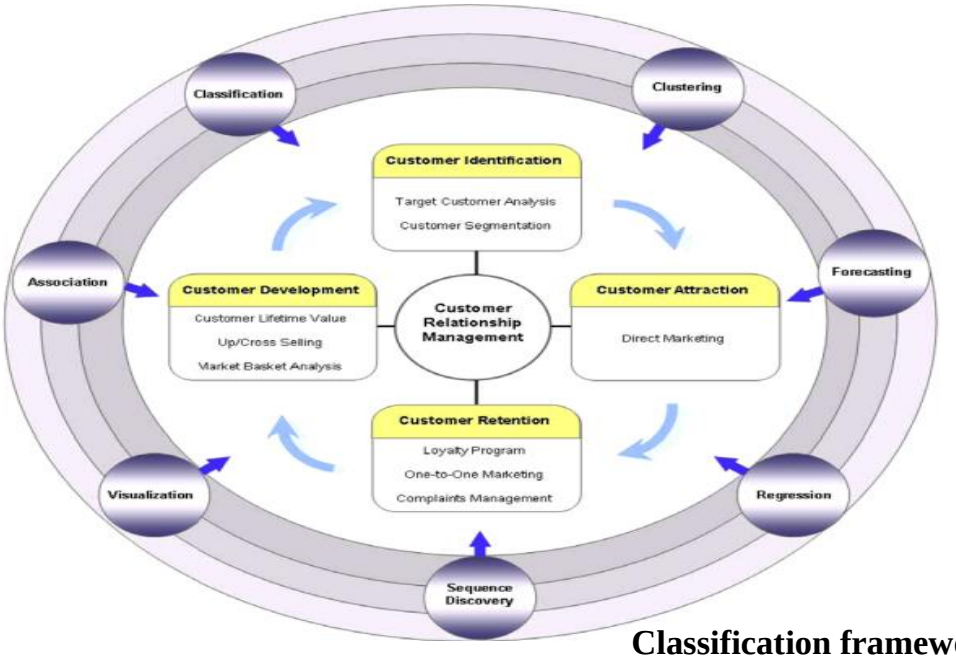
2) Có thể ước lượng xác suất vỡ nợ từ các phương pháp data mining không? (trả lời câu hỏi 2).

Bài báo dựa vào 23 thuộc tính X1 – X23 [1] và sử dụng 6 kỹ thuật data mining để phân lớp (cho vay hay không cho vay). Để ước lượng xác suất vỡ nợ, phương pháp "Sorting Smoothing Method (SSM)" được đề xuất với công thức ước lượng xác suất như sau:

$$P_i = \frac{Y_{i-n} + Y_{i-n+1} + \cdots + Y_{i-1} + Y_i + Y_{i+1} + \cdots + Y_{i+n-1} + Y_{i+n}}{2n+1}$$

Trong đó, Pi là xác suất ước lượng, Yi là biến nhị phân thể hiện có nguy cơ vỡ nợ hay không (0 – không có nguy cơ, 1 – có nguy cơ), n là số lượng dữ liệu quan sát.

Một review về hệ thống quản lý khách hàng (customer relationship management) từ năm 2000 đến 2006 được trình bày ở công trình [2]. Bài báo giới thiệu khái niệm, đặc điểm của customer relational management (CRM). Nội dung chính của bài báo là trình bày các công trình nghiên cứu liên quan đến CRM và giới thiệu CRM framework – framework hỗ trợ phân tích tính cách và hành vi khách hàng để hỗ trợ cho việc triển khai các chiến lược quản lý khách hàng, tìm ra nhóm khách hàng mang lại lợi ích lớn cho công ty. CRM bao gồm bốn module: xác định, thu hút, duy trì, phát triển khách hàng.



Classification framework for data mining techniques in CRM

Một ứng dụng phổ biến của bài toán prediction là dự đoán khả năng sống sót của bệnh nhân ung thư vú được trình bày ở công trình [3]. Ở công trình bày, nhóm tác giả Abdelghani Bellaachia sử dụng 3 kỹ thuật là mô hình xác suất (Naïve bayes), mạng neural lan truyền ngược (back-propagated neural network), cây quyết định (C4.5 decision tree) để xây dựng mô hình dự báo. Sau thực nghiệm với dataset SEER[1], kỹ thuật C4.5 cho dự báo đạt độ chính xác cao nhất. Trước đó, Dursun Delen et al. (2004) [5] thực hiện một so sánh 3 kỹ thuật data mining là cây quyết định (C5), mạng neural (neural network), chuỗi hồi quy (logistic regression) trong dự đoán khả năng sống sót của bệnh nhân ung thư vú. Kết quả thực nghiệm trên dữ liệu SEER (từ năm 1973 – 2000) cho thấy C5 cho kết quả tốt nhất đạt 93.6%, mạng neural với độ chính xác 91.2% và logistic regression là 89.2%. Ở thực nghiệm này, nhóm tác giả sử dụng kỹ thuật k-fold cross-validation (chia dataset thành 10 phần và đánh giá chéo) để đánh giá mô hình.

Một ứng dụng của bài toán dự báo trong lĩnh vực y khoa nữa được Sellappan Palaniappani et al. (2008) [6] giới thiệu là hệ thống dự đoán bệnh tim thông minh (Intelligent Heart Disease Prediction System). Intelligent Heart Disease Prediction System (IHDPS) sử dụng phương pháp CRISP-DM[2] để xây dựng mô hình khai thác dữ liệu. Hệ thống này bao gồm 6 bước: xác định vấn đề (business understanding), xác định dữ liệu (data understanding), chuẩn bị dữ liệu (data preparation), xây dựng mô hình dự báo (modeling), đánh giá mô hình (evaluation), triển khai ứng dụng (deployment). Mô hình dự báo trong hệ thống này được xây dựng với 3 kỹ thuật là Naïve bayes, mạng neural, cây quyết định.

Bài toán dự báo ngoài những hệ thống mới được xây dựng, các nhà nghiên cứu cũng thường xuyên review các phương pháp, kỹ thuật đã sử dụng. Công trình [8] là một bài review như thế. Bài báo này trình bày những vấn đề và một số hướng nghiên cứu trong tương lai về dự đoán trong y học lâm sàng (chuẩn đoán bệnh).

[1] https://seer.cancer.gov/statfacts/html/breast.html
[2] http://www.sv-europe.com/crisp-dm-methodology/

Tại hội nghị nghiên cứu của đại học De La Salle, hệ thống phân loại âm nhạc được trình bày [10]. Các tác giả sử dụng data mining và SQL (Standard Query Language) để phân tích dự đoán. Thuật toán J48, BFTree, Random Tree được lựa chọn để xây dựng mô hình dự đoán. Kết quả thực nghiệm cho thấy, thuật toán J48 có độ chính xác dự đoán cao nhất. J48 là thuật toán tốt để phân loại âm nhạc trong thực nghiệm của tác giả.

[11] [12] là các luận văn có thể trích dẫn các khái niệm, review, một số công trình mới của họ.

1.2. Nhận xét

Bài toán dự báo có rất nhiều ứng dụng trong các lĩnh vực khác nhau. Một số kỹ thuật data mining được sử dụng phổ biến trong xây dựng mô hình dự đoán là cây định danh, mô hình xác suất, mô hình hồi quy, mạng neural. Những bài toán dự báo thông tin chưa biết từ những thông tin đã biết (các thuộc tính đã biết) thường sử dụng thuật toán cây quyết định và đạt độ chính xác cao hơn các thuật toán khác [10] [3] [5].

2. Các công cụ dự báo trong chuỗi dữ liệu thời gian

Dự báo chuỗi dữ liệu thời gian là một lĩnh vực quan trọng trong bài toán dự đoán. Bài toán này quan sát thêm một yếu tố là thời gian, dự báo thông tin trong tương lai. Có nghĩa là dựa vào các thông tin đã biết (các giá trị trước đó và hiện tại), các nhà nghiên cứu sẽ xây dựng một số mô hình có khả năng dự báo trước các thông tin cần biết ở một thời điểm xác định (t+1, t+2,…). Phần sau trình bày một số phương pháp dự báo trong chuỗi dữ liệu thời gian của một số nghiên cứu đã được xuất bản.

2.1. Các công trình nghiên cứu đã có

Antti Sorjamaa et al. (2007) [13] đã đề xuất một phương pháp dự báo dài hạn trong chuỗi thời gian (long term for time series), có nghĩa là sẽ dự báo ở nhiều mốc thời gian khác nhau trong tương lai (khác với short term là chỉ dự báo một mốc thời gian kế tiếp trong tương lai). Phương pháp này kết hợp chiến lược dự đoán và lựa chọn input đầu vào của các kỹ thuật: k-nearest neighbors approximation method (k-NN), mutual information

(MI), nonparametric noise estimation (NNE). Phương pháp này đã áp dụng thành công trên thực tế với ứng dụng dự báo nguồn điện cần sử dụng của Ba Lan.

Có 2 chiến lược để dự báo dài hạn trong chuỗi thời gian là chiến lược dự báo đệ quy (recursive prediction strategy) và chiến lược dự báo trực tiếp (direct prediction strategy):

- **Chiến lược dự báo đệ quy**: việc xây dựng mô hình dự đoán là xây dựng một hàm số có dạng:

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, \ldots, y_{t-M+1})$$

Để dự đoán giá trị tiếp theo, hàm số được sử dụng có dạng:

$$\hat{y}_{t+2} = f_1(\hat{y}_{t+1}, y_t, y_{t-1}, \ldots, y_{t-M+2})$$

Có thể thấy, để dự đoán được giá trị ở một mốc thời gian nào đó, chiến lược dự đoán này phải tính được giá trị ở mốc thời gian trước đó.

- **Chiến lược dự báo trực tiếp**: Mô hình dự đoán có dạng:

$$\hat{y}_{t+h} = f_h(y_t, y_{t-1}, \ldots, y_{t-M+1}) \quad \text{with } 1 \leqslant h \leqslant H.$$

Chiến lược này xây dựng hàm số có thể tính giá trị dự báo trực tiếp tại một mốc thời gian mà không cần phải tính các mốc thời gian trước đó.

Ở bài báo này, các tác giả sử dụng chiến lược thứ 2 – chiến lược dự báo trực tiếp. Để xây dựng các mô hình dự báo, nhóm tác giả cần phải lựa chọn dữ liệu đầu vào thật tốt và dựa vào 3 tiêu chí: **1) tối thiểu hóa các ước lượng sai số của k-NN 2) tối đa hóa mối quan hệ tương hỗ giữa giá trị input và output 3) tối thiểu hóa độ nhiễu.** Sau khi chọn ra được các input cần thiết, mô hình dự báo được xây dựng với kỹ thuật LS-SVM[3] là một mô hình phi tuyến. Kết quả thực nghiệm cho thấy, độ lỗi dự báo MSE khá thấp và độ chính xác dự báo cao.

[3] J.A.K. Suykens, T.V. Gestel, J.D. Brabanter, B.D. Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002

Năm 2002, hệ thống suy diễn mờ dựa với tên là Dynamic evolving network fuzzy inference system (DENFIS) [15] được giới thiệu cho bài toán dự báo chuỗi thời gian. Hệ thống này có đặc điểm rất thông minh là các luật mờ (luật suy diễn) được tạo ra và cập nhật trong quá trình hoạt động. Tại mỗi thời điểm dự báo, output của DENFIS được tính toán thông qua hệ thống suy diễn mờ dựa trên tập mờ (tập luật mờ có thể được cập nhật trong quá trình hoạt động).

Thuật toán SVM là một mô hình học có giám sát và được sử dụng phổ biến trong các bài toán dự báo chuỗi thời gian. Francis E.H. Tay [16] đã xây dựng hệ thống dự đoán tài chính với SVM. Mục tiêu của bài báo là so sánh độ chính xác của hệ thống dự đoán SVM với mạng neural lan truyền ngược. Hệ thống được thử nghiệm ở Chicago Mercantile Market. Sau thực nghiệm, tác giả nhận thấy hệ thống dự báo với SVM cho độ lỗi ít hợn mạng neural lan truyền ngược. Năm 2003, Kyoung-jae Kim [18] cũng đã thực nghiệm tương tự và cũng đạt kết quả tốt với SVM.

Một bài toán dự báo liên quan đến kinh tế, đó là dự báo giá điện hằng ngày được trình bày ở công trình [17]. Việc dự báo giá điện là hết sức cần thiết đối với các nhà sản xuất điện tiêu thụ trong thị trường đầy cạnh tranh. Nó giúp các công ty có thể thực hiện những chiến lược phù hợp để tối ưu hóa lợi ích của mình. Bài báo đề xuất 2 công cụ dự báo chuỗi thời gian là dynamic regression and transfer function models.

- Hướng tiếp cận hồi quy động: giá điện tại thời điểm t liên quan đến giá ở các thời điểm trước đó t-1, t-2,… và nhu cầu tiêu thụ điện ở thời điểm t, t-1 t-2,… Mô hình dự đoán giá điện ở thời điểm t như sau:

$$p_t = c + \omega^d(B)\, d_t + \omega^p(B)\, p_t + \varepsilon_t$$

  Trong đó, pt là giá điện tại thời điểm t, c là hằng số, dt là nhu cầu tiêu thụ điện tại thời điểm t, $\omega^p(B)$ và $\omega^d(B)$ là các hàm đa thức thể hiện mối liên hệ giữa pt, dt, ε_t là độ lỗi.

- Hướng tiếp cận dựa trên hàm chuyển đổi: mô hình hàm chuyển đổi được thể hiện dưới dạng biểu thức sau:

$$p_t = c + \omega^d(B)\, d_t + N_t$$

Trong đó, pt là giá điện tại thời điểm t

c là hằng số

dt là nhu cầu điện tại thời điểm t

$\omega^d(B) = \sum_{l=0}^{K} \omega_l^d B^l$ là hàm số thể hiện mỗi quan hệ dt và pt

Bài báo thực nghiệm trên dữ liệu thực tế ở Tây Ban Nha và California. Dựa trên thực nghiệm cho thấy độ lỗi dự báo ở Tây Ban Nha khoảng 5%, California khoảng 3%. Nhìn chung với 2 hướng tiếp cận đề xuất, bài báo cũng đã xây dựng một hệ thống dự báo tiềm năng.

Năm 2001, tác giả G. Peter Zhang [19] xây dựng mô hình dự báo lai (hybrid) dựa trên sự kết hợp mô hình tuyến tính dự báo chuỗi thời gian Autoregressive integrated moving average (ARIMA) và mạng neural network nhân tạo (ANN). Kết quả thực nghiệm cho thấy sự kết hợp này mang lại hiệu quả dự báo tốt hơn.

- Mô hình ARIMA: trong mô hình này, giá trị của biến tương lai được giả sử là một hàm tuyến tính của các giá trị đã biết trước đó và giá trị lỗi ngẫu nhiên. Mô hình này là một hàm số như sau:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

Với yt và εt là giá trị thực và độ lỗi ngẫu nhiên tại thời điểm t

$\phi_i \ (i = 1, 2, \ldots, p)$ and $\theta_j \ (j = 0, 1, 2, \ldots, q)$ là tham số của mô hình và thường có giá trị nguyên. Dựa vào các giá trị lịch sử, thay đổi tham số mô hình sao cho đạt mô hình mong muốn.

2.2. Nhận xét

Bài toán dự báo chuỗi thời gian là một bài toán quan trọng và khó khăn trong dự đoán. Các bài toán này tính toán các giá trị tương lai (ở một thời điểm hoặc nhiều thời điểm trong tương lai). Các kỹ thuật được sử dụng phổ biến là thuật toán SVM, SVD, mạng neural,… Ngoài ra, các hướng tiếp cận lai được chú trọng nghiên cứu trong thời gian gần đây như mạng neural mờ, kết hợp phương pháp ARIMA và mạng neural. Để đánh giá các mô hình dự báo này, người ta dựa vào các độ đo độ lỗi (giá trị chênh lệch giữa giá trị thực và giá trị dự báo) như MSE, RMSE, MAE,…

TÀI LIỆU THAM KHẢO

[1] I-Cheng Yeh, Che-hui Lien (2009), "*The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*". Expert Systems with Applications, Volume 36, Issue 2, Part 1, March 2009, Pages 2473-2480.

[2] E.W.T. Ngai, Li Xiu, D.C.K. Chau (2009), "*Application of data mining techniques in customer relationship management: A literature review and classification*". Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 2592-2602.

[3] Abdelghani Bellaachia, Erhan Guven (2006), "*Predicting Breast Cancer Survivability Using Data Mining Techniques*". Department of Computer Science, The George Washington University, Washington DC.

[4] Chris Rygielski, Jyun-Cheng Wang, David C. Yen (2002), "*Data mining techniques for customer relationship management*". Technology in Society, Volume 24, Issue 4, November 2002, Pages 483-502.

[5] Dursun Delen, Glenn Walker, Amit Kadam (2004), "*Predicting breast cancer survivability: a comparison of three data mining methods*". Artificial Intelligence in Medicine, Volume 34, Issue 2, June 2005, Pages 113-127.

[6] Sellappan Palaniappan, Rafiah Awang (2008), "*Intelligent heart disease prediction system using data mining techniques*". International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[7] Boris Milovic, Milan Milovic (2012), "*Prediction and Decision Making in Health Care using Data Mining*". International Journal of Public Health Science (IJPHS), Vol. 1, No. 2, December 2012, pp. 69-78.

[8] Riccardo Bellazzi, Blaz Zupan (2008), "*Predictive data mining in clinical medicine: Current issues and guidelines*". International Journal of Medical Informatics, Volume 77, Issue 2, February 2008, Pages 81-97.

[9] Giovanni Seni, John F. Elder, *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010.

[10] Remedios de Dios Bulos, Georgio F. Go, Giselle O. Ling, Timothy C. Uy, Lawrence J. Yap (2014), "*Predictive Analysis Using Data Mining Techniques and SQL*". DLSU Research Congress, De La Salle University, Manila, Philippines, March 6-8, 2014.

[11] Ing. Stephen Nabareseh (2017). *Predictive analytics: a data mining technique in customer churn management for decision making*. Tomas Bata University in Zlín.

[12] Patricia Elizabeth Nalwoga Lutu (2010). *Dataset Selection for Aggregate Model Implementation in Predictive Data Mining*. The University of Pretoria.

[13] Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, Amaury Lendasse (2007), "*Methodology for long-term prediction of time series*". Neurocomputing, Volume 70, Issues 16–18, October 2007, Pages 2861-2869.

[14] Cédric Richard, José Carlos M. Bermudez, Paul Honeine (2009), "Online Prediction of Time Series Data With Kernels". IEEE Transactions on Signal Processing, Volume 57 Issue 3.

[15] Nikola K. Kasabov, Qun Song (2002), "DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction". IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 10, NO. 2, APRIL 2002.

[16] Francis E.H. Tay, Lijuan Cao (2001), "*Application of support vector machines in financial time series forecasting*". Omega, Volume 29, Issue 4, August 2001, Pages 309-317.

[17] Francisco J. Nogales, Javier Contreras, Antonio J. Conejo, Rosario Espínola (2002), "*Forecasting Next-Day Electricity Prices by Time Series Models*". IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 17, NO. 2, MAY 2002.

[18] Kyoung-jae Kim (2003), "*Financial time series forecasting using support vector machines*". Neurocomputing, Volume 55, Issues 1–2, September 2003, Pages 307-319.

[19] G. Peter Zhang (2003), "*Time series forecasting using a hybrid ARIMA and neural network model*". Neurocomputing, Volume 50, January 2003, Pages 159-175.