ỨNG DỤNG LUẬT KẾT HỢP TRONG KHAI PHÁ DỮ LIỆU CHỨNG KHOÁN

APPLYING ASSOCIATION RULES IN STOCK DATA MINING

Nguyễn Văn Chức[1], Nguyễn Hữu Phi[2]

[1]Trường Đại học Kinh tế, Đại học Đà Nẵng; chuc.nv@due.edu.vn

[2]Lớp 38H12K14, Trường Đại học Kinh tế, Đại học Đà Nẵng; nguyenhuuphi2508@gmail.com

Tóm tắt - Thị trường chứng khoán Việt Nam đang phát triển mạnh mẽ trong những năm gần đây. Số lượng các công ty niêm yết trên thị trường chứng khoán tăng lên nhanh chóng đã thu hút rất nhiều nhà đầu tư. Cùng với sự phát triển mạnh mẽ của thị trường chứng khoán, khối lượng dữ liệu sinh ra từ các giao dịch chứng khoán không ngừng tăng lên theo thời gian. Trong khối lượng dữ liệu khổng lồ này, chứa đựng nhiều tri thức tiềm ẩn rất có giá trị đối với các nhà đầu tư chứng khoán. Bài báo này tập trung nghiên cứu về kỹ thuật luật kết hợp trong khai phá dữ liệu và ứng dụng kỹ thuật này nhằm phát hiện các tri thức tiềm ẩn (các mối quan hệ, tương quan) về thay đổi (tăng, giảm) giá và khối lượng giao dịch của các mã chứng khoán. Dựa vào các tri thức phát hiện được từ mô hình khai phá dữ liệu, một công cụ sẽ được xây dựng nhằm hỗ trợ cho các nhà đầu tư chứng khoán ra quyết định một cách hiệu quả và tin cậy hơn.

Từ khóa - chứng khoán; khai phá dữ liệu; luật kết hợp; mô hình dự đoán; giao dịch.

Abstract - Vietnam stock market has been developing strongly in recent years. The rapid increase in the number of fast-growing companies posted on the stock market has attracted more investors. As a result, the volume of data generated by stock transaction continues to grow rapidly. The large data volume contains a lot of potential information that is useful to security investors. This paper focuses on studying association rule technique in data mining to apply this technique to discover potential knowledge (relationships, correlations) about the change (increase, decrease) of prices and transaction volume among stock codes Based on the knowledge discovered from data mining model, we have built a tool to support security investors in making wise and reliable decisions.

Key words - stock; data mining; association rule; predictive model; transaction.

1. Đặt vấn đề

Hiện nay, thị trường chứng khoán đang phát triển mạnh và mở rộng nhanh chóng, ngày càng thu hút một lượng lớn các nhà đầu tư chứng khoán. Dữ liệu về giao dịch chứng khoán phát sinh từng ngày, từng giờ và tăng lên một cách nhanh chóng theo thời gian. Nguồn dữ liệu khổng lồ này chứa rất nhiều tri thức tiềm ẩn (mối quan hệ, xu hướng) liên quan đến giá cả, khối lượng giao dịch, chỉ số tăng giảm của các mã chứng khoán đang giao dịch. Vấn đề đặt ra là làm sao có thể khai phá khối lượng dữ liệu lớn về giao dịch chứng khoán nhằm phát hiện các tri thức tiềm ẩn nhằm giúp cho các nhà đầu tưchứng khoán ra quyết định đầu tư có hiệu quả và tin cậy hơn. Bài báo này tập trung nghiên cứu về luật kết hợp trong khai phá dữ liệu và ứng dụng kỹ thuật này nhằm tìm ra các mối quan hệ (tương quan) về giá và khối lượng lượng giao dịch của các mã chứng khoán đang hoạt động trên sàn giao dịch HOSE. Dựa vào các tri thức phát hiện được từ kỹ thuật luật kết hợp, một công cụ sẽ được xây dựng nhằm giúp cho các nhà đầu tư ra quyết định đầu tư một cách hiệu quả và tin cậy hơn trong đầu tư chứng khoán.

2. Giới thiệu về luật kết hợp trong khai phá dữ liệu

Trong lĩnh vực Data Mining, mục đích của luật kết hợp (Association Rule - AR) là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như dưới đây [1].

Cho cơ sở dữ liệu giao dịch T gồm tập các giao dịch $t_1, t_2, \ldots, t_n$.

$T = \{t_1, t_2, \ldots, t_n\}$. Mỗi giao dịch t_i bao gồm tập các đối tượng I (gọi là itemset).

$I = \{i_1, i_2, \ldots, i_m\}$. Một itemset gồm k items, gọi là k-itemset.

Mục đích của luật kết hợp là tìm ra sự kết hợp (tương quan) giữa các items.

Những luật kết hợp này có dạng: $\mathbf{X \rightarrow Y}$

Hai tiêu chí rất quan trọng trong việc đánh giá luật kết hợp đó là độ hỗ trợ (support) và độ tin cậy (confidence).

Công thức tính độ hỗ trợ và độ tin cậy của luật kết hợp X→Y [2]:

$$Support\ (X \rightarrow Y) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

$$Confidence\ (X \rightarrow Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

Trong đó: $n(X)$: Số giao dịch chứa X;

N: Tổng số giao dịch;

Các luật kết hợp có độ hỗ trợ và độ tin cậy lớn hơn hoặc bằng độ hỗ trợ tối thiểu (min_sup) và độ tin cậy tối thiểu (min_conf) gọi là các luật mạnh.min_sup và min_conf gọi là các giá trị ngưỡng (threshold), được xác định trước khi sinh các luật kết hợp [2], [3].

3. Ứng dụng luật kết hợp trong khai phá dữ liệu chứng khoán

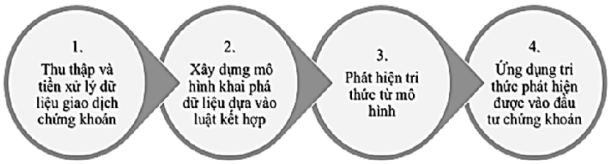
3.1. Mô tả ứng dụng

Mục đích của ứng dụng: Ứng dụng thuật toán Apriori phát hiện luật kết hợp xây dựng mô hình khai phá dữ liệu giúp phát hiện các mối quan hệ về biến động giá và khối lượng giao dịch của các mã chứng khoán trong dữ liệu giao dịch chứng khoán. Từ kết quả của mô hình khai phá dữ liệu dựa trên luật kết hợp, một công cụ được xây dựng nhằm giúp nhà đầu tư có thể sử dụng các tri thức phát hiện được, hỗ trợ ra quyết định đầu tư chứng khoán hiệu quả và tin cậy hơn.

Dữ liệu đầu vào: Dữ liệu được thu thập trên sàn giao dịch chứng khoán HOSE gồm các đặc trưng quan trọng liên quan tới các giao dịch chứng khoán như: mã chứng khoán, ngày giao dịch, giá đóng cửa, mở cửa, cao nhất, thấp nhất, khối lượng giao dịch; thông tin các công ty có niêm yết trên thị trường chứng khoán.

Đầu ra: Các tri thức phát hiện được dưới dạng luật kết hợp MaCKA→ MaCKB [Sup, Conf] thể hiện mối quan hệ liên quan đến sự tăng (giảm) theo các yếu tố như giá chứng khoán hoặc khối lượng giao dịch của các mã chứng khoán.

3.2. Quy trình triển khai luật kết hợp trong khai phá dữ liệu đầu tư chứng khoán

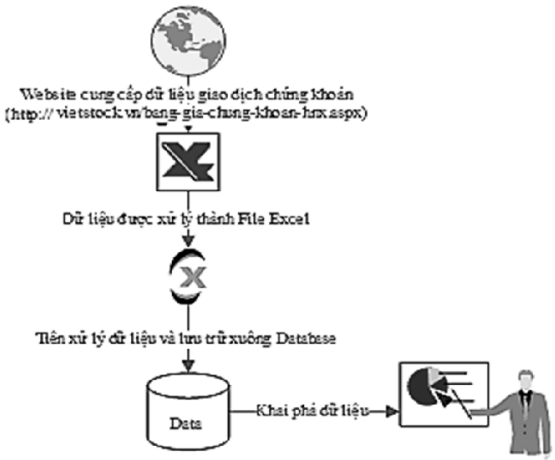


Hình 1. *Quy trình triển khai ứng dụng luật kết hợp khai phá dữ liệu chứng khoán*

Bước 1: Thu thập và tiền xử lý dữ liệu

Dữ liệu dùng để xây dựng mô hình khai phá dựa vào luật kết hợp được thu thập từ sàn giao dịch chứng khoán có cung cấp dữ liệu giao dịch hàng ngày.

Quy trình thu thập dữ liệu được mô tả trong Hình 2.



Hình 2. *Quy trình thu thập dữ liệu*

Dữ liệu thu thập ban đầu để xây dựng mô hình gồm nhiều thuộc tính, sau quá trình tiền xử lý dữ liệu, loại bỏ các thuộc tính không ảnh hưởng tới mô hình. Dữ liệu thu thập được hơn 100.000 mẫu giao dịch từ các giao dịch chứng khoán trên http://banggia.vietstock.vn/bang-gia-chung-khoan-hnx.aspx trong khoảng thời gian từ năm 2010 trở về sau [4] theo cấu trúc như Bảng 1.

Bảng 1. *Cấu trúc của dữ liệu giao dịch chứng khoán*

STT	Thuộc tính	Kiểu DL	Giá trị của thuộc tính	Giải thích
1	MaCK	Text	Các mã chứng khoán giao dịch	Mã chứng khoán
2	NgayGD	Datetime	Ngày giao dịch	Ngày giao dịch
3	MoCua	Numeric	Các giá trị số về giá mở cửa của các mã chứng khoán	Giá mở cửa
4	CaoNhat	Numeric	Các giá trị số về giá cao nhất của các mã chứng khoán	Giá cao nhất
5	ThapNhat	Numeric	Các giá trị số về giá thấp nhất của các mã chứng khoán	Giá thấp nhất
6	DongCua	Numeric	Các giá trị số về giá đóng cửa của các mã chứng khoán	Giá đóng cửa
7	KhoiLuong	Numeric	Các giá trị số về khối lượng giao dịch của các mã chứng khoán	Khối lượng giao dịch

Các bước tiền xử lý dữ liệu:

Lấy dữ liệu:

Chọn khoảng thời gian cần lấy dữ liệu;

Chọn loại khảo sát (theo giá hoặc theo khối lượng);

Thu thập dữ liệu giao dịch tương ứng với điều kiện đã thiết lập.

Tính toán thay đổi*:*

Tính toán sự giảm – tăng - đứng (không tăng không giảm) theo giá hoặc khối lượng giao dịch cho các mã chứng khoán theo điều kiện.

Tính toán và xử lý những giao dịch được thực hiện cùng nhau theo ngày.

Mã hóa dữ liệu:

Sau khi tính toán thay đổi, tiến hành mã hóa sự tăng, giảm, đứng của giá hoặc khối lượng giao dịch như sau:

Biến thiên giảm: mã hóa bằng -1

Biến thiên tăng: mã hóa bằng 1

Không biến thiên (không tăng, không giảm): mã hóa bằng 0

Kết quả cuối cùng của tiền xử lý dữ liệu như Hình 3 (với dữ liệu mẫu là 4 mã cổ phiếu ABT, ACB, ACL, AGC)

	ABT	ACB	ACL	AGC
	-1	0	1	1
	1	-1	0	1
	-1	0	0	1
	1	-1	1	1
	-1	-1	-1	-1
	1	-1	1	1
	-1	1	-1	-1
	0	0	-1	1
	0	-1	-1	-1
	1	-1	-1	1
	-1	1	-1	-1
	1	-1	1	1
	-1	-1	1	-1
	-1	1	-1	-1

Hình 3. *Kết quả tiền xử lý dữ liệu*

Bước 2: Xây dựng mô hình khai phá dữ liệu dựa vào luật kết hợp

Mô hình khai phá dữ liệu dựa vào luật kết hợp được triển khai trên môi trường lập trình Visual Studio 2010 và hệ quản trị cơ sở dữ liệu SQL SERVER 2008R2 với khả năng quản trị cơ sở dữ liệu lớn, hiệu suất cao và an toàn.

Sau khi thực hiện các thao tác tiền xử lý dữ liệu phù hợp với mô hình khai phá dữ liệu, sử dụng thuật toán Apriori để tìm các luật thể hiện các mối quan hệ về sự thay đổi giá hoặc khối lượng giao dịch của các mã chứng khoán.

Kết quả các luật được phát hiện như Hình 5 và Hình 6.

Bước 3: Phát hiện tri thức từ mô hình

Từ mô hình phát hiện luật kết hợp, các tri thức được phát hiện dưới dạng các luật:

IF X_1(a) AND X_2(a) … AND X_n(a) THEN Y(a)[Sup, Conf]

Trong đó:

X_1,X_2...,X_n: là các mã chứng khoán được chọn để dự đoán ra mã Y;

Y: mã chứng khoán cần dự đoán;

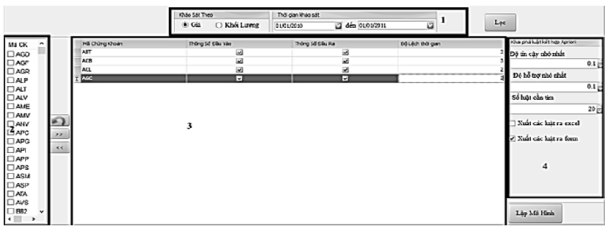
a: Trạng thái biến động giá chứng khoán (-1: giảm giá, 0: đứng giá, 1: tăng giá) hoặc khối lượng giao dịch của các mã chứng khoán;

Sup: độ hỗ trợ của luật kết hợp;

Conf: độ tin cậy của luật kết hợp.

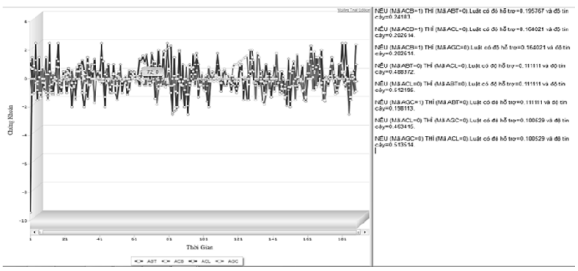
Bước 4: Ứng dụng tri thức phát hiện được vào đầu tư chứng khoán

Dựa vào tri thức phát hiện được từ mô hình luật kết hợp trên dữ liệu đầu tư chứng khoán đã xây dựng, một hệ thống giao tiếp được xây dựng cho phép nhà đầu tư có thể sử dụng các tri thức này vào việc đầu tư chứng khoán hiệu quả và tin cậy hơn như Hình 4.



Hình 4. *Công cụ khai phá luật kết hợp*

Chú thích: (1) Khoảng thời gian và đối tượng phân tích, (2) Các mã chứng khoán có tồn tại giao dịch trong khoảng thời gian được lựa chọn, (3, 4) Thông số cho mô hình.



Hình 5. *Kết quả luật kết hợp phát hiện được từ mô hình (theo giá)*

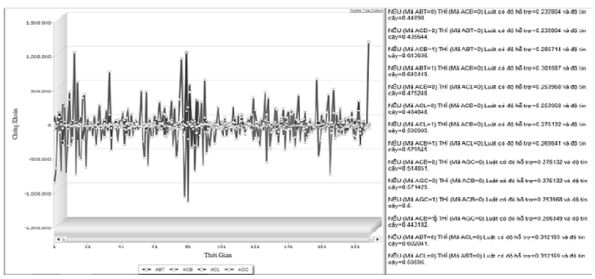
Chú thích:

(1) Biểu đồ thể hiện biến động giá của các mã chứng khoán, (2) Các luật kết hợp phát hiện được từ mô hình.

Từ mô hình trên (Hình 4), ta có 2 luật được trích xuất ra từ tập luật được xây dựng từ mô hình như sau:

Luật 1: NẾU (Mã ABT=0) THÌ (Mã ACL=0). Luật có độ hỗ trợ 0.111111 và độ tin cậy 0.488372.

Luật 2: NẾU (Mã ACL=0) THÌ (Mã ABT=0). Luật có độ hỗ trợ 0.111111 và độ tin cậy 0.512195.



Hình 6. *Kết quả luật kết hợp phát hiện được từ mô hình (theo khối lượng giao dịch)*

Luật 1: NẾU (Mã ACL=0) THÌ (Mã ACB=0). Luật có độ hỗ trợ0.253968 và độ tin cậy0.484848.

Luật 2: NẾU (Mã ACL=1) THÌ (Mã ACB=0). Luật có độ hỗ trợ0.275132 và độ tin cậy0.590909.

4. Kết luận và hướng phát triển

Khai phá dữ liệu ngày càng được sử dụng rộng rãi trong quá trình phát hiện tri thức trên khối lượng dữ liệu lớn nhằm hỗ trợ ra quyết định. Luật kết hợp là kỹ thuật được sử dụng phổ biến nhằm phát hiện các mối quan hệ (tương quan) tiềm ẩn trong khối lượng dữ liệu lớn bởi tính đơn giản, hiệu quả và nhất là nhất là khả năng biểu diễn tri thức phát hiện được dưới dạng các luật X →Y rất dễ hiểu và dễ sử dụng. Bài báo đã tìm hiểu về kỹ thuật luật kết hợp, từ đó nghiên cứu ứng dụng kỹ thuật này vào xây dựng mô hình khai phá dữ liệu nhằm tìm ra mối quan hệ về sự thay đổi giá chứng khoán (tăng, giảm, đứng giá) cũng như sự thay đổi (tăng, giảm) về khối lượng giao dịch của các mã chứng khoán. Trên cơ sở các tri thức phát hiện được từ mô hình luật kết hợp, một công cụ đã được xây dựng nhằm giúp cho các nhà đầu tư chứng khoán dễ dàng sử dụng các tri thức này hỗ trợ cho việc ra quyết định đầu tư của mình hiệu quả và tin cậy hơn. Cùng với kinh nghiệm và năng lực của các nhà đầu tư, các tri thức phát hiện được từ mô hình luật kết hợp sẽ hỗ trợ tốt hơn cho nhà đầu tư trong việc ra quyết định trong việc đầu tư chứng khoán hiệu quả và "có lý trí" hơn.

Hạn chế của nghiên cứu là do dữ liệu về chứng khoán phát sinh liên tục với khối lượng lớn, thêm vào đó, dữ liệu về giao dịch chứng khoán được cung cấp chưa đồng nhất về cấu trúc nên việc thu thập và tiền xử lý dữ liệu rất phức tạp dẫn đến làm giảm hiệu suất của mô hình. Trong thời gian tới sẽ nghiên cứu phát triển mô hình theo hướng nâng cao hiệu suất của mô hình, phát triển công cụ thu thập và tiền xử lý dữ liệu trực tuyến từ các sàn giao dịch chứng khoán cũng như kết hợp nhiều mô hình khai phá dữ liệu như phân lớp dữ liệu, phân cụm liệu, dự báo chuỗi thời gian… nhằm hỗ trợ tốt hơn cho các nhà đầu tư chứng khoán.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Đức Thuần, Nhập môn khai phá dữ liệu và quản trị tri thức, NXB Thông tin và truyền thông, 2013.

[2] Jiawei Han and Micheline Kamber, Datamining: Concepts and Techniques, Simon Fraser University, 2011.

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami; Mining Association Rules Between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993

[4] http://hsx.vietstock.vn/bang-gia-chung-khoan-hsx.aspx

[5] http://bis.net.vn/forums/data+mining

[6] http://www.stockta.com/

(BBT nhận bài: 19/10/2015, phản biện xong: 01/11/2015)