Tạp chí Tin học và Điều khiển học, T.22, S.1 (2006), 45–52

PHÁT HIỆN PHẦN TỬ NGOẠI LAI TRONG CƠ SỞ DỮ LIỆU QUAN HỆ NHỜ PHÂN TÍCH HỒI QUY

PHẠM HẠ THỦY[1], HOÀNG XUÂN HUẤN[2]

[1] *Trung tâm Tin học Kiểm toán Nhà nước*
[2] *Khoa CNTT, Trường Đại học Công nghệ Hà Nội*

Abstract. The aim of this paper is to present a method of outlier discovery in relational databases having some functional dependence between some set of attributes. In particular, when this functional dependence is linear, we could easily to use techniques of linear regression for detecting outliers of this type. This approach is illustrated by an example of detecting frauds and mistakes in audit activity.

Tóm tắt. Trong bài này chúng tôi giới thiệu một phương pháp phát hiện phần tử ngoại lai trong cơ sở dữ liệu quan hệ mà trong nó có sự phụ thuộc hàm số giữa một số thuộc tính này với một số thuộc tính khác. Khi sự phụ thuộc hàm này là tuyến tính, chúng ta dễ dàng sử dụng kỹ thuật hồi qui tuyến tính để phát hiện các phần tử ngoại lai dạng này. Phương pháp này được minh họa bằng một ví dụ áp dụng trong việc phát hiện gian lận và sai sót trong hoạt động kiểm toán.

1. GIỚI THIỆU

Các kỹ thuật học máy để phát hiện phần tử ngoại lai (outlier) đang được quan tâm nghiên cứu và được ứng dụng rộng rãi để khai thác tri thức từ dữ liệu (Data mining) nhằm trợ giúp quyết định(xem [1,2,3,5,6,7]). Mô tả một cách đơn giản, một đối tượng trong tập dữ liệu D được xem là ngoại lai khi nó khác biệt nhiều so với các đối tượng khác. Có hai loại nguyên nhân gây nên sự khác biệt này: loại thứ nhất là các dữ liệu được thu thập hoặc tạo sinh theo một quy luật khác với các dữ liệu khác và được xem là dữ liệu sai hay dữ liệu không hợp lệ, loại thứ hai là dữ liêu hợp lệ nhưng có những đặc điểm riêng biệt so với đa số dữ liệu và do đó cho ta nhưng thông tin đáng quan tâm. Để tìm ra các đối tượng khác biệt này, trước hết ta cần có một cách đo độ khác biệt của các đối tượng theo một cách đánh giá tính tương đồng nào đó (xem [1]) và sau đó là phát triển các thuật toán để tìm các đối tượng có độ khác biệt cao trong tập dữ liệu để khảo sát rõ hơn.

Để xác định mức độ khác biệt giữa các đối tượng, hai phương pháp tiếp cận chính đang được dùng là phương pháp phân tích dựa trên khoảng cách và phương pháp thống kê, trong đó phương pháp thống kê được dùng rộng rãi và có hiệu quả hơn.

Các thuật toán tìm kiếm phần tử ngoại lai theo tiếp cận thống kê đều cần biết các phân bố xác suất liên quan tới tập dữ liệu và mức độ khác biệt của các đối tượng dữ liệu được đánh giá qua xác suất xuất hiện của chúng theo các phân bố này.

Trong thực tiễn, ta thường phải làm việc với các cơ sở dữ liệu (CSDL) quan hệ với các lược đồ có cả thuộc tính số và định danh, chẳng hạn các bảng tổng hợp kế toán hoặc tổng

hợp kết quả sản xuất kinh doanh của các doanh nghiệp. Trong các thuộc tính nhận giá trị kiểu số của các bảng này có thể có một (hoặc nhiều) thuộc tính Y mà giá trị của dữ liệu của thuộc tính này phụ thuộc vào giá trị của một nhóm thuộc tính $X_1, ..., X_k$ khác theo một quan hệ hàm chứa sai số ngẫu nhiêu nào đó. Trong các trường hợp đó, các phụ thuộc hàm này có thể xác định nhờ phân tích hồi quy và ta có thể xác định các phần tử ngoại lai nhờ khảo sát các giá trị quan trắc này.

Trong bài này, chúng tôi đưa ra cách tiếp cận nhận dạng phần tử ngoại lai trong tập dữ liệu quan hệ có một thuộc tính mà giá trị của dữ liệu của thuộc tính này có phụ thuộc hàm ngẫu nhiên vào giá trị của nhóm thuộc tính độc lập khác, nhờ đánh giá sai số giữa số liệu quan trắc với giá trị hàm. Trên cơ sở đó, đề xuất hai lược đồ mở rộng ứng dụng cho trường hợp có nhiều phụ thuộc hàm ngẫu nhiên. Trong thực tế ta thường biết được dạng hàm phụ thuộc tham số của các phụ thuộc này (chẳng hạn phụ thuộc tuyến tính). Khi đó ta dễ dàng áp dụng các kỹ thuật phân tích hồi quy để xác định các phụ thuộc hàm nhờ đó tìm kiếm các phần tử ngoại lai theo tiếp cận này. Phương pháp này có thể áp dụng để trợ giúp xác định các hồ sơ cần kiểm tra trong hoạt động kiểm toán nhà nước.

Mục 2, chúng tôi đưa ra các định nghĩa về phần tử ngoại lai của tập dữ liệu có một phụ thuộc hàm (ngẫu nhiên). Các thuật toán tìm kiếm phần tử ngoại lai nhờ phân tích hồi quy tuyến tính được trình bày ở mục 3. Mục 4 giới thiệu các lược đồ giải quyết cho các trường hợp có nhiều phụ thuộc hàm và cuối cùng là phần kết luận.

2. PHẦN TỬ NGOẠI LAI TRONG CƠ SỞ DỮ LIỆU QUAN HỆ

Trong mục này, sau khi mô tả khái niệm phần tử ngoại lai, chúng tôi định nghĩa tập dữ liệu có phụ thuộc hàm ngẫu nhiên trong CSDL quan hệ và đưa ra các định nghĩa phần tử ngoại lai của các tập này.

2.1. Khái niệm phần tử ngoại lai

Trong một tập dữ liệu, thường tồn tại các đối tượng không tuân theo một hình thức hoặc một mô hình dữ liệu chung với các đối tượng dữ liệu còn lại do dữ liệu sai hay chứa đựng các thông tin đặc biệt. Các đối tượng khác biệt đó cần được xử lý đặc biệt để lấy thông tin hoặc loại bỏ khi nó là dữ liệu sai và được gọi là các phần tử ngoại lai. Tùy theo cách xét mô hình dữ liệu mà có các cách định nghĩa phần tử ngoại lai khác nhau. Barnet và Levis [2] mô tả "Một phần tử ngoại lai là một đối tượng xuất hiện không nhất quán với tập dữ liệu còn lại." còn Hawkins [3] đưa ra định nghĩa trực quan về phần tử ngoại lai là "Một đối tượng mà nó lệch hướng rất nhiều với đối tượng khác do đó dẫn đến sự nghi ngờ rằng chúng được tạo ra bởi một kỹ thuật khác."

Tùy theo thông tin quan tâm và đặc thù của tập dữ liệu mà người ta đưa ra các định nghĩa khác nhau cho phần tử ngoại lai. Nhiều tác giả (xem [5, 6]) đưa ra khái niệm về khoảng cách giữa các phần tử trong không gian dữ liệu và xem một phần tử là ngoại lai khi mật độ của dữ liệu tại điểm đang xét thấp hơn các điểm khác theo cách nhìn nào đó. Theo quan điểm thống kê thì một số tác giả khác (xem [2]) xem các điểm không cùng phân bố ngẫu nhiên với các dữ liệu khác là ngoại lai.

Trong thực tế khi làm việc với các cơ sở dữ liệu quan hệ, các phần tử (bản ghi) trong một tập dữ liệu quan hệ phải tuân theo những ràng buộc (qui tắc) cho trước nào đó thì những

phần tử không tuân theo các ràng buộc này sẽ được coi là ngoại lai. Trong bài này chúng tôi xét trường hợp phát hiện phần tử ngoại lai của tập dữ liệu quan hệ mà trong nó có sự phụ thuộc hàm ngẫu nhiên giữa các thuộc tính số. Trước hết chúng ta đưa ra định nghĩa một tập dữ liệu có phụ thuộc hàm ngẫu nhiên.

2.2. Tập dữ liệu có phụ thuộc hàm ngẫu nhiên

Xét tập dữ liệu D trong CSDL quan hệ r ứng với lược đồ $R(A_1, A_2, ..., A_n)$, có các miền giá trị của A_i là D_i tương ứng.

Giả sử D gồm N phần tử $t_1, t_2, ..., t_N$ trong đó:

$$\begin{array}{ll} t_1: & t_1(A_1), \quad t_1(A_2), ..., t_1(A_n) \\ t_2: & t_2(A_1), \quad t_2(A_2), ..., t_2(A_n) \\ ... & \\ t_N: & t_N(A_1), \quad t_N(A_2), ..., t_N(A_n) \end{array}$$

với $t_z(A_i)$ là giá trị thuộc tính A_i của đối tượng t_z. R có thể có thuộc tính định danh và số. Ở đây các thuộc tính được xét là thuộc tính số. Ta sẽ xem các giá trị $t_i(A_k)$ là một giá trị của biến ngẫu nhiên X_k nào đó.

Định nghĩa sự phụ thuộc hàm ngẫu nhiên

Tập dữ liệu quan hệ D gọi là có phụ thuộc hàm ngẫu nhiên nếu tồn tại một thuộc tính A_j và nhóm k thuộc tính (không mất tổng quát ta xem là $A_1, ..., A_k$) sao cho giá trị của dữ liệu ở các thuộc tính tương ứng có biểu diễn:

$$Y = f(X_1, ..., X_k) + W, \tag{1}$$

trong đó f là hàm thực k biến và W là nhiễu ngẫu nhiên có kỳ vọng bằng không (nhiễu), Y là ký hiệu để chỉ biến ngẫu nhiên X_j. Ta gọi f là hàm quan hệ của X_j đối với các biến $X_1, ..., X_k$.

Hàm f này có thể biết trước hoặc biết dạng hàm và sẽ được xấp xỉ bởi hàm hồi quy sẽ trình bày trong mục sau.

Ví dụ, theo quy luật lợi nhuận bình quân, trong tập dữ liệu của một báo cáo tổng hợp tài chính của một hãng sản xuất thì doanh số của hãng phụ thuộc hàm ngẫu nhiên vào chi phí đầu tư, nguyên vật liệu và lương công nhân,...

Với một tập dữ liệu D có thể có nhiều phụ thuộc hàm ngẫu nhiên tương ứng với các nhóm biến khác nhau, để đơn giản ta sẽ dùng từ phụ thuộc hàm thay cho phụ thuộc hàm ngẫu nhiên như trong các tài liệu về xác suất thống kê.

Bây giờ ta sẽ đưa ra các định nghĩa phần tử ngoại lai cho các tập dữ liệu có phụ thuộc hàm theo nghĩa trên.

2.3. Các định nghĩa về phần tử ngoại lai

Giả sử tập dữ liệu D có phụ thuộc hàm cho bởi (1) và hàm f đã biết nhờ xấp xỉ ngẫu nhiên bởi hàm hồi quy. Với mỗi đối tượng $t_i \in D$ $(i = 1, 2, ..., N)$ ta định nghĩa độ ngoại lai của đối tượng này như sai số tương đối của Y xấp xỉ bởi f.

Định nghĩa 1. (Độ ngoai lai)

Độ ngoại lai g_i của đối tượng t_i ứng với phụ thuộc hàm (1) được tính bởi:

$$g_i = \frac{|Y_i - f(x_1^i, ..., x_k^i)|}{|Y_i|}, \tag{2}$$

trong đó, Y_i là giá trị thuộc tính A_j và x_p^i là giá trị thuộc tính A_p của đối tượng t_i.

Bây giờ giả sử δ là số dương khá bé cho trước, ta có định nghĩa phần tử ngoại lai mức δ như sau:

Định nghĩa 2. (Ngoại lai mức δ)

Đối tượng dữ liệu t_p của D được gọi là phần tử ngoại lai mức δ nếu độ ngoai lai của nó không bé hơn δ:

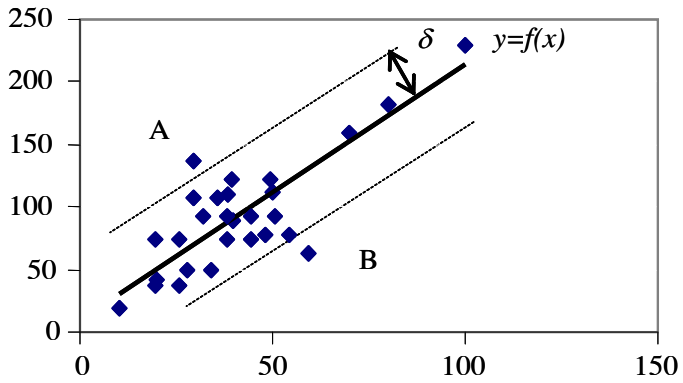
$$g_p \geq \delta. \tag{3}$$

Mức δ thường được xác định trước bởi người dùng. Bây giờ ta xét cách nhìn khác, với m là số tự nhiên nhỏ hơn N, ta định nghĩa phần tử ngoại lai $top(m)$ của D như sau.

Định nghĩa 3. (Ngoại lai $top(m)$ của D)

Đối tượng dữ liệu t_p được gọi là phần tử ngoại lai $top(m)$ của D nếu g_p thuộc vào m giá trị lớn nhất.

Hình 1 minh họa một tập dữ liệu có một phụ thuộc hàm tuyến tính của giá trị y của thuộc tính X_j đối với giá trị x của thuộc tính X_i. Các đối tượng dữ liệu ứng với các điểm A, B là phần tử ngoại lai mức δ và cũng là phần tử ngoại lai $top(2)$, còn đối tượng dữ liệu ứng với B là phần tử ngoại lai $top(1)$ của D.



Hình 1. Các cặp dữ liệu ứng với các điểm A, B là ngoại lai

Bây giờ ta sẽ mô tả các thuật toán tìm các phần tử ngoại lai theo các định nghĩa đã nêu.

3. CÁC THUẬT TOÁN TÌM PHẦN TỬ NGOẠI LAI

Mục này sẽ giới thiệu lược đồ tổng quát của thuật toán tìm phần tử ngoại lai của tập dữ liệu có một quan hệ phụ thuộc hàm (mục sau sẽ dành cho trường hợp có nhiều quan hệ), sau đó đi sâu hơn vào thuật toán dùng hồi quy tuyến tính và mô tả ví dụ ứng dụng. Thuật toán trình bày dưới đây đã định hướng cho một loạt ứng dụng trong hoạt động kiểm toán của Kiểm toán Nhà nước.

3.1. Lược đồ chung

Giả sử ta có CSDL trong đó có các tập dữ liệu có phụ thuộc hàm. Các phụ thuộc này có thể đã biết đầy đủ, nhưng thông thường thì chỉ biết dạng hàm phụ thuộc. Ở đây ta xét

trường hợp thứ hai, các phụ thuộc hàm và dạng hàm hồi quy được lấy từ ý kiến chuyên gia hoặc nhờ phân tích tương quan và hồi quy tuyến tính.

Để tìm các phần tử ngoại lai trong CSDL này, trước hết ta cần tách riêng các tập dữ liệu có phụ thuộc hàm. Việc tách này có thể được thực hiện bởi chuyên gia, chẳng hạn từ các bảng tổng hợp của các đối tượng kiểm toán cùng loại.

Với mỗi tập dữ liệu có phụ thuộc hàm, ta phân tích hồi quy để tìm hàm hồi quy của biến phụ thuộc (xem [4,8]). Các phụ thuộc hàm và dạng hàm hồi quy cũng được lấy từ ý kiến chuyên gia hoặc phân tích tương quan và hồi quy.

Khi đã có các hàm hồi quy thì tính các độ ngoại lai của mỗi đối tượng và xác định các phần tử ngoại lai theo các định nghĩa nêu ở Mục 2.3 tùy theo δ hoặc m được chọn trước, các tham số này do người dùng chọn theo kiểu hỗ trợ quyết định (xem [7]). Lược đồ này được đặc tả như sau:

Procedure Tìm kiếm phần tử ngoại lai
Begin
Tách các tập DL có phụ thuộc hàm; // theo ý kiến chuyên gia hoặc kiểm định
Xác định dạng hàm hồi quy; // cho mỗi tập dữ liệu tách được
Xác định các hàm hồi quy; // cho mỗi phụ thuộc hàm của tập DL tương ứng
Tính độ ngoại lai cho các đối tượng;//theo giá trị tham số được chọn.
Xác định các phần tử ngoại lai;// theo các tham số và định nghĩa được chọn.
End

3.2. Trường hợp hồi quy tuyến tính

Trong lược đồ trên, khâu then chốt là xác định hàm hồi quy. Khi các phụ thuộc hàm là tuyến tính, ta dễ dàng có các thủ tục xác định hàm hồi quy cho thuật toán, các thủ tục này thường đã được cài đặt trong các phần mềm thống kê. Để tiện dùng, chúng tôi mô tả tóm tắt cho trường hợp hồi qui tuyến tính đơn với ví dụ ở Mục 3.3, thủ tục hồi quy tuyến tính bội có thể xem chi tiết trong [8].

Thủ tục hồi quy tuyến tính đơn

Giả sử tập dữ liệu D có các giá trị trên các trường X và Y có phụ thuộc tuyến tính đơn:

$$Y = aX + b. \tag{3}$$

Ký hiệu các giá trị của tập N đối tượng dữ liệu của D tương ứng trên các trường này là $\{x_1, x_2, ..., x_N\}$ và $\{y_1, y_2, ..., y_N\}$. Thủ tục xác định hàm hồi quy của Y theo (3) như sau.

Bước 1. Tính hệ số tương quan giữa X và Y theo công thức:

$$\rho_{X,Y} = (M_{XY} - M_X M_Y)/\Big[(\sqrt{M_X^2 - (M_X)^2)(M_Y^2 - (M_Y)^2)}\Big], \tag{4}$$

trong đó,

$$M_{XY} = \sum_{i=1}^{N} \frac{x_i y_i}{N}; M_X = \sum_{i=1}^{N} \frac{x_i}{N}; M_Y = \sum_{i=1}^{N} \frac{y_i}{N}; M_X^2 = \sum_{i=1}^{N} \frac{x_i^2}{N}; M_Y^2 = \sum_{i=1}^{N} \frac{y_i^2}{N}.$$

Nếu $\rho_{X,Y} \geq \alpha$ với $\alpha > 0$ cho trước thì ta xem là có tương quan, $\rho_{X,Y}$ càng lớn, mối tương quan càng chặt chẽ.

Bước 2. Xác định các hệ số hồi quy a, b nhờ phương pháp bình phương tối thiểu tức là cực tiểu hóa sai số trung bình phương E để được hàm hồi quy $Y = aX + b$

$$E = \sum_{i=1}^{N} \frac{(ax_i + b - y_i)^2}{N}. \tag{5}$$

Các công thức hiện để tính a, b cũng như tính toán sai số, khoảng tin cậy của ước lượng hồi qui có thể xem [8]. Sai số trong ước lượng hồi quy có cũng ảnh hưởng tới độ ngoại lai của dữ liệu và được xét theo trường hợp cụ thể.

Sau đây chúng tôi đưa ra một ví dụ thực tế minh họa cho việc ứng dụng phương pháp phát hiện phần tử ngoại lai được trình bày ở trên.

3.3. Khi tập dữ liệu có nhiều phụ thuộc hàm

Khi một tập dữ liệu D có nhiều quan hệ phụ thuộc hàm thì ta có nhiều "tiêu chuẩn" để xác định phần tử ngoại lai và cần áp dụng lược đồ đa tiêu chuẩn để xác định mức độ ngoại lai cho các đối tượng.

Giả sử ta có k quan hệ phụ thuộc hàm với độ ngoại lai của đối tượng dữ liệu t_i theo phụ thuộc hàm thứ j là g_i^j. Sau đây là hai phương pháp đơn giản để xác định phần tử ngoại lai đa tiêu chuẩn.

a) Phương pháp tổng hợp theo trọng số

Với m khá lớn, ta tìm các phần tử ngoại lai $top(m)$ của mỗi phụ thuộc hàm. Tùy theo mỗi phụ thuộc mà ta gán cho một trọng số dương ρ_j cho phụ thuộc hàm thứ j và độ ngoại lai tổng hợp cho đối tượng dữ liệu t_i sẽ là:

$$G_i = \sum_{j=1}^{N} \rho_i g_i^j. \tag{6}$$

Các trọng số ρ_j có thể chọn thay đổi theo hướng trợ giúp quyết định.

b) Phương pháp ngưỡng bội

Mỗi phụ thuộc hàm thứ j xác định các phần tử ngoại lai mức δ_j tương ứng. Trong tập các phần tử ngoại lai tìm được, phần tử nào là ngoại lai của nhiều phụ thuộc hàm hơn thì có mức ngoại lai cao hơn. Sau đó có thể tìm m phần tử ngoại lai có mức cao nhất để xem xét.

4. THỰC NGHIỆM

Mô hình phát hiện phần tử ngoại lai nhờ phân tích hồi quy mô tả ở trên đã được chúng tôi áp dụng thử nghiệm để phát hiện những hiện tượng bất thường trong sản xuất kinh doanh thông qua báo cáo tài chính của doanh nghiệp.

Trong hoạt động kiểm toán, người ta không thể kiểm tra hết các chứng từ, sổ sách của mọi doanh nghiệp. Hiện nay, các kiểm toán viên vẫn dùng phương pháp thủ công rà soát vĩ mô các báo cáo để xác định đối tượng và các tài liệu cần kiểm toán theo phương pháp chuyên gia. Vấn đề đặt ra là cần sử dụng công cụ tin học để hạn chế các đối tượng và tài liệu cần xem xét để kiểm toán. Trong mô hình này, các phần tử ngoại lai là các đối tượng mà các kiểm toán viên nên lưu ý kỹ hơn. Để áp dụng được, trong các bản báo cáo cần tách các dữ liệu của các doanh nghiệp cùng loại để xác định các phụ thuộc hàm. Khi bảng tổng hợp là các dữ liệu tổng hợp định kỳ của cùng một doanh nghiệp thì ta có thể đặt vấn đề chúng có

phụ thuộc tuyến tính giữa doanh số và chi phí trong mỗi thời kỳ và phần tử ngoại lai là đối tượng có nghi vấn cần kiểm toán... Ta xét một ví dụ cụ thể sau.

Giả sử người kiểm toán viên thực hiện bước khảo sát để tìm ra các nội dung trọng tâm cần xem xét tại một Doanh nghiệp A. Người kiểm toán viên nghiên cứu bảng kê tổng hợp về doanh thu, chi phí nguyên vật liệu chính (phải mua ngoài và được hoàn thuế giá trị VAT), chi phí tiền lương, VAT trong các tháng của Doanh nghiệp được trình bày trong Bảng 1:

Bảng 1. Bản kê tổng hợp trong năm của doanh nghiệp A

THANG	DOANHTHU	CHI_NVL	TIENLUONG	VAT
1	1,415,420,800	566,333,320	495,627,280	40,043,332
2	1,425,358,000	570,308,200	499,105,300	40,321,574
3	1,445,760,000	578,469,000	506,246,000	40,892,830
4	1,450,267,320	580,271,928	507,823,562	41,019,035
5	1,465,890,000	586,521,000	513,291,500	41,456,470
6	1,500,540,000	600,381,000	525,419,000	42,426,670
7	1,510,567,000	604,391,800	528,928,450	42,707,426
8	1,515,680,000	680,437,000	530,718,000	48,030,590
9	1,525,678,000	606,437,000	534,217,300	42,850,590
10	1,615,244,000	716,262,600	565,565,400	50,538,382
11	1,680,800,700	672,485,280	588,510,245	47,473,970
12	1,550,526,000	620,375,400	542,914,100	43,826,278
Tổng	18,101,731,820	7,382,673,528	6,338,366,137	521,587,147

Bảng 2. Kết quả tính toán xác định ngoại lai về chi phí NVL của doanh nghiệp A

THANG	DOANHTHU x	CHI_NVL y	Giá trị Hồi qui $y = ax + b$	Độ ngoại lai g_i	So sánh với $g_i - \delta$
1	1,415,420,800	566,333,320	567,134,319	0.001	-0.04
2	1,425,358,000	570,308,200	572,269,511	0.003	-0.04
3	1,445,760,000	578,469,000	582,812,540	0.008	-0.03
4	1,450,267,320	580,271,928	585,141,763	0.008	-0.03
5	1,465,890,000	586,521,000	593,215,009	0.011	-0.03
6	1,500,540,000	600,381,000	611,120,898	0.018	-0.02
7	1,510,567,000	604,391,800	616,302,495	0.020	-0.02
8	**1,515,680,000**	**680,437,000**	**618,944,712**	**0.090**	**0.05**
9	1,525,678,000	606,437,000	624,111,323	0.029	-0.01
10	**1,615,244,000**	**716,262,600**	**670,395,850**	**0.064**	**0.02**
11	**1,680,800,700**	**672,485,280**	**704,273,223**	**0.047**	**0.01**
12	1,550,526,000	620,375,400	636,951,887	0.027	-0.01
TBình	1,508,477,652	615,222,794	615,222,794		

Để phát hiện những hiện tượng bất thường trong kê khai chi phí NVL, ta áp dụng phương pháp phát hiện phần tử ngoại lai theo tương quan tuyến tính. Mối tương quan tuyến tính này là đã được kiểm nghiệm từ thực tế. Ở đây ta chọn mối tương quan giữa doanh thu và chi phí NVL.

Sử dụng phương pháp tìm hàm hồi qui, ta xác định được mối quan hệ giữa DOANHTHU

và CHI_NVL như sau: CHI_NVL = 0.517∗DOANHTHU - 164, 304, 864.

Nếu chọn $\delta = 4\%$. Kết quả tính toán được trình bày trong Bảng 2.

Các phần tử ngoại lai (các tháng cần tập trung kiểm toán) sẽ là: T8, T10, T11. Trong đó số liệu tháng 8 là phần tử ngoại lai top(1), số liệu tháng 8, 10 là các phần tử ngoại lai top(2) và số liệu các tháng 8,10,11 là các phần tử ngoại lai top(3).

Trong các tháng này, khi kiểm tra chi tiết, các tháng 8, tháng 10 đã có hiện tượng khai thấp doanh thu để giảm nộp thuế doanh thu và đồng thời vẫn thanh toán tiền VAT được hoàn trả. Như vậy nếu kiểm toán viên chọn $\delta = 4\%$ và chọn top (3) thì có 3 phần tử nghi vấn để kiểm tra và tỷ lệ phát hiện chính xác là 2/3. Nếu chọn top(2) hoặc top(1) thì ta có tỉ lệ phát hiện chính xác là 100%. (Nếu $\delta = 6\%$ thì các phần tử ngoại lai sẽ là số liệu các tháng 8, 10 ta cũng được kết quả phát hiện chính xác là 100%). Khi chọn tham số δ cần lưu ý tới sai số hồi quy sao cho nó không ảnh hưởng nhiều tới độ ngoại lai.

Thuật toán trên đã được cài đặt chạy thử nghiệm trên môi trường IDEA- một phần mềm kiểm toán tổng quát của CASEWARE (có chứa ngôn ngữ lập trình VBA) dùng để thực hiện, phát triển những phần mềm dùng trong kiểm toán.

5. KẾT LUẬN

Trên đây chúng tôi đưa ra cách tiếp cận thống kê mới để tìm các phần tử ngoại lai trong cơ sở dữ liệu quan hệ có phụ thuộc hàm ngẫu nhiên. Các lược đồ tính toán có thể dùng để trợ giúp xác định đối tượng kiểm toán trong Kiểm toán Nhà nước.

Trong thời gian tới chúng tôi sẽ mở rộng ứng dụng thử nghiệm và nghiên cứu các mô hình tìm phần tử ngoại lai đa tiêu chuẩn có hiệu quả nhờ áp dụng các lý thuyết mạnh hơn như tập mờ hoặc tập thô.

TÀI LIỆU THAM KHẢO

[1] A. Arning, R. Agrawal, P. Raghavan, A linear method for deviation detection in large databases, *Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining*, Portland, Oregon, August, 1996, p. 164–169.

[2] V. Barnett, and T. Lewis, *Outliers in Statistical Data*, John Wiley, 3rd edition, 1994.

[3] D. Hawkins, *Identification of Outliers*, Chapman and Hall, London, 1980.

[4] D. Freedman, R. Pisani, and R. Purves, *Statistics*, W. W. Norton, New York, 1978.

[5] E. Knorr, and R. Ng, Algorithms for mining distance-based outliers in large datasets, *Proc. of the VLDB Conference*, New York, USA, September 1998, p.392–403.

[6] E. M. Knorr, "Outliers and data mining: finding exceptions in data", Doctoral thesis, Dept. of Computer science, University of British Columbia, 2002.

[7] E. Turban, *Decision Support and Expert Systems Management Support Systems*, Prentice Hall, 1995.

[8] Nguyễn Cao Văn, Trần Thái Bình, *Lý thuyết Xác suất và Thống kê toán,* Nhà xuất bản Giáo dục, 2002.

Nhận bài ngày 12 - 8 - 2005
Nhận lại sau sửa ngày 15 - 12 -2005