Tạp chí Tin học và Điều khiển học, T.22, S.3 (2006), 267–274

MỘT MỞ RỘNG THUẬT TOÁN PHÂN CỤM k-MEANS CHO DỮ LIỆU HỖN HỢP

HOÀNG XUÂN HUẤN[1], NGUYỄN THỊ XUÂN HƯƠNG[2]

[1]*Khoa Công nghệ thông tin, Trường Đại học Công nghệ*
[2]*Trường Đại học Dân lập Hải Phòng*

Abstract. Partitioning a large set of data objects into homogenous clusters is an important problem in data mining. Among clustering algorithms, k-means is well known by its advantages and especially efficient for clustering very large data sets. This algorithm and its prime variants are only to cluster numerical data sets. There are some extensions to this algorithm for clustering data sets with mixed numeric and categorical values but they omit some its advantages. In this paper, we present an improvement of k-means algorithm for clustering a data set with mixed numeric and categorical values. This algorithm conserves advantages of k-means algorithm.

Tóm tắt. Phân một tập dữ liệu thành các tập con sao cho các đối tượng trong cùng tập con tương tự nhau, còn các đối tượng thuộc các tập con khác nhau thì khác nhau theo một nghĩa nào đó là một bài toán quan trọng trong khám phá tri thức từ dữ liệu. Trong số các thuật toán phân cụm, thuật toán k-mean có nhiều ưu điểm và được sử dụng rộng rãi, đặc biệt khi phân cụm các tập dữ liệu lớn. Ban đầu thuật toán này và các biến thể của nó chỉ làm việc với dữ liệu số, sau đó đã có một số mở rộng để làm việc với dữ liệu định danh hoặc dữ liệu hỗn hợp, nhưng các thuật toán này làm mất đi một số ưu điểm của thuật toán nguyên thủy. Trong bài này, chúng tôi đề xuất một thuật toán mở rộng của thuật toán k-means cho dữ liệu hỗn hợp gọi là thuật toán k-tâm. Thuật toán này kế thừa các ưu điểm của thuật toán k-means.

1. GIỚI THIỆU

Thuật toán phân cụm k-means đầu tiên do MacQueen đề xuất năm 1967 [10] và các biến thể của nó [1] như PAM; CLARA; CLARANS) là các thuật toán phân hoạch tập dữ liệu gồm N đối tượng có n thuộc tính số thành k ($k < N$) tập con. Các thuật toán loại này đơn giản, độ phức tạp thấp và dễ song song hóa (xem [2]) nên được sử dụng rộng rãi (xem [3, 4, 9]).

Để mở rộng phạm vi sử dụng cho các dữ liệu có thuộc tính hỗn hợp hoặc định danh, trong [6, 7] Huang đề xuất các thuật toán k-prototypes và k-modes. Trong các thuật toán này, khái niệm mode được dùng làm tâm cho các tập giá trị của thuộc tính định danh và mêtric rời rạc được dùng để xác định độ tương tự trong các thuộc tính này. Nhược điểm của các thuật toán này là ở mỗi lần phân cụm lại, khi có một đối tượng thay đổi cụm thì phải tính lại mode (hay là tâm) của các cụm liên quan nên phức tạp hơn k-means và không song song hóa được. Một cách tiếp cận khác là mã hoá các thuộc tính định danh bằng số để phân cụm như là dữ liệu số nhưng khó giải thích kết quả và khó áp dụng với các tập dữ liệu lớn (xem [8]) vì số giá trị số để mã hóa sẽ lớn.

Trong bài này, chúng tôi trình bày một mở rộng thuật toán k-means cho dữ liệu hỗn hợp và gọi là thuật toán k-tâm, trong đó khái niệm mode được dùng để xác định tâm cho tập dữ liệu có thuộc tính định danh. Với mêtric và hàm mục tiêu được xét, thuật toán hội tụ tới điểm cực tiểu địa phương của hàm mục tiêu.

Ngoài phần kết luận, bài báo được trình bày như sau. Mục 2, giới thiệu tóm tắt thuật toán k-means, khái niệm mode và các thuật toán k-modes, k-prototypes cùng các vấn đề liên quan. Mục 3 trình bày thuật toán k-tâm.

2. THUẬT TOÁN k-MEANS VÀ CÁC VẤN ĐỀ LIÊN QUAN

Trong phần này chúng tôi giới thiệu thuật toán ***k*-means** (MacQueen, 1967), khái niệm mode và cách dùng nó trong các thuật toán k-modes, k-prototypes và những vấn đề liên quan. Khái niệm mode trong Mục 2.2.1 là mở rộng trực tiếp khái niệm này của Huang [6, 7].

2.1. Thuật toán *k*-means

Thuật toán k-means dùng để chia tập dữ liệu gồm N đối tượng trong không gian số học n chiều $D = \left\{x^i\right\}_{i=1}^{N}$ thành k ($k < N$) tập con. Trên R^n chọn một mêtric d, chẳng hạn mêtric Euclide, thuật toán thực hiện như sau.

Bước 1. Chọn k phần tử ban đầu $\{z^j\}_{j=1}^{k}$ của D làm tâm các cụm con.

Bước 2. Với mỗi $i = 1, ..., N$, xếp x^i vào cụm C_j nếu:

$$d(x^i, z^j) = \min\{d(x^i, z^q) \,|\, q \leqslant k\}. \tag{1}$$

Bước 3. Tính trung bình cộng của các phần của các cụm C_j làm tâm mới:

$$z^j \leftarrow \frac{1}{|c_j|} \sum_{x \in C_j} x, \tag{2}$$

trong đó $|C_j|$ là số phần tử của cụm C_j.

Bước 4. Trở lại Bước 2 để xếp lại các cụm con nhờ tâm mới cho tới khi các cụm không thay đổi.

Nếu mêtric được dùng là mêtric Euclide thì thuật toán hội tụ tới cực tiểu địa phương của hàm:

$$E = \sum_{j=1}^{k} \sum_{x \in C_j} d^2(x, z^j). \tag{3}$$

2.2. Mode của tập dữ liệu và các thuật toán *k*-modes, *k*-prototypes

Để mở rộng thuật toán cho các đối tượng dữ liệu có chứa thuộc tính định danh, trong [6, 7] Huang xét D là tập N đối tượng $\{x^i\}_{i=1}^{N}$ trong đó $x^i = (x_1^i, ..., x_m^i, x_{m+1}^i, ..., x_n^i)$ là phần tử của quan hệ r trên với lược đồ quan hệ $R = \{A_1, ..., A_n\}$ và $x_j^i \in \mathrm{Dom}(A_j)$ với mỗi $j \leqslant m$ là các giá trị thực còn với $m+1 \leqslant j \leqslant n$ là các giá trị định danh. Các thuật toán trong [6, 7] dựa trên khái niệm mode của tập dữ liệu có thuộc tính định danh và dùng mode thay cho trọng tâm của mỗi tập dữ liệu.

2.2.1. Mode của tập dữ liệu

Để tiện trình bày định nghĩa mode của tập dữ liệu hỗn hợp, chúng tôi đưa thêm định nghĩa j-mode với $j \leqslant n$.

Định nghĩa. Giả sử C là tập con của tập dữ liệu hỗn hợp D.

i) Với mọi $j \leqslant n$, j-mode của C (ký kiệu là j-mode(C)) là giá trị có tần suất nhiều nhất trong thuộc tính A_j của C nếu A_j là thuộc tính định danh và là trung bình cộng của các giá trị thuộc tính A_j của C khi A_j là thuộc tính số. Nếu A_j là thuộc tính định danh và có nhiều giá trị có tần suất như nhau trong C thì j-mode(C) có thể không duy nhất và ta chọn giá trị nào cũng được.
ii) Mode của tập hợp C ký hiệu là mode(C) là phần tử $z = (z_1, ..., z_n)$ trong đó

$$z_j = j\text{-mode}(C),\ \forall j \leqslant n. \tag{4}$$

Chú ý. Khi dữ liệu tập dữ liêu là thuộc tính định danh thì mode của tập C ở đây trùng với định nghĩa mode của Huang trong [6, 7].

2.2.2. Các thuật toán k-modes và k-prototypes

Thuật toán k-modes dùng để phân cụm dữ liệu định danh còn thuật toán k-prototypes là mở rộng của nó cho dữ liệu hỗn hợp gồm thuộc tính số và thuộc tính định danh. Trong thuật toán k-prototypes, Huang cũng dùng mêtric xác định bởi công thức (8) trong Mục 3.1 nhưng các trọng số ρ_i của thuộc tính số đều bằng 1 còn các trọng số ρ_i của thuộc tính định danh đều bằng nhau.

Sau khi định nghĩa mêtric cho các dữ liệu hỗn hợp, thuật toán k-prototypes tương tự k-means chỉ khác một điểm là khi phân bố D vào các cụm thì cần tính lại tâm (mode) của các cụm liên quan ngay khi có một đối tượng đổi cụm. Thuật toán thực hiện như sau.

Bước 1. Chọn ngẫu nhiên k phần tử $\{z^j\}_{j=1}^k$ của D làm mode cho các cụm $\{C_j\}_{j=1}^k$ tương ứng và phân mỗi đối tượng $x \in D$ vào cụm C_j mà x gần mode của nó nhất.

Bước 2. Tính lại mode cho mỗi cụm.

Bước 3. Lặp lại việc tái phân bố D theo các mode mới nhưng khi có phần tử x chuyển từ cụm C^i sang cụm C^j thì tính lại tâm của các cụm này. Thủ tục lặp được thực hiện đến khi các cụm không đổi.

Nhận xét. So với thuật toán k-means, việc tính lại mode của các cụm liên quan mỗi khi có thay đổi cụm của một đối tượng trong Bước 3 làm số phép tính tăng trong mỗi lần lặp và không song song hóa thuật toán được. Mặt khác, trong thực tế có những thuộc tính nhận giá trị định danh có thứ tự thì dùng mêtric rời rạc để xác định mêtric không phản ánh được bản chất dữ liệu.

3. THUẬT TOÁN *k*-TÂM

Trong mục này ta xét tập dữ liệu hỗn hợp D như trong mục trước nhưng phân biệt các thuộc tính làm ba loại: định danh, thứ tự và số. Sau khi xây dựng mêtric để xác định mức tương đồng, chúng tôi mô tả thuật toán và chứng minh tính hội tụ của nó.

3.1. Mêtric trên dữ liệu hỗn hợp

Trong lược đồ quan hệ R, miền giá trị của các thuộc tính A_j có thể là tập số thực, định danh hay là tập có thứ tự.

Định nghĩa 3.1.1. Giả sử Dom(A_j) là miền giá trị của thuộc tính A_j. Ta có các khái niệm sau:
i) Thuộc tính định danh: A_j được gọi là thuộc tính định danh nếu Dom(A_j) là tập không có thứ tự, tức là $\forall a, b \in \text{Dom}(A_j)$, hoặc $a = b$ hay $a \neq b$.

ii) Thuộc tính số: A_j được gọi là thuộc tính số nếu $\mathrm{Dom}(A_j)$ là tập số thực.
iii) Thuộc tính thứ tự: Nếu $\mathrm{Dom}(A_j)$ là tập hữu hạn và có thứ tự hoàn toàn thì A_j được gọi là thuộc tính có thứ tự (chẳng hạn: $\mathrm{Dom}(A_j)$={không đau, hơi đau, đau và rất đau}.

Trên miền giá trị $\mathrm{Dom}(A_j)$ của một thuộc tính A_j ta xác định các khoảng cách như sau.

Định nghĩa 3.1.2. $\forall x, y \in \mathrm{Dom}(A_j)$ ta có hàm $d_j(x, y)$ xác định bởi:
i) Nếu A_j là thuộc tính số thì $d_j(x, y) = |x - y|$. (5)
ii) Nếu A_j là thuộc tính thứ tự và $\mathrm{DOM}(A_j) = \{a_j^1, ..., a_j^k\}$ với $a_j^1 < a_j^2 < ... < a_j^k$, ta lấy một hàm đơn điệu $f_j : \mathrm{DOM}(A_j) \to [0, 1]$ sao cho $f_j(a_j^1) = 0; f_j(a_j^k) = 1$ (Hàm này có thể là: $f_j(a_j^i) = \dfrac{i-1}{k-1}$). Khi đó,

$$d_j(x, y) = |f_j(x) - f_j(y)|. \tag{6}$$

iii) Nếu A_j là thuộc tính định danh thì

$$d_j(x, y) = \begin{cases} 0 \text{ khi } x = y, \\ 1 \text{ khi } x \neq y. \end{cases} \tag{7}$$

Bây giờ ta định nghĩa khoảng cách trên D.

Định nghĩa 3.1.3. Giả sử $x = (x_1, ..., x_n)$ và $y = (y_1, ..., y_n)$ là hai đối tương dữ liệu hỗn hợp trên D, khoảng cách $d(x, y)$ được tính bởi công thức:

$$d(x, y) = \sqrt{\sum_{j=1}^{n} \rho_j^2 d_j^2(x_j, y_j)}, \tag{8}$$

trong đó các $d_j(x_j, y_j)$ được tính theo các công thức (5) - (7) và ρ_j là các trọng số dương cho bởi các chuyên gia tùy ý theo mức quan trọng của thuộc tính.

Với định nghĩa trên, ta luôn có thể xem các thuộc tính thứ tự có miền giá trị là đoạn $[0, 1]$ để tìm mode (các giá trị trên thuộc tính này của D là tập con) và nó cũng được xem là thuộc tính số khi không xảy ra nhầm lẫn. Định lý sau là mở rộng tính chất của mode trong [6, 7].

Định lý 1. *Nếu xem miền giá trị của các thuộc tính có thứ tự là đoạn* $[0, 1]$ *và mode của tập hợp xác định như đã nói ở trên thì với mọi tập dữ liệu hỗn hợp* C*, mode(*C*) cực tiểu hàm:*

$$E(x) = \sum_{y \in C} d^2(y, x), \tag{9}$$

trong đó x *là phần tử của quan hệ* r *trên lược đồ quan hệ* $R = \{A_1, ..., A_n\}$.

Chứng minh. Với mỗi đối tượng $z \in r$, ta có thể xem $z = (z^n, z^c)$ trong đó z^n là hình chiếu của z lên quan hệ có các thuộc tính số, z^c là hình chiếu của z lên quan hệ có thuộc tính định danh và ký hiệu tích vô hướng $\langle x^n, y^n \rangle = \sum\limits_j \rho_j x_j^n y_j^n$ (tổng lấy trên các thuộc tính của x^n); chuẩn $\|x^n\|$ và khoảng cách $d(x^n, y^n)$ là chuẩn Euclide và khoảng cách sinh bởi tích vô hướng này; $d(x^c, y^c)$ xác định bởi (8) với tổng lấy trên các thuộc tính (định danh) của chúng. Khi đó $E(x)$ có thể biểu diễn như sau.

$$E(x) = \sum_{y \in C} \|x^n - y^n\|^2 + \sum_{y \in C} d^2(y^c, x^c). \tag{10}$$

Mặt khác, ký hiệu M là mode(C) ta có:

$$\sum_{j\in C}\|y^n - x^n\|^2 = \sum_{y\in C}\left(\|y^n - M^n\|^2 + \|M^n - x^n\|^2 + 2\langle y^n - M^n, M^n - x^n\rangle\right)$$
$$= \sum_{y\in C}\|y^n - M^n\|^2 + |C|\|M^n - x^n\|^2 + 2\sum_{y\in C}\langle y^n - M^n, M^n - x^n\rangle$$
$$= \sum_{y\in C}\|y^n - M^n\|^2 + |C|\|M^n - x^n\|^2,$$

trong đó $|C|$ là số phần tử của C. Tổng này đạt cực tiểu khi x là mode(C). Tính cựu tiểu của tổng thứ hai trong (10) suy từ định nghĩa của mode và Định lý 1 trong [7] với chú ý rằng khi A_j là thuộc tính định danh thì từ (7) ta có

$$d_j^2(x, y) = d_j(x, y). \tag{11}$$

Định lý được chứng minh. ■

Bây giờ ta mô tả thuật toán.

3.2. Thuật toán *k*-tâm

Sau khi đã mở rộng các miền giá trị của thuộc tính có thứ tự và xác định khoảng cách giữa các đối tượng như đã nêu, thuật toán k-tâm phân cụm dữ liệu hỗn hợp thực hiện như thuật toán k-means ở Mục 3. Thuật toán được đặc tả như sau.

Proceduce *k*-tâm
Begin
Chọn các trọng số ρ_j, các hàm f_j, xác định k.
Chọn k phần tử ban đầu $\{z^j\}_{j=1}^k$ của D làm tâm các cụm
Xếp mỗi $x \in D$ vào cụm C_j mà nó gần tâm nhất;
For $j = 1, ..., k$ **do** $z_j \leftarrow$ mode(C_j);
Repeat
Phân bố lại cụm theo tâm mới // như k-means;
Cập nhật lại tâm cho các cụm // nhờ tính mode
Until các cụm không đổi;
Xác định các cụm;
End

Dễ dàng nhận được định lý sau về tính hội tụ của thuật toán.

Định lý 2. *Thuật toán k-tâm kết thúc sau một số hữu hạn bước lặp.*

Chứng minh. Xét hàm:

$$E = \sum_{j=1}^{k}\sum_{x\in c^j} d^2(x, z^j). \tag{12}$$

Ta sẽ chứng minh hàm E đơn điệu giảm trong quá trình lặp tái phân bố lại và vì số cách phân cụm là hữu hạn nên thuật toán kết thúc sau hữu hạn bước. Thực vậy, trong quá trình này E chỉ thay đổi trong hai trường hợp:
1) Khi có đối tượng x đổi từ cụm C_i sang cụm C_j mà chưa tính lại tâm.
2) Sau khi cập nhật tâm ở mỗi lần lặp.
Trường hợp thứ nhất. Ta dễ dàng thấy rằng mỗi khi có đối tượng x đổi từ cụm C_i sang cụm C_j mà chưa tính lại tâm thì E sẽ giảm một lượng là $\Delta E = d^2(x, z^i) - d^2(x, z^j)$.

Trường hợp thứ hai. Trong mỗi lần lặp, giả sử trước khi phân bố lại, tập dữ liệu gồm K cụm $C_1, ..., C_k$ với tâm $z^1, ..., z^k$ tương ứng và sau khi tái phân bố lại thành các cụm $\underline{C}_1, ..., \underline{C}_k$ với tâm $\underline{z}^1, ..., \underline{z}^k$ tương ứng.

Khi đó dễ thấy

$$\sum_{j=1}^{k}\sum_{x\in c^j} d^2(x, z^j) \leqslant \sum_{j=1}^{k}\sum_{x\in \underline{c}^j} d^2(x, z^j) \leqslant \sum_{j=1}^{k}\sum_{x\in \underline{c}^j} d^2(x, \underline{z}^j),$$

việc cập nhật lại tâm mới sẽ làm giảm E theo Định lý 1 áp dụng cho các cụm $\underline{C}_i$ tương ứng trong các số hạng ở vế phải.

Như vậy E giảm thực sự khi có ít nhất một đối tượng dữ liệu đổi cụm trong lần lặp nên không có cách phân cụm nào bị lặp lại. Mặt khác số cách phân cụm là hữu hạn nên thuật toán phải dừng sau hữu hạn bước. ■

Nhận xét

1) Khi thuật toán kết thúc, các đối tượng tâm có thể không thuộc tập D do ta đã chuyển đổi giá trị của thuộc tính có thứ tự. Để tìm phần tử đại diện cho mỗi cụm, ta lấy phần tử thuộc cụm gần với tâm của nó nhất.

2) Chứng minh Định lý 2 cũng cho thấy thuật toán hội tụ tới điểm cực tiểu địa phương của E mà không đảm bảo là cực tiểu toàn cục. Để tăng hiệu quả của thuật toán ta có thể kết hợp với thuật toán di truyền (xem [11]) hoặc khởi tạo tâm ban đầu bằng phương pháp chuyên gia.

3) So với các thuật toán k-prototypes thuật toán này có những ưu điểm sau.

• Như thuật toán k-means, thuật toán k-tâm song song hóa được như trong [2].

• Trong mỗi lần lặp để phân bố lại các cụm, ta chỉ phải cập nhật tâm một lần mà không phải cập nhật mỗi khi có đối tượng dữ liệu đổi cụm như k-prototypes.

• Nhờ phân biệt thuộc tính thứ tự và thuộc tính định danh, chất lượng cụm tốt hơn.

• Trong thực tế, mỗi thuộc tính số có thể dùng những đơn vị đo khác nhau nên việc chọn trọng số ρ_i cho mỗi thuộc tính thứ i là cần thiết.

4. VÍ DỤ ỨNG DỤNG

Chúng tôi thử nghiệm phân cụm cho hỗ trợ chẩn trị y học ở bệnh viện Việt- Tiệp Hải phòng. Dưới đây là bài toán tổng quát và kết quả ứng dụng.

Phát biểu bài toán tổng quát

Có một tập hồ sơ bệnh án của một loại bệnh với các triệu chứng (bao gồm cả kết quả xét nghiệm sinh hóa) đã biết. Theo kinh nghiệm chuyên gia có thể chia làm k nhóm có đặc điểm gần nhau để theo dõi và điều trị theo các chế độ cụ thể và cần tìm bệnh án điển hình cho mỗi nhóm. Các triệu chứng có thể là giá trị số (nhiệt độ, chỉ số sinh hóa...) hoặc sắp thứ tự (không đau, hơi đau, đau, rất đau...) hay là thuộc tính định danh như chảy máu tiêu hóa, rối loạn tiêu hóa, vàng da... Khi số triệu chứng nhiều thì phân nhóm của thầy thuốc gặp khó khăn và nhiều trường hợp không thống nhất. Về bản chất thì đây là bài toán có giám sát với kết luận mờ, tuy vậy nó vẫn có thể xem là bài toán phân cụm nửa giám sát nhờ biết trước một số trường hợp thuộc loại bệnh (cụm nào) để làm tâm khởi tạo. Sau khi phân cụm, tâm cụm sẽ dùng làm trường hợp điển hình để xác định các trường hợp mới.

Trường hợp cụ thể

Chúng tôi thử nghiệm cho loại bệnh xơ gan theo 42 hồ sơ bệnh án ở bệnh viện. Mỗi bệnh nhân có thể có 34 triệu chứng trong đó có 12 thuộc tính định danh như: phản ứng Rivaltas, nhu mô gan, phù, chảy máu tiêu hóa..., có 8 thuộc tính có thứ tự: suy nhược cơ thể (mức độ), đau sườn phải (mức độ), bụng chướng (mức độ),... và 14 chỉ số sinh hóa khác: độ dãn TMC, tỷ lệ prothrombin, almubin... Vì số thuộc tính khá nhiều nên phân loại mức độ bệnh khó, đặc biệt đối với các bệnh nhân tiến triển ở mức độ trung gian thì kết quả phân loại của các bác sĩ không giống nhau. Hiện các bác sĩ có hai cách chia nhóm bệnh để nghiên cứu điều trị:

Cách 1. Chia thành 3 nhóm bệnh:

1. Xơ gan còn bù.
2. Xơ gan giai đoạn mất bù.
3. Xơ gan mất bù dẫn đến k - gan.

Cách 2. Chia thành 4 nhóm bệnh:

1. Xơ gan tiềm tàng.
2. Xơ gan giai đoạn mất bù (mức thấp, khác nhóm 2 ở trên).
3. Xơ gan mất bù dẫn đến hội chứng chảy máu tiêu hóa do tăng áp lực tính mạch cửa.
4. Xơ gan mất bù dẫn đến k - gan.

Kết quả thực nghiệm

Thực nghiệm cho thấy việc chọn các trọng số ρ_i trong công thức 8 rất quan trọng, ảnh hưởng đến chất lượng phân cụm. Việc xác định này phụ thuộc vào kiểu giá trị của thuộc tính (số thực, nguyên và định danh), đơn vị đo và mức quan trọng của thuộc tính để xác định các loại bệnh (được xác định bởi các chuyên gia). Nếu dùng phương pháp k-prototypes thì kết quả không chấp nhận được.

Việc chọn tâm cụm đầu tiên cũng ảnh hưởng đến chất lượng phân cụm nên chúng tôi khởi tạo các tâm cụm theo gợi ý của chuyên gia (phương pháp nửa giám sát).

Kết quả phân cụm thực nghiệm theo hai cách chia nhóm trên so với cách đánh giá của các bác sĩ ở hồ sơ được giới thiệu ở Bảng 1 và Bảng 2.

Bảng 1. Kết quả phân 3 cụm, có 34 hồ sơ phân trùng loại

Cụm	Nhóm 1	Nhóm 2	Nhóm 3
Kết quả k - tâm	7	29	6
Phân cụm của hồ sơ	8	27	7
Số kết quả trùng	5	24	5

Bảng 2. Kết quả phân 4 cụm, có 30 hồ sơ phân trùng

Cụm	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4
Kết quả k-tâm	9	14	13	6
Phân cụm của hồ sơ	8	18	11	5
Số kết quả trùng	6	12	8	4

Theo đánh giá của Bác sĩ chuyên khoa thì kết quả như vậy là chấp nhận được, các trường hợp khác là do bệnh ở mức độ tiến triển chưa rõ ràng nên cách phân loại của chuyên gia trong các trường hợp này cũng không nhất quán.

5. KẾT LUẬN

Trên đây chúng tôi đưa ra một cải tiến thuật toán k-means cho dữ liệu hỗn hợp, thuật toán này dễ song song hóa và dễ áp dụng cho tập dữ liệu rất lớn. Kết quả ứng dụng của thuật toán mới ở mức thử nghiệm cho nghiên cứu của bệnh viện. Cũng như k-means, nó có nhược điểm là kết quả phân cụm phụ thuộc nhiều vào khởi tạo ban đầu. Để khắc phục nhược điểm này, ta có thể chọn khởi tạo theo kiểu nửa giám sát như đã nêu trên hoặc kết hợp với thuật toán di truyền. Đối với phân cụm cho các loại bệnh, nên phát triển theo hướng phân cụm mờ vì các triệu chứng bệnh tiến triển không rõ ràng.

TÀI LIỆU THAM KHẢO

[1] P. Andritsos, *Data Clusting Techniques*, Department of Computer Science, University Toronto, 2002.

[2] S. Kantabutra, A. L. Couch , Parallel k-meanss clustering algorithm on NOWs, *Technical Journal* **1** (2000) 243–248.

[3] C. D. Looney, Pattern recognition using neural network, *Theory and Algorithm for Engineers and Scientist*, New york, Oxford, 1997.

[4] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, 2001.

[5] Hoàng Xuân Huấn, Case-based reasoning with rough features, VNU, *Journal of Science, Nat., Sci.& Tech.* **21** (2005) 24–32.

[6] Z. HUANG, Clustering large data sets with mixed numeric and categorical values, *Proc. 1st Conference of PAKDD* (1997) 21–34.

[7] Z. HUANG, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery* **2** (1998) 283–304.

[8] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, *IEEE Trans. Knowledge and Data Eng.* **14** (2002) 673–690.

[9] P. Lingras, Unsupervised rough set classification using gas, *Journal of Intelligent Information System* **16** (2001) 215–228.

[10] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, *Proc. of 5th Berkely Symposium on Mathematical Statistics and Probility* (1967) 281–297.

[11] Z. Michalewics, *Genetic algorithms + Data structures = Evolutionary Programs*, Berlin, Springer-Verlag, 1996.

[12] Hoàng Hải Xanh, "Về các phương pháp phân cụm dữ liệu trong data ming", Luận văn thạc sỹ, ĐHQG Hà Nội, 2005.

Nhận bài ngày 29 - 9 - 2005
Nhận lại sau sửa ngày 30 - 6 - 2006