BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN TẤN PHƯƠNG

NGHIÊN CỨU ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG PHÂN TÍCH SỐ LIỆU DÂN CƯ

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TSKH TRẦN QUỐC CHIẾN**

Phản biện 1: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 2: **GS.TS. NGUYỄN THANH THUỶ**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 10 tháng 9 năm 2011.

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Trong vài thập niên gần đây, cùng với sự thay đổi và phát triển không ngừng của ngành công nghệ thông tin, luồng thông tin được chuyển tải mau lẹ đến chóng mặt, ước tính cứ khoảng 20 tháng lượng thông tin trên thế giới lại tăng gấp đôi. Những người ra quyết định trong các tổ chức tài chính, thương mại, khoa học…không muốn bỏ sót bất cứ thông tin nào, họ thu thập, lưu trữ tất cả mọi thông tin vì cho rằng trong nó ẩn chứa những giá trị nhất định nào đó.

Hiện nay lượng dữ liệu mà con người thu thập và lưu trữ trong các kho dữ liệu là rất lớn, những kỹ thuật truyền thống không đủ khả năng làm việc với dữ liệu thô, không thể phân tích bằng tay vì phải tốn rất nhiều thời gian để khám phá ra thông tin có ích, phần lớn dữ liệu chưa bao giờ được phân tích như nhận định của Usama Fayyad:*"Hố sâu khả năng sinh ra dữ liệu và khả năng sử dụng dữ liệu"*. Giải pháp duy nhất giúp phân tích tự động khối lượng dữ liệu lớn đó là kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu ứng dụng rộng trên toàn thế giới, với kỹ thuật KDD, tác giả muốn nghiên cứu ứng dụng trong phân tích số liệu dân cư ở Việt Nam để phát hiện những tri thức về tăng trưởng dân số.

Vấn đề tăng trưởng dân số quá nhanh ở Việt Nam trong những thập niên gần đây được sự quan tâm rất lớn của các cấp lãnh đạo, điển hình là việc chính phủ Việt Nam đưa ra chính sách kế hoạch hoá gia đình "Mỗi gia đình chỉ có 1 hoặc 2 con". Đã có nhiều biện pháp xử lý những gia đình vi phạm chính sách kế hoạch hoá gia đình, nhưng qua đợt thống kê dân số gần đây nhất vào năm 2009 còn rất nhiều gia đình

vi phạm chính sách kế hoạch hoá gia đình (sinh trên 2 con). Những gia đình vi phạm chính sách có những đặc điểm chung nào?

Với lượng lớn dữ liệu thu thập được qua mỗi đợt thống kê dân số tại Việt Nam, việc ứng dụng khai phá dữ liệu trong phân tích số liệu dân cư là cần thiết để phát hiện những đặc điểm chung về các gia đình vi phạm chính sách kế hoạch hoá gia đình, hỗ trợ lãnh đạo ban dân số kế hoạch hoá gia đình các cấp đưa ra biện pháp phù hợp, tôi quyết định chọn đề tài:

"Nghiên cứu ứng dụng khai phá dữ liệu trong phân tích số liệu dân cư".

2. Mục đích nghiên cứu

Mục đích của đề tài là tìm hiểu các kỹ thuật khai phá dữ liệu, nghiên cứu ứng dụng kỹ thuật khai phá dữ liệu trong phân tích số liệu dân cư, nhằm phát hiện các đặc điểm chung của những gia đình vi phạm chính sách kế hoạch hóa gia đình, hỗ trợ cho các cấp lãnh đạo có những nhận định để đưa ra biện pháp phù hợp.

3. Đối tượng và phạm vi nghiên cứu

- Tìm hiểu lý thuyết về phát hiện tri thức và khai phá dữ liệu
- Quản lí và tổ chức lưu trữ cơ sở dữ liệu từ số liệu thống kê dân số tại tỉnh Quảng Nam.
- Nghiên cứu một số mã nguồn mở áp dụng trong khai phá dữ liệu.
- Áp dụng kỹ thuật khai phá dữ liệu trên cơ sở dữ liệu lưu trữ.

4. Phương pháp nghiên cứu

- Thu thập số liệu thống kê dân số từ nguồn dữ liệu thống kê dân số tại tỉnh Quảng Nam
- Chọn phương pháp khai phá dữ liệu thích hợp.
- Lựa chọn công nghệ cài đặt chương trình.

- Phân tích và kiểm định kết quả đạt được.

5. Ý nghĩa khoa học và thực tiễn

- Cung cấp một cách nhìn tổng quan về phát hiện tri thức và khai phá dữ liệu.
- Áp dụng các thuật toán khai phá dữ liệu trên cơ sở dữ liệu thống kê dân số ở Việt Nam. (Dữ liệu thu thập từ nguồn dữ liệu thống kê dân số tại tỉnh Quảng Nam)
- Tìm ra các đặc điểm chung của những gia đình vi phạm chính sách kế hoạch hóa gia đình hỗ trợ các nhà lãnh đạo có những nhận định cụ thể.
- Chương trình được sử dụng cho lãnh đạo ban dân số kế hoạch hóa gia đình các cấp.

6. Cấu trúc của luận văn

Chương 1: Giới thiệu khái niệm, tính chất, các bước trong quá trình khai phá dữ liệu. Phương pháp, dạng cơ sở dữ liệu có thể khai phá và những thách thức trong quá trình khai phá dữ liệu.

Chương 2: Trình bày khái niệm và các bước trong quá trình khai phá dữ liệu bằng luật kết hợp, trình bày thuật toán Apriori. Trình bày khái niệm và các bước trong quá trình khai phá dữ liệu bằng cây quyết định, trình bày thuật toán C4.5

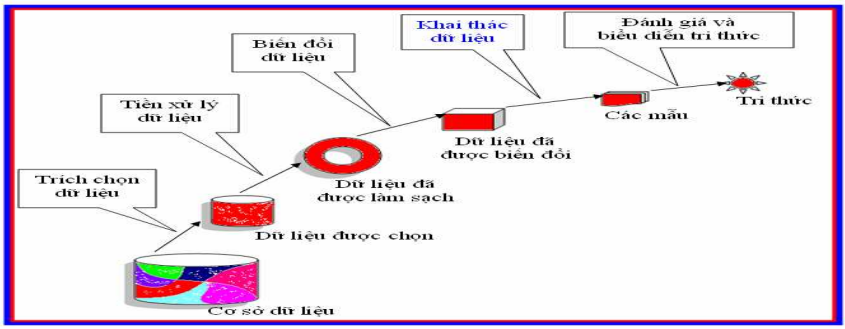
Chương 3: Xây dựng hệ thống cây quyết định trong phân tích số liệu dân cư.

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. GIỚI THIỆU CHUNG VỀ KHÁM PHÁ TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

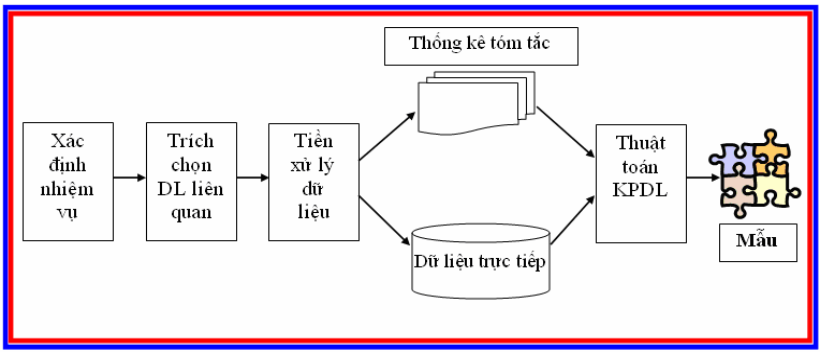
Hiện nay, lượng dữ liệu mà con người thu thập, lưu trữ trong các kho dữ liệu là rất lớn, những kỹ thuật truyền thống không đủ khả năng làm việc với dữ liệu thô. Vậy làm thế nào chúng ta có thể trích lọc được những thông tin có ích từ một kho dữ liệu rất lớn. Để giải quyết vấn đề đó, kỹ thuật khám phá tri thức trong cơ sở dữ liệu đã ra đời.

1.2. QUÁ TRÌNH KHÁM PHÁ TRI THỨC



Hình 1.1: Các bước trong quá trình khám phá tri thức.

1.3. QUÁ TRÌNH KHAI PHÁ DỮ LIỆU



Hình 1.2: Quá trình khai phá dữ liệu

1.4. CÁC PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU

1.4.1. Theo quan điểm của học máy

1.4.2. Theo các lớp bài toán cần giải quyết

1.5. CÁC DẠNG CƠ SỞ DỮ LIỆU CÓ THỂ KHAI PHÁ

- Cơ sở dữ liệu quan hệ
- Cơ sở dữ liệu đa chiều
- Cơ sở dữ liệu giao tác
- Cơ sở dữ liệu quan hệ - hướng đối tượng
- Dữ liệu không gian và thời gian
- Cơ sở dữ liệu đa phương tiện …

1.6. MỘT SỐ THÁCH THỨC TRONG KHAI PHÁ DỮ LIỆU

- Các cơ sở dữ liệu lớn
- Số chiều lớn (số thuộc tính của dữ liệu quá nhiều)
- Thay đổi dữ liệu và tri thức
- Dữ liệu bị thiếu hoặc nhiễu
- Quan hệ giữa các trường phức tạp
- Giao tiếp giữa người sử dụng với các tri thức đã có
- Tích hợp với các hệ thống khác…

1.7. KẾT LUẬN

Quá trình nghiên cứu tổng quan về khai phá dữ liệu giúp chúng ta hiểu được các bước trong qui trình khai phá dữ liệu, phương pháp, dạng dữ liệu có thể khai phá và những vấn đề cần giải quyết trong khai phá dữ liệu.

CHƯƠNG 2
KHAI PHÁ DỮ LIỆU BẰNG LUẬT KẾT HỢP VÀ PHÂN LỚP

2.1 KHAI PHÁ DỮ LIỆU BẰNG LUẬT KẾT HỢP

2.1.1. Khái niệm về tập phổ biến và luật kết hợp

Trước khi đi vào tìm hiểu kỹ thuật khai thác dữ liệu bằng luật kết hợp, ta có một số khái niệm cơ bản như sau:

Hạng mục (Item): là một thuộc tính nào đó (i_k) của đối tượng đang xét trong cơ sở dữ liệu. ($i_k : k \in \{1...m\}$, với m là số thuộc tính của đối tượng).

Tập các hạng mục (Itemset) $I = \{i_1, i_2, ..., i_m\}$: là tập hợp các thuộc tính của đối tượng đang xét trong cơ sở dữ liệu.

Giao dịch (transaction): là tập các hạng mục trong cùng một đơn vị tương tác, mỗi giao dịch được xử lý một cách nhất quán mà không phụ thuộc vào các giao dịch khác.

Cơ sở dữ liệu giao dịch D: là tập các giao dịch mà mỗi giao dịch được đánh nhãn với một định danh duy nhất (cơ sở dữ liệu giao dịch $D = \{T_1, T_2, ..., T_n\}, T_i \subseteq I$).

Một giao dịch $T \in D$ hỗ hợ một tập $X \subseteq I$ nếu nó chứa tất cả các mục của X.

Độ hỗ trợ (supp) của tập các hạng mục X trong cơ sở dữ liệu giao dịch D là tỷ lệ giữa số các giao dịch chứa X trên tổng số giao dịch trong D.

$$Supp(X) = \frac{\text{Số lượng giao dịch chứa X}}{\text{Tổng số giao dịch}} \qquad (2.1)$$

Tập các hạng mục phổ biến X hay tập phổ biến là tập các hạng mục có độ hỗ trợ thoả mãn độ hỗ trợ tối thiểu (minsupp) (minsupp là một giá trị do người dùng xác định trước).

Nếu tập mục X có $Supp\ (X) \geq$ minsupp thì ta nói X là một tập các mục phổ biến.

Tập phổ biến tối đại là tập phổ biến và không tồn tại tập nào bao nó là tập phổ biến.

Tập phổ biến đóng là tập phổ biến và không tồn tại tập nào bao nó có cùng độ hỗ trợ như nó.

Vấn đề khám phá luật kết hợp được phát biểu như sau: Cho trước 2 thông số độ hỗ trợ θ và độ tin cậy β. Đánh số tất cả các mẫu trong D có độ hỗ trợ và độ tin cậy lớn hơn hay bằng θ và β tương ứng.

Luật kết hợp cho biết phạm vi mà trong đó sự xuất hiện các mục X nào đó trong các giao dịch của cơ sở dữ liệu giao dịch D sẽ kéo theo sự xuất hiện tập những mục Y cũng trong giao dịch đó. Mỗi luật kết hợp được đặc trưng bởi hai thông số là độ hỗ trợ và độ tin cậy (supp, conf).

Luật kết hợp $X \rightarrow Y$ tồn tại một độ tin cậy confidence (c/conf). Độ tin cậy conf được định nghĩa là khả năng giao dịch T hỗ trợ X thì cũng hỗ trợ Y. Ta có công thức tính độ tin cậy conf như sau:

$$Conf\ (X \rightarrow Y) = \frac{Supp\ (X \cup Y)}{Supp\ (X)} \qquad (2.2)$$

Khai phá dữ liệu bằng luật kết hợp phân thành hai bài toán con :

Bài toán 1: Tìm tất cả các tập mục mà có độ hỗ trợ lớn hơn độ hỗ trợ tối thiểu do người dùng xác định. Các tập mục thoả mãn độ hỗ trợ tối thiểu được gọi là các tập mục phổ biến.

Bài toán 2 : Dùng các tập mục phổ biến để sinh ra các luật mong muốn. Ý tưởng chung là nếu gọi XY và X là các tập mục phổ biến, thì chúng ta có thể xác định luật nếu $X \rightarrow Y$ với tỷ lệ độ tin cậy :

$$Conf\ (X \rightarrow Y) = \frac{Supp\ (XY)}{Supp\ (X)} \qquad (2.3)$$

Nếu conf($X \rightarrow Y$) $\geq$ minconf thì luật kết hợp $X \rightarrow Y$ được giữ lại (Luật này sẽ thoả mãn độ hỗ trợ tối thiểu vì X là phổ biến).

❖ Các tính chất của tập mục phổ biến

Tính chất 1:

Với X và Y là tập các mục, nếu $X \subseteq Y$ thì : $Supp(X) \geq Supp(Y)$. Điều này là rõ ràng vì tất cả các giao dịch của D hỗ trợ Y thì cũng hỗ trợ X.

Tính chất 2 :

Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến. Nếu tập mục X không có độ hỗ trợ tối thiểu trên D nghĩa là $Supp(X) <$ minsupp thì mọi tập Y chứa tập X sẽ không phải là một tập phổ biến vì $Supp(Y) \leq Supp(X) <$ minsupp (theo tính chất 1)

Tính chất 3:

Các tập con của tập phổ biến cũng là tập phổ biến. Nếu tập mục Y là tập phổ biến trên D, nghĩa là $Supp(Y) \geq$ minsupp thì tập con X của Y là tập phổ biến trên D vì $Supp(X) \geq Supp(Y) >$ minsupp.

❖ Các tính chất của luật kết hợp

Tính chất 1:

Nếu $X \rightarrow Z$ và $Y \rightarrow Z$ thì $X \cup Y \rightarrow Z$ chưa chắc xảy ra vì chúng còn phụ thuộc vào độ hỗ trợ của mỗi trường hợp.

Tính chất 2:

Nếu $X \cup Y \rightarrow Z$ thì $X \rightarrow Z$ và $Y \rightarrow Z$ chưa chắc xảy ra vì chúng còn phụ thuộc vào độ tin cậy trong mỗi trường hợp.

Tính chất 3:

Nếu $X \rightarrow Y$ và $Y \rightarrow Z$ thì $X \rightarrow Z$ chưa chắc xảy ra vì chúng còn phụ thuộc vào độ tin cậy.

Tính chất 4:

Nếu $A \to (L - A)$ không thoả mãn độ tin cậy cực tiểu thì luật $B \to (L - B)$ cũng không thỏa mãn, với các tập thoả L, A, B và $B \subseteq A \subset L$.

2.1.2. Các ứng dụng khai thác tập phổ biến và luật kết hợp

2.1.3. Một số hướng tiếp cận trong khai thác luật kết hợp

2.1.4. Thuật toán khai phá dữ liệu bằng luật kết hợp

2.1.4.1. Qui trình khai phá dữ liệu bằng luật kết hợp

Bước 1. Tìm tất cả các tập phổ biến theo ngưỡng minsupp

Bước 2. Tạo ra các luật từ các tập phổ biến.

Đối với tập phổ biến S, tạo ra các tập con khác rỗng của S. Với mỗi tập con khác rỗng A của S: Luật $A \to (S - A)$ là luật kết hợp cần tìm nếu $Conf(A \to (S - A)) = Supp(S) / Supp(A) \geq$ minconf.

2.1.4.2. Thuật toán Apriori khai phá dữ liệu bằng luật kết hợp

Bài toán đặt ra:

- Tìm tất cả các tập mục có độ hỗ trợ minsupp cho trước.

- Sử dụng các tập mục phổ biến để sinh ra các luật kết hợp với độ tin cậy minconf cho trước.

**** Quá trình thực hiện để tìm tất cả các tập mục phổ biến với minsupp cho trước:***

Bước 1: Thực hiện nhiều lần duyệt lặp đi lặp lại, trong đó tập k-mục được sử dụng cho việc tìm tập (k + 1)-mục.

Bước 2 : Các lần duyệt sau sử dụng kết quả tìm được ở bước trước đó để sinh ra các tập mục ứng viên, kiểm tra độ phổ biến các ứng viên trên cơ sở dữ liệu và loại bỏ các ứng viên không phổ biến

Bước 3 : Thực hiện lặp để tìm L_3, …., L_k cho đến khi không tìm thấy tập mục phổ biến nào nữa.

Giải thuật **Apriori**

Các ký hiệu :

L_k : tập tất cả k-mục phổ biến (tức tập tất cả k-mục có độ hỗ trợ lớn hơn độ hỗ trợ tối thiểu). Mỗi phần tử của tập này có 2 trường : tập mục (itemset) và số mẫu tin hỗ trợ (support-count).

C_k : Tập tất cả k-mục ứng viên, mỗi phần tử trong tập này cũng có 2 trường là tập mục (itemset) và số mẫu tin hỗ trợ (support-count).

|D| : Tổng số giao dịch trên D.

Count: Biến để đếm tần suất xuất hiện của tập mục đang xét tương ứng, giá trị khởi tạo bằng 0.

Nội dung thuật toán Apriori được trình bày như sau:

Input: Tập các giao dịch D, độ hỗ trợ tối thiểu minsupp

Output: L- tập mục phổ biến trong D

Thuật toán:

L_1={ tập 1-mục phổ biến}// tìm tập phổ biến 1 hạng mục

For (lần lượt duyệt các mẫu tin từ đầu đến cuối trong tập L_k) do

Begin

C_{k+1}=apriori-gen(L_k);//sinh ra tập ứng viên (k+1) hạng mục

For (mỗi một giao dịch $T \in D$) do //duyệt csdl để tính support

Begin

C_T=subset(C_{k+1}, T); //lấy tập con của T là ứng viên trong C_{k+1}

For (mỗi một ứng viên $c \in C_T$) do

c.count++; //tăng bộ đếm tần suất 1 đơn vị

end;

$L_{k+1} = \{c \in C_{k+1} | \frac{\text{c.count}}{|D|} \geq \text{minsupp}\}$

End;

Return $\cup_k L_k$

+ Trong giai đoạn thứ nhất đếm support cho các mục và giữ lại các mục mà supp của nó lớn hơn hoặc bằng minsupp.

+ Trong các giai đoạn thứ k ($k \geq 1$), mỗi giai đoạn gồm có 2 pha:

- Trước hết tất cả các tập T_i trong tập L_k được sử dụng để sinh ra các tập ứng viên C_{k+1}, bằng cách thực hiện hàm Apriori_gen.
- Tiếp theo CSDL D sẽ được quét để tính độ hỗ trợ cho mỗi ứng viên trong C_{k+1}.

Thuật toán sinh tập ứng viên của hàm Apriori_gen với đối số L_k sẽ cho kết quả là tập hợp của tất cả các L_{k+1}.

Thuật toán hàm Apriori_gen

Input: tập mục phổ biến L_k có kích thước k-mục

Output: tập ứng viên C_{k+1}

Thuật toán:

```
Function apriori-gen(Lk: tập mục phổ biến có kích thước k)
 Begin
  For (mỗi Ti∈ Lk) do
   For (mỗi Tj ∈ Lk) do
    Begin
     If (Ti và Tj chỉ khác nhau 1 hạng mục) then
         C= Ti ∪ Tj ;// hợp Ti với Tj sinh ra ứng viên c
     If subset(c, Lk) then //kiểm tra tập con không phổ biến c trong Lk
         Remove (c)// xoá ứng viên c
     Else Ck+1 = Ck+1 ∪ {c}; // kết tập c vào Ck+1
     End;
  Return Ck+1
End;
```

2.2 KHAI PHÁ DỮ LIỆU BẰNG PHÂN LỚP DỮ LIỆU

2.2.1. Khái niệm sự phân lớp

Phân lớp dữ liệu là kỹ thuật dựa trên tập huấn luyện để phân lớp dữ liệu mới.

- Mục đích: Gán các mẫu vào các lớp với độ chính xác cao nhất để dự đoán những nhãn phân lớp cho các bộ dữ liệu mới.
- Đầu vào: Một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu.
- Đầu ra: Mô hình cây quyết định dựa trên tập huấn luyện và những nhãn phân lớp.

2.2.2. Quá trình phân lớp

2.2.3. Phân lớp bằng phương pháp quy nạp cây quyết định

2.2.3.1. Khái niệm cây quyết định

2.2.3.2. Tạo cây quyết định

Tạo cây quyết định bao gồm 2 giai đoạn: Tạo cây và tỉa cây

- *Tạo cây:* ở thời điểm bắt đầu tất cả những mẫu huấn luyện đều ở gốc, sau đó phân chia mẫu dựa trên các thuộc tính được chọn.

- *Tỉa cây:* là xác định và xóa những nhánh mà có phần tử hỗn loạn hoặc những phần tử nằm ngoài các lớp cho trước.

2.2.3.3. Sử dụng cây quyết định

Kiểm tra giá trị thuộc tính của từng nút bắt đầu từ nút gốc của cây quyết định và suy ra các luật tương ứng.

**** Thuật toán quy nạp cây quyết định:***

1. Cây được xây dựng đệ quy từ trên xuống dưới.
2. Ở thời điểm bắt đầu, tất cả những mẫu huấn luyện ở gốc.
3. Thuộc tính được phân loại theo giá trị.
4. Những mẫu huấn luyện được phân chia đệ quy dựa trên thuộc tính mà nó chọn lựa.

5. Kiểm tra những thuộc tính được chọn dựa trên nền tảng của heuristic hoặc một định lượng thống kê.

2.2.3.4. Giải thuật qui nạp cây quyết định C4.5

Ý tưởng giải thuật C4.5 như sau:

Đầu vào: Một tập hợp các mẫu huấn luyện. Mỗi mẫu huấn luyện bao gồm các thuộc tính với giá trị phân loại của nó.

Đầu ra: Cây quyết định có khả năng phân loại đúng đắn các mẫu huấn luyện và cho cả các bộ chưa gặp trong tương lai.

Giải thuật:

```
Function induce_tree (tập_mẫu_huấn_luyện, tập_thuộc_tính)
  begin
  if mọi mẫu trong tập_mẫu_huấn_luyện đều nằm trong cùng
  một lớp then
      return một nút lá được gán nhãn bởi lớp đó
  else if tập_thuộc_tính là rỗng then
        return nút lá được gán nhãn bởi tuyển của tất cả các lớp
        trong tập_mẫu_huấn_luyện
      else
      begin
        chọn một thuộc tính P, lấy nó làm gốc cho cây hiện tại;
        //(thuộc tính P có độ đo GainRatio lớn nhất )
        xóa P ra khỏi tập_thuộc_tính;
        với mỗi giá trị V của P
          begin
              tạo một nhánh của cây gán nhãn V;
              Đặt vào phân_vùng V các mẫu trong
              tập_mẫu_huấn_luyện có giá trị V tại thuộc tính P;
              Gọi induce_tree(phân_vùngV, tập_thuộc_tính)
              //gắn kết quả vào nhánh V
          end
    end
end
```

2.2.3.5. Một số vấn đề cần giải quyết trong việc phân lớp dữ liệu

**** Việc chọn thuộc tính nào để phân chia các mẫu?***

Ta có thể chọn bất kỳ thuộc tính nào làm nút của cây, điều này có khả năng xuất hiện nhiều cây quyết định khác nhau cùng biểu diễn một tập mẫu

Thuộc tính được chọn là thuộc tính cho độ đo tốt nhất, có lợi nhất cho quá trình phân lớp.

Độ đo để đánh giá chất lượng phân chia là độ đo đồng nhất.

- Information Gain
- Information Gain Ratio
- Gini Index
- X2 – số thống kê bảng ngẫu nhiên
- G – thống kê (statistic)

**** Điều kiện để dừng việc phân chia:***

1. Tất cả những mẫu huấn luyện thuộc về cùng một lớp.
2. Không còn thuộc tính còn lại nào để phân chia tiếp.
3. Không còn mẫu nào còn lại.

**** Độ lợi thông tin (Information Gain) trong cây quyết định:***

Information Gain (Gain): là đại lượng được sử dụng để lựa chọn thuộc tính có độ lợi thông tin lớn nhất để phân lớp. Độ đo Information Gain được tính dựa vào 2 độ đo info (I) và entropy (E).

Info là độ đo thông tin kỳ vọng để phân lớp một mẫu trong tập dữ liệu. Giả sử cho P, N là hai lớp và S là tập dữ liệu chứa p phần tử của lớp P và n phần tử của lớp N. Khối lượng thông tin cần để quyết định một mẫu tùy ý trong S thuộc lớp P hoặc N được định nghĩa như sau:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \qquad (2.6)$$

Entropy là khái niệm để đo tính thuần nhất của một tập huấn luyện. Giả sử rằng sử dụng thuộc tính A để phân hoạch tập hợp S thành những tập hợp $\{S_1, S_2, ... ,S_v\}$. Nếu S_i chứa những p_i mẫu của lớp P và n_i mẫu của N, entropy hay thông tin mong đợi cần để phân lớp những đối tượng trong tất cả các cây con S_i là:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) \qquad (2.7)$$

Độ lợi thông tin nhận được bởi việc phân nhánh trên thuộc tính A là:

$$Gain(A) = I(p, n) - E(A) \qquad (2.8)$$

Ta nhận thấy độ đo Gain có xu hướng chọn các thuộc tính có nhiều giá trị, tuy nhiên thuộc tính có nhiều giá trị không phải lúc nào cũng cho việc phân lớp tốt nhất, vì vậy ta cần chuẩn hóa độ đo Gain, việc chọn thuộc tính không chỉ dựa vào độ đo Gain mà còn phụ thuộc vào độ đo GainRation.

SplitInfo là độ đo thông tin trung bình của từng thuộc tính, để hạn chế xu hướng chọn thuộc tính có nhiều giá trị, thông tin trung bình của thuộc tính A được tính:

$$\text{SplitInfo(A)} = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right) \qquad (2.9)$$

Việc chọn thuộc tính để phân nhánh dựa vào độ đo GainRation

$$\text{GainRatio(A)} = \text{Gain(A)} / \text{SplitInfo(A)} \qquad (2.10)$$

Đây là công thức tính độ đo GainRatio cho thuộc tính A trên cơ sở dữ liệu D, sau đó ta chọn thuộc tính nào có độ đo GainRatio lớn nhất để phân lớp theo thuộc tính đó.

**** Vấn đề quá khớp trong phân lớp***

**** Vấn đề phân lớp cây quyết định trong cơ sở dữ liệu lớn***

2.3 KẾT LUẬN

Hai phương pháp khai phá dữ liệu bằng luật kết hợp và phân lớp mà chúng ta tìm hiểu trên đây, ở mỗi phương pháp có các thuật toán điển hình, chúng tiếp cận khai phá dữ liệu khác nhau, mỗi phương pháp có ưu và khuyết điểm riêng tùy thuộc vào dạng dữ liệu, miền dữ liệu, khối lượng dữ liệu...Như chúng ta đã phân tích ở trên, ưu điểm khai phá dữ liệu bằng phương pháp phân lớp dữ liệu đối với khối lượng dữ liệu lớn, chính vì thế mà chúng ta áp dụng thuật toán C4.5 để phân lớp dữ liệu dân cư. Thuật toán này là 1 trong số 10 thuật toán "nổi tiếng nhất – best known" trong Data Mining, được trao phần thưởng tại ICDM'06-Hong Kong.

CHƯƠNG 3
ỨNG DỤNG TRONG PHÂN TÍCH SỐ LIỆU DÂN CƯ

3.1 MÔ TẢ BÀI TOÁN

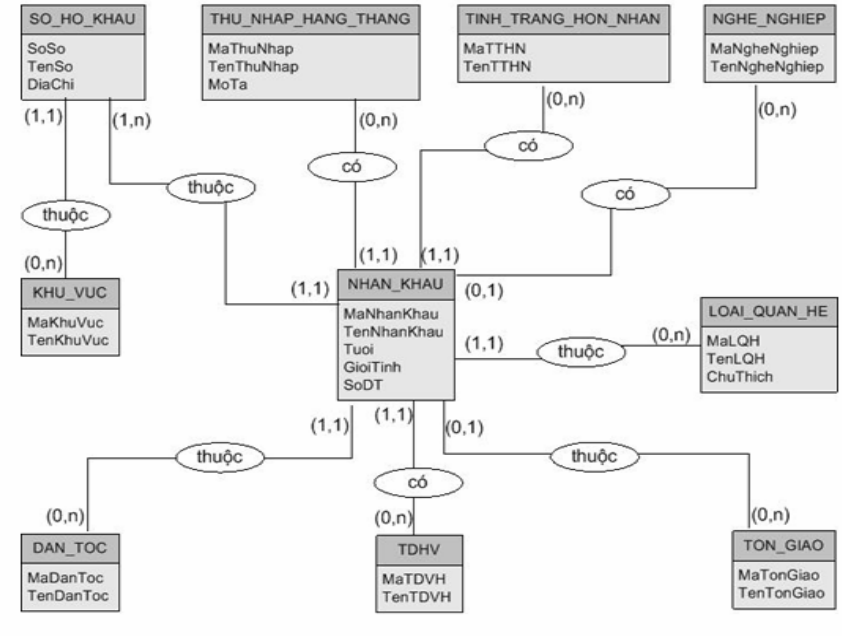
Qua khảo sát thực tế, việc thu thập dữ liệu dân cư trên toàn quốc được thực hiện theo chu kỳ 5 năm và có một số địa phương còn thực hiện việc khảo sát và cập nhật thường xuyên theo từng tháng, từng quí, từng năm nhằm thống kê dân số theo độ tuổi, giới tính, trình độ văn hóa, mức độ tăng trưởng dân số...theo từng vùng và trên cả nước. Đây là công việc cần thiết, giúp các nhà lãnh đạo có nhận định nên hỗ trợ những yếu tố nào và hạn chế những yếu tố nào, tạo điều kiện thuận lợi ổn định xã hội và phát triển đất nước.

Với mong muốn ứng dụng khai phá dữ liệu trong phân tích số liệu dân cư để tìm ra những đối tượng thường hay vi phạm kế hoạch hóa gia đình, hỗ trợ cho ban lãnh đạo DS-KHHGĐ các cấp tập trung vận động, tuyên truyền và giáo dục cho những đối tượng có thể vi phạm kế hoạch hóa gia đình góp phần thực hiện chiến lược dân số cho giai

đoạn tới đạt kết quả tốt hơn. Tác giả đã thu thập một khối lượng lớn thông tin qua các cuộc tổng điều tra dân số, thực hiện phân tích, lưu trữ dữ liệu dưới hệ quản trị CSDL quan hệ SQL Server 2005 và sử dụng thuật toán C4.5 khai phá dữ liệu bằng mô hình cây quyết định.

3.2 PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- ❖ Xác định các thực thể
- ❖ Mô hình thực thể kết hợp(ERD)



Mô hình thực thể kết hợp

- ❖ Chuyển mô hình ERD thành mô hình quan hệ

Theo phân tích dữ liệu lưu trữ và mối quan hệ của các bảng cơ sở dữ liệu đồng thời qua khảo sát thực tế, ta thấy việc có vi phạm hay không vi phạm kế hoạch hóa gia đình phụ thuộc vào nhiều thuộc tính

khác nhau. Như trình độ học vấn, khu vực sinh sống, thu nhập, giới tính của con…

❖ Xét các thuộc tính:

1. Trình độ học vấn (TH cơ sơ, TH phổ thông, THCN)
2. Khu vực sinh sống (Thành thị, Nông thôn, Miền núi)
3. Thu nhập (Thấp, Trung bình, Cao)
4. Giới tính của 2 con (1 trai 1 gái, 2 trai, 2 gái)

Từ dữ liệu lưu trữ ta rút trích các mẫu dữ liệu theo bảng sau:

Bảng3.3 Một số mẫu dữ liệu trong cơ sở dữ liệu dân cư (S)

STT	Họ và tên	Trình độ học vấn	Thu nhập	Nơi ở	Giới tính	Vi phạm
1	Hà Lương	TH phổ thông	Trung bình	Thành thị	1 trai, 1 gái	Không
2	Phạm Văn Chánh	TH cơ sở	Cao	Nông thôn	2 gái	Có
3	Nguyễn Công Trạng	TH phổ thông	Trung bình	Miền núi	1 trai, 1 gái	Không
4	Võ Bé	TH CN trở lên	Thấp	Thành thị	2 trai	Không
5	Lê Thanh Tùng	TH phổ thông	Thấp	Thành thị	2 gái	Có
6	Đỗ Ngọc Thái	TH cơ sở	Trung bình	Nông thôn	2 trai	Có
7	Nguyễn Long	TH CN trở lên	Thấp	Miền núi	2 gái	Có
8	Trương Ngọc Lộc	TH phổ thông	Cao	Thành thị	2 gái	Không
9	Nguyễn Hưu Tuân	TH cơ sở	Thấp	Miền núi	2 trai	Có
10	Lê Thanh Tùng	TH cơ sở	Cao	Miền núi	1 trai, 1 gái	Không
11	Nguyễn Minh Kế	TH phổ thông	Thấp	Nông thôn	2 trai	Không
12	Lê Văn Thắng	TH CN trở lên	Cao	Nông thôn	1 trai, 1 gái	Không
13	Huỳnh Thi Chung	TH phổ thông	Thấp	Thành thị	2 trai	Không
14	Phạm Thị Hoang	TH Phổ thông	Trung bình	Miền núi	2 gái	Có
15	Đoàn Văn Ngự	TH cơ sở	Thấp	Nông thôn	1 trai, 1 gái	Có
16	Phạm Hùng	TH CN trở lên	Cao	Miền núi	2 gái	Không
17	Võ Trung Thông	TH CN trở lên	Thấp	Thành thị	1 trai, 1 gái	Không
18	Lê Đức Sơn	TH phổ thông	Cao	Nông thôn	2 trai	Không
19	A Viết Ngai	TH cơ sở	Thấp	Miền núi	1 trai, 1 gái	Có
20	Phạm Văn Cảm	TH cơ sở	Cao	Nông thôn	1 trai, 1 gái	Không

Để xây dựng cây quyết định, tại mỗi nút của cây thì thuật toán đều đo lượng thông tin nhận được trên các thuộc tính và chọn thuộc tính có lượng thông tin tốt nhất làm nút phân tách trên cây nhằm để đạt được cây có ít nút nhưng có khả năng dự đoán cao.

Ta tính độ đo GainRatio cho các thuộc tính theo bảng dữ liệu mẫu trên để xác định thuộc tính nào được chọn trong quá trình tạo cây quyết định.

Bộ mẫu dữ liệu của chúng ta có 02 miền giá trị {c, k} (c ứng với "Có vi phạm" và k ứng với "không vi phạm kế hoạch hóa gia đình")

Lượng thông tin trên tất cả mẫu dữ liệu S theo bảng trên:

$$I(S) = -\frac{12}{20}\log_2\frac{12}{20} - \frac{8}{20}\log_2\frac{8}{20} = 0{,}970$$

+ Xét thuộc tính *trình độ học vấn*

STT	Trình độ học vấn	c_i	k_i	$I(c_i,k_i)$
1	TH cơ sở	5	2	0,86
2	TH phổ thông	2	6	0,81
3	TH chuyên nghiệp trở lên	1	4	0,72

Ta có:

$$E(\text{Trình độ học vấn}) = \frac{7}{20} * I(c_1,k_1) + \frac{8}{20} * I(c_2,k_2) + \frac{5}{20} * I(c_3,k_3) = 0{,}805$$

Trong đó:

$$I(c_1,k_1) = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0{,}86$$

$$I(c_2,k_2) = -\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8} = 0{,}81$$

$$I(c_3,k_3) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0{,}72$$

Do đó:

Gain(Trình độ học vấn) = I(S) – E(Trình độ học vấn) = 0,165

Tính độ đo SplitInfo cho thuộc tính trình độ học vấn như sau:

$$\text{SplitInfo(Trình độ học vấn)} = -\frac{7}{20}\log_2\frac{7}{20} - \frac{8}{20}\log_2\frac{8}{20} - \frac{5}{20}\log_2\frac{5}{20} = 1{,}558$$

Vậy độ đo GainRatio cho thuộc tính trình độ học vấn:

GainRatio(Trình độ học vấn)= Gain(Trình độ học vấn)/

SplitInfo(Trình độ học vấn) =0,165/1,558=**0,106**

Tương tự:

+ Tính Entropy cho thuộc tính *Thu nhập*

STT	Thu nhập	c_i	k_i	$I(c_i,k_i)$
1	Thấp	5	4	0,99
2	Trung bình	2	2	1,0
3	Cao	1	6	0,59

E(Thu nhập) = $\frac{9}{20}$ *$I(c_1,k_1)$ + $\frac{4}{20}$ *$I(c_2,k_2)$ + $\frac{7}{20}$ *$I(c_3,k_3)$= 0,852

Gain(Thu nhập) = 0,970 – 0,852 = 0,118

SplitInfo(Thu nhập)= $-\frac{9}{20}\log_2\frac{9}{20}-\frac{4}{20}\log_2\frac{4}{20}-\frac{7}{20}\log_2\frac{7}{20}=1{,}512$

GainRatio(Thu nhập)=0,118/1,512=**0,078**

+ Tính Entropy cho thuộc tính *Khu vực sinh sống*

STT	Khu vực	c_i	k_i	$I(c_i,k_i)$
1	Nông thôn	3	4	0,99
2	Miền núi	4	3	0,99
3	Thành thị	1	5	0,65

E(Khu vực) = $\frac{7}{20}$ *$I(c_1,k_1)$ + $\frac{7}{20}$ *$I(c_2,k_2)$ + $\frac{6}{20}$ *$I(c_3,k_3)$=0,888

Gain(Khu vực) = 0,970 – 0,888 = 0,082

SplitInfo(Khu vực)= $-\frac{7}{20}\log_2\frac{7}{20}-\frac{7}{20}\log_2\frac{7}{20}-\frac{6}{20}\log_2\frac{6}{20}=1{,}581$

GainRatio(Khu vực)=0,082/1,581=**0,051**

+ Tính Entropy cho thuộc tính *Giới tính của con*

STT	Khu vực	c_i	k_i	$I(c_i,k_i)$
1	2 trai	2	4	0,92
2	2 gái	4	2	0,92
3	1 trai, 1 gái	2	6	0,81

E(Giới tính) = $\frac{8}{20}$ *$I(c_1,k_1)$+ $\frac{6}{20}$ *$I(c_2,k_2)$+ $\frac{6}{20}$ *$I(c_3,k_3)$ = 0,876

Gain(Giới tính) = 0,970 – 0,876 = 0,094

SplitInfo(Giới tính)= $-\frac{8}{20}\log_2\frac{8}{20}-\frac{6}{20}\log_2\frac{6}{20}-\frac{6}{20}\log_2\frac{6}{20}=1{,}570$

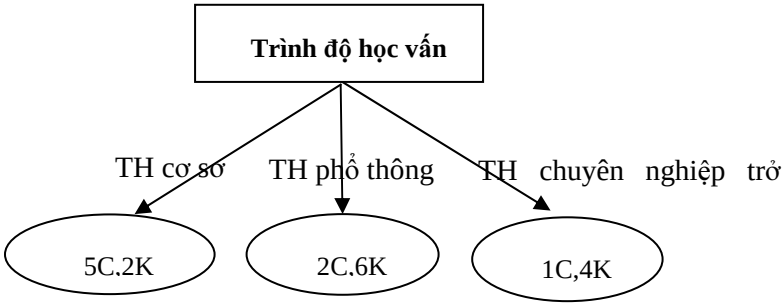
GainRatio(Giới tính của con)=0,094/1,570=**0,060**

Độ đo GainRatio của các thuộc tính được sắp xếp giảm dần:

STT	Thuộc tính	Độ đo GainRatio
1	Trình độ học vấn	0,106
2	Thu nhập	0,078
3	Giới tính của con	0,060
4	Khu vực sinh sống	0,051

Thuộc tính có độ đo GainRatio lớn nhất là "Trình độ học vấn".

Cây phân nhánh theo thuôc tính "Trình độ học vấn" như sau:



Hình 3.2. Cây quyết định tại thuộc tính "Trình độ học vấn"

Nhận xét:

Sau khi phân nhánh cây theo thuộc tính "Trình độ học vấn", ở các nút con vẫn chưa nút nào có tất cả các mẫu thuộc về một lớp. Vì vậy ta lập bảng dữ liệu phân theo giá trị tương ứng theo từng nút và tiếp tục phân nhánh cây quyết định theo từng nút này.

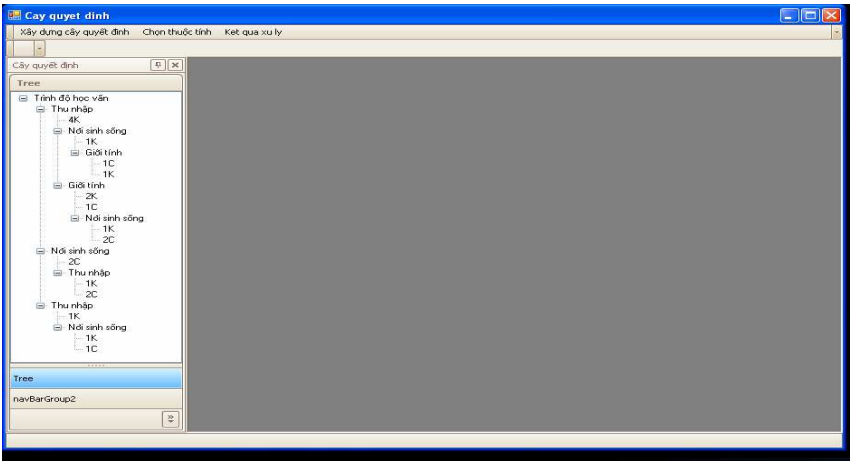
Tương tự chúng ta áp dụng thuật toán C4.5 tạo cây quyết định cho các mẫu với thuộc tính trình độ học vấn "TH cơ sở", "TH phổ thông", "TH chuyên nghiệp trở lên" và các cây con của chúng cho đến khi tất cả các nút có các mẫu cùng một lớp. Dựa vào cây quyết định cuối cùng ta rút ra các luật.

3.3 CÀI ĐẶT CHƯƠNG TRÌNH

Qua phân tích thiết kế hệ thống, CSDL dân cư được lưu trữ dưới Hệ quản trị CSDL SQL Server 2005.

Chúng ta sử dụng ngôn ngữ lập trình C# để cài đặt thuật toán cây quyết định C4.5

❖ Màn hình chính khi thực hiện chương trình:



Hình 3.10: Màn hình chính của chương trình

Các chức năng chính của chương trình:

Khi chọn menu "Xây dựng cây quyết định", chương trình thực hiện tạo cây quyết định theo tập dữ liệu mẫu.

Khi chọn menu "Chọn thuộc tính", màn hình chọn thuộc tính hiển thị, cho phép chúng ta chọn thuộc tính nào cần khai phá, khi chọn nút "thực hiện" chương trình sẽ thực hiện tạo cây quyết định theo các thuộc tính được chọn.

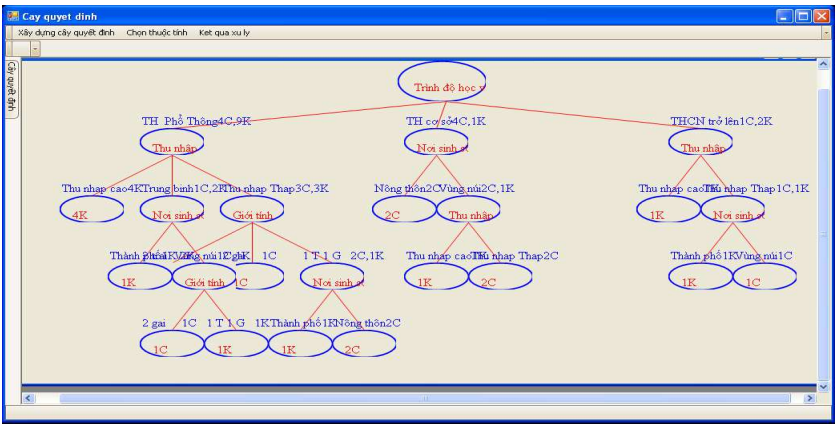
Khi nhấn chuột phải chọn nút nào trên cây quyết định thì menu popup xuất hiện gồm 2 mục: "Xây dựng cây quyết định", " Xây dựng cây thứ cấp":

- Khi chúng ta chọn mục “Xây dựng cây quyết định”: chương trình sẽ thực hiện vẽ toàn bộ cây quyết định ra màn hình.

- Khi chúng ta chọn mục “Xây dựng cây thứ cấp”: chương trình sẽ thực hiện vẽ cây theo nút được chọn ra màn hình.

Khi chọn menu “Kết quả xữ lý”, màn hình nhập liệu hiển thị cho phép chúng ta nhập vào tập dữ liệu kiểm tra. Nhấn nút “thực hiện” để nhận kết quả.

❖ Màn hình cây quyết định tổng quát:



Hình 3.11: Màn hình cây quyết định tổng quát

3.4 THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Sau khi cài đặt và thử nghiệm chương trình, ta có các nhận xét sau:

- Cài đặt thành công thuật toán C4.5, đưa ra cây quyết định theo mẫu dữ liệu đúng theo phân tích của đề tài.

- Đưa tập dữ liệu vào kiểm tra cho ra kết quả đúng theo phân tích của đề tài.

KẾT LUẬN

1. Đánh giá chung về tình hình nghiên cứu

Với khai phá dữ liệu, đây là một lĩnh vực nghiên cứu mới về việc phát hiện tri thức từ cơ sở dữ liệu lớn bằng các thuật toán đã và đang thu hút các nhà nghiên cứu và người dùng trong ngành tin học.

Luận văn đã tập trung vào việc tìm hiểu phương pháp khai phá dữ liệu bằng luật kết hợp và phân lớp, với khối lượng dữ liệu dân cư tương đối lớn nên ứng dụng khai phá dữ liệu bằng phương pháp phân lớp dữ liệu chiếm ưu thế hơn, chúng ta dùng phương pháp phân lớp dữ liệu để phân lớp và dự đoán.

2. Kết quả đạt được từ nghiên cứu

Qua đề tài này, bản thân đã tìm hiểu được một số kiến thức hữu ích nhằm tích lũy kinh nghiệm cho công việc hằng ngày. Những kết quả đạt được từ đề tài này:

- Tìm hiểu lý thuyết về khai phá dữ liệu, cụ thể là phương pháp khai phá dữ liệu bằng luật kết hợp và phân lớp tạo cây quyết định.
- Cài đặt thành công thuật toán C4.5 trên ngôn ngữ lập trình C#. Kết quả cho ra tương đối giống kết quả khảo sát thực tế.

Luận văn có ý nghĩa thực tiễn cao, gần gũi, thiết thực. Ứng dụng của đề tài không chỉ cho lĩnh vực hỗ trợ chính sách kế hoạch hóa gia đình mà còn có thể đáp ứng trong nhiều lĩnh vực khác, những lĩnh vực có tính chất dự đoán dựa vào khối lượng dữ liệu lớn.

3. Hướng phát triển của đề tài

Luận văn đã đạt được một số kết quả nhất định như trên, tuy nhiên vẫn còn có nhiều điều cần cải tiến và nghiên cứu thêm, như :

- Nâng cấp hoàn chỉnh chương trình để ứng dụng vào thực tế.
- Mở rộng lý thuyết và ứng dụng sang mô hình phân tán.