BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUẢN LÝ VÀ CÔNG NGHỆ HẢI PHÒNG

ĐỒ ÁN TỐT NGHIỆP

NGÀNH : CÔNG NGHỆ THÔNG TIN

Sinh viên : Lưu Thế Dũng

Giảng viên hướng dẫn : TS Lê Văn Phùng

HẢI PHÒNG – 2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUẢN LÝ VÀ CÔNG NGHỆ HẢI PHÒNG

XÁC ĐỊNH PHẦN TỬ NGOẠI LAI DỰA VÀO PHỤ THUỘC HÀM ĐẶC BIỆT TRONG CƠ SỞ DỮ LIỆU QUAN HỆ VÀ ỨNG DỤNG

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

NGÀNH: CÔNG NGHỆ THÔNG TIN

Sinh viên : Lưu Thế Dũng

Giảng viên hướng dẫn : TS Lê Văn Phùng

HẢI PHÒNG – 2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUẢN LÝ VÀ CÔNG NGHỆ HẢI PHÒNG

NHIỆM VỤ ĐỀ TÀI TỐT NGHIỆP

Sinh viên : Lưu Thế Dũng **Mã SV**: 1512101007

Lớp : CT2001C

Ngành : CÔNG NGHỆ THÔNG TIN

Tên đề tài: Xác định phần tử ngoại lai dựa vào phụ thuộc hàm đặc biệt trong cơ sở dữ liệu quan hệ và ứng dụng

NHIỆM VỤ ĐỀ TÀI

1. **Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp**

   - Tìm hiểu về phụ thuộc hàm và phần tử ngoại lai trong cơ sở dữ liệu quan hệ.
   - Tìm hiểu về phương pháp phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong cơ sở dữ liệu quan hệ.
   - Ứng dụng tìm phần tử ngoại lai để kiểm tra xếp loại học lực và danh hiệu cho học sinh trường THPT ở Hải Phòng.

2. **Các tài liệu, số liệu cần thiết**

   - Số liệu: Bảng điểm của lớp học trường THPT Kiến Thụy.

3. **Địa điểm thực tập tốt nghiệp**

   - Công ty Cổ Phần Thiết Bị Điện , Điện Tử - Bách Khoa.

CÁN BỘ HƯỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

Họ và tên : Lê Văn Phùng

Học hàm, học vị : Tiến sĩ

Cơ quan công tác : Viện Công nghệ Thông tin,

Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Nội dung hướng dẫn:

- Tìm hiểu về phụ thuộc hàm;
- Tìm hiểu về phát hiện phần tử ngoại lai;
- Ứng dụng phần tử ngoại lai.

Đề tài tốt nghiệp được giao ngày 18 tháng 10 năm 2021

Yêu cầu phải hoàn thành xong trước ngày 30 tháng 12 năm 2021

Đã nhận nhiệm vụ ĐTTN	Đã giao nhiệm vụ ĐTTN
Sinh viên	*Giảng viên hướng dẫn*
	TS.Lê Văn Phùng

Hải Phòng, ngày tháng năm 2021

TRƯỞNG KHOA

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

PHIẾU NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN TỐT NGHIỆP

Họ và tên giảng viên: Lê Văn Phùng

Đơn vị công tác: Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Họ và tên sinh viên : Lưu Thế Dũng

Ngành: Công nghệ Thông tin

Nội dung hướng dẫn:

- Tìm hiểu về phụ thuộc hàm và phần tử ngoại lai trong cơ sở dữ liệu quan hệ.
- Tìm hiểu về phương pháp phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong cơ sở dữ liệu quan hệ.
- Ứng dụng phần tử ngoại lai đối với phụ thuộc hàm dạng đặc biệt để kiểm tra xếp loại học lực và danh hiệu cho học sinh trường THPT ở Hải Phòng.

1.Tinh thần thái độ của sinh viên trong quá trình làm đề tài tốt nghiệp

- Học sinh có tinh thần cố gắng cao trong quá trình làm đồ án tốt nghiệp , từ sưu tập tài liệu, tìm hiểu tài liệu, tổng hợp tư liệu, phân tích số liệu thực tế tại nơi ứng dụng.
- Đảm bảo đúng tiến độ thực hiện đồ án theo quy định của nhà trường và hướng dẫn của giáo viên hướng dẫn.

2.Đánh giá chất lượng của đồ án/khóa luận (so với nội dung yêu cầu đã đề ra trong nhiệm vụ Đ.T. T.N trên các mặt lý luận, thực tiễn, tính toán số liệu…)

- Đồ án tốt nghiệp của sinh viên đã đáp ứng đầy đủ những vấn đề cốt yếu nhất của nội dung đề tài theo yêu cầu đề cương đồ án tốt nghiệp đã đặt ra.

- Phần lý thuyết đã cơ bản đáp ứng được yêu cầu tổng quan kiến thức chung và tìm hiểu sâu về kiến thức hẹp để áp dụng thực tế.
- Phần thực hành thử nghiệm lập trình tuy còn đơn giản nhưng đã thể hiện được khả năng vận dụng những kiến thức học được vào giải quyết bài toán thực tế.

3. Ý kiến của giảng viên hướng dẫn tốt nghiệp

Đạt ☐ Không đạt ☐ Điểm:……...

Hải Phòng, ngày 22 tháng 12 năm 2021

Giảng viên hướng dẫn

(Ký và ghi rõ họ tên)

TS. Lê Văn Phùng

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập - Tự do - Hạnh phúc

PHIẾU NHẬN XÉT CỦA GIẢNG VIÊN CHẤM PHẢN BIỆN

Họ và tên giảng viên: Lương Thanh Nhạn

Đơn vị công tác: Trường Đại học Y Dược Hải Phòng

Họ và tên sinh viên: Lưu Thế Dũng Ngành: Công nghệ thông tin

Đề tài tốt nghiệp: **Xác định phần tử ngoại lai dựa vào phụ thuộc hàm đặc biệt trong cơ sở dữ liệu quan hệ và ứng dụng.**

1. Phần nhận xét của giảng viên chấm phản biện

Đồ án đã thực hiện được các nội dung sau:

- Trình bày tổng quan về phụ thuộc hàm và phần tử ngoại lai trong cơ sở dữ liệu quan hệ.
- Tìm hiểu về phương pháp phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong cơ sở dữ liệu quan hệ.
- Ứng dụng tìm phần tử ngoại lai để kiểm tra xếp loại học lực và danh hiệu cho học sinh trường THPT Kiến Thụy ở Hải Phòng.
- Sử dụng ngôn ngữ lập trình PHP, hệ quản trị cơ sở dữ liệu MySQL, tác giả đã xây dựng phần mềm kiểm tra kết quả xếp loại học lực, danh hiệu học sinh của giáo viên chủ nhiệm so với qui định của Bộ giáo dục và đào tạo.

  Như vậy, sinh viên đã biết vận dụng các kiến thức đã học để giải quyết bài toán thực tế. Đồ án tốt nghiệp cơ bản đã đáp ứng đầy đủ các yêu cầu đề ra trong đề cương đã phê duyệt.

2. Những mặt còn hạn chế

- Tên đề mục 1.4: Mô hình phát hiện các phần tử ngoại lai trong dữ liệu và trong CSDL quan hệ nên sửa lại vì trong nội dung này không trình bày về mô hình mà chỉ đưa ra định nghĩa, phân loại và ứng dụng.

- Phần mềm chưa áp dụng được cho các lớp chuyên và còn tình huống chưa giải quyết triệt để(không hiển thị chú thích lỗi khi có 2 tính toán sai của một học sinh)
- Dữ liệu thực của hệ thống còn hạn chế, khóa luận chưa thực hiện đánh giá kết quả thực nghiệm.
- Lỗi chính tả(trang 17, 22, 23, ...)
- Nhiều câu viết chưa rõ nghĩa, thiếu chủ ngữ(trang 12, 48, ...)
- Định dạng toàn khóa luận chưa đồng nhất.

3. Ý kiến của giảng viên chấm phản biện

Được bảo vệ ☐ Không được bảo vệ ☐ Điểm:

Hải Phòng, ngày …… tháng ….. năm 2022

Giảng viên chấm phản biện

(Ký và ghi rõ họ tên)

TS. Lương Thanh Nhạn

LỜI CẢM ƠN

Lời đầu tiên em xin chân thành cảm ơn các Thầy, Cô trong khoa Công nghệ Thông tin cùng toàn thể các Thầy, Cô trường Đại học Quản lý và Công nghệ Hải Phòng đã giảng dạy, truyền đạt cho em những kiến thức quý báu và tạo điều kiện thuận lợi cho em trong suốt quá trình học tập tại trường cũng như trong thời gian thực hiện đồ án tốt nghiệp.

Đặc biệt, em muốn gửi lời cảm ơn tới Tiến Sỹ– Lê Văn Phùng giảng viên trực tiếp hướng dẫn tận tình chỉ bảo giúp em khắc phục những khó khăn, thiếu sót để có thể hoàn thành các phần trong đồ án tốt nghiệp từ tìm hiểu lý thuyết cho tới thực hành sử dụng công cụ.

Xin cảm ơn các bạn bè, người thân đã luôn đồng hành cùng tôi trong suốt thời gian qua và cho tôi chỗ dựa vững chắc để tôi đạt được những kết quả như ngày hôm nay.

Với hiểu biết tìm tòi của bản thân và sự chỉ bảo hướng dẫn tận tình của giảng viên, em đã cố gắng hoàn thành đồ án một cách tốt nhất có thể nhưng cũng không thể tránh được thiếu sót. Kính mong nhận được sự đóng góp ý kiến từ thầy cô để em có thể nâng cao cũng như bổ sung thêm kiến thức cho bản thân, hoàn thiện đồ án với một kết quả tốt và hoàn chỉnh hơn.

Em xin chân thành cảm ơn!

Hải Phòng, ngày tháng năm 2021.

Sinh viên thực hiện

Lưu Thế Dũng

Mục Lục

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

MỞ ĐẦU

Thế kỷ XXI được xem là một kỷ nguyên của nền kinh tế tri thức. Các công nghệ khám phá tri thức được áp dụng rộng rãi trong nhiều lĩnh vực và đã đem lại những thành tựu to lớn. Nhưng các công nghệ khám phá tri thức thường nhằm mục đích tìm kiếm, khám phá các dạng và mẫu thường gặp. Chủ yếu tập trung vào các hướng: Tìm kiếm các luật kết hợp, nhận dạng và phân lớp mẫu…Còn lĩnh vực khám phá phần tử ngoại lai mới bước đầu được sự quan tâm nghiên cứu. Mặc dù nó được ứng dụng trong nhiều lĩnh vực của cuộc sống: như phát hiện những thẻ bất thường trong hệ thống ngân hàng, những tuyến đường bất ổn không hợp lý trong giao thông, ứng dụng trong hệ thống an ninh, dự báo thời tiết, trong thị trường chứng khoán, trong lĩnh vực thể thao,...

Đồ án này thực hiện những công việc như sau:

- Tìm hiểu về phụ thuộc hàm và phần tử ngoại lai trong cơ sở dữ liệu quan hệ.
- Tìm hiểu về phương pháp phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong cơ sở dữ liệu quan hệ.
- Ứng dụng tìm phần tử ngoại lai để kiểm tra xếp loại học lực và danh hiệu cho học sinh trường THPT ở Hải Phòng.

Chương 1
Tổng quan về phụ thuộc hàm và phần tử ngoại lai trong CSDL quan hệ

1.1. Phụ thuộc hàm

1.1.1 Định nghĩa phụ thuộc hàm

Phụ thuộc hàm (functional dependency) là một công cụ dùng để biểu diễn một cách hình thức các ràng buộc. Phương pháp biểu diễn này có rất nhiều ưu điểm, và đây là một công cụ cực kỳ quan trọng, gắn chặt với lý thuyết thiết kế cơ sở dữ liệu (CSDL).

Phụ thuộc hàm được ứng dụng trong việc giải quyết các bài toán tìm khóa, tìm phủ tối thiểu và chuẩn hóa CSDL.

Khái niệm về phụ thuộc hàm trong một quan hệ là rất quan trọng trong việc thiết kế mô hình dữ liệu. Năm 1970 E.F Codd đã mô tả phụ thuộc hàm trong mô hình dữ liệu quan hệ, nhằm giải quyết việc phân rã không mất thông tin.

Định nghĩa:

Cho R = {a_1, a_2,.., a_n} là tập các thuộc tính, r = {h_1, h_2,.., h_m} là một quan hệ trên R, và A, B $\subseteq$ R (A, B là tập cột hay tập thuộc tính). Khi đó ta nói A xác định hàm cho B hay B phụ thuộc hàm vào A trong r

(ký pháp A $\xrightarrow[r]{f}$ B) nếu:

($\forall$ h_i, h_j $\in$ r) (($\forall$a $\in$ A) (h_i(a) = h_j(a)) $\Rightarrow$ ($\forall$b $\in$ B) (h_i(b) = h_j(b)))

nghĩa là đối số trùng nhau thì hàm có cùng giá trị [2].

Người ta còn viết (A, B) hay A $\rightarrow$ B thay cho $\xrightarrow[r]{f}$ B

Lúc đó tập hợp tất cả (A, B) như thế xác định một họ f trên R.

Nhận xét:

Ta có thể thấy rằng B mà phụ thuộc hàm vào A, nếu hai dòng bất kỳ mà các giá trị của tập thuộc tính A mà bằng nhau từng cặp một, thì kéo theo các giá trị trên tập thuộc tính B cũng phải bằng nhau từng cặp một.

Ý nghĩa: Khái niệm phụ thuộc hàm miêu tả một loại ràng buộc (phụ thuộc dữ liệu) xẩy ra tự nhiên nhất giữa các tập thuộc tính.

Ví dụ :

Xét một quan hệ :

THISINH

SBD	Họtên	Điạchỉ	Tỉnh	Khu vực
HP0001	Bùi văn An	14 Kiến An	Hải Phòng	3
HP0002	Nguyễn Hải Đăng	15 Cát Hải	Hải Phòng	3
HP0003	Nguyễn vân Anh	Văn Lãng	Lạng Sơn	1
HP0004	Vũ thúy Liên	52 Quang Trung	Nam Định	2

Bảng 1 Bảng quan hệ THÍ SINH

Trong quan hệ THISINH dựa vào định nghĩa phụ thuộc hàm của quan hệ , có thể kết luận:

{Tỉnh} $\xrightarrow[r]{f}$ { Khuvực}

{SBD} $\xrightarrow[r]{f}$ { Họtên, Địachỉ, Tỉnh, Khuvực}

1.1.2 Hệ tiên đề Armstrong

Gọi F là tập xác định các phụ thuộc hàm đối với lược đồ quan hệ R và X → Y là một phụ thuộc hàm. X, Y ⊆ R. Nói rằng *X → Y được suy diễn logic từ F* nếu mối quan hệ r trên R đều thoả mãn phụ thuộc hàm của F thì cũng thoả mãn X → Y. Chẳng hạn F ={A → B, B → C} thì A → C suy ra từ F. Gọi F^+ *là bao đóng* (closure) của F, tức là tập tất cả các phụ thuộc hàm được suy diễn logic từ F. Nếu F = F^+ thì F là *họ đầy đủ* (full family) của các phụ thuộc hàm [3].

Để có thể xác định khoá của một lược đồ quan hệ và các suy diễn logic giữa các phụ thuộc hàm cần thiết phải tính được F^+ từ F. Do đó đòi hỏi phải có các hệ tiên đề. Tập các quy tắc của hệ tiên đề được Armstrong (1974) đưa ra, được gọi là hệ tiên đề Armstrong.

Định nghĩa:

Cho R = {a_1,...,a_n} là tập các thuộc tính.

X, Y, Z ⊆ R. *Hệ tiên đề Armstrong* bao gồm 3 tính chất cơ bản sau:

- A1 (phản xạ) : Nếu Y ⊆ X thì X → Y
- A2 (tăng trưởng) : Nếu Z ⊆ R và X → Y thì XZ → YZ.

Trong đó ký hiệu XZ là hợp của hai tập X và Z thay cho ký hiệu $X \cup Z$.

- A3 (bắc cầu) : Nếu $X \rightarrow Y$ và $Y \rightarrow Z$ thì $X \rightarrow Z$.

Nhận xét:

- Việc nghiên cứu phụ thuộc hàm không lệ thuộc vào các quan hệ (bảng) cụ thể. Vì vậy, áp dụng được các công cụ toán nhằm sáng tỏ *cấu trúc logic của mô hình dữ liệu quan hệ*
- Có nhiều quan hệ khác nhau nhưng các họ đầy đủ các phụ thuộc hàm của chúng lại như nhau

Ví dụ:

Cho r_1, r_2 là các quan hệ:

r_1 =

A	B
0	0
1	1
2	1
3	2

r_2=

A	B
0	0
1	1
2	1
3	1

Có thể thấy rằng r_1 và r_2 khác nhau nhưng $F_{r1} = F_{r2}$ vì chỉ có $A \rightarrow B$

1.2. Các dạng phụ thuộc hàm đặc biệt loại đơn giản

Có một số phụ thuộc hàm có dạng rất đặc biệt. Chúng ta sẽ xét ở đây hai dạng rất đặc biệt trong số đó [4].

1.2.1. Dạng các phụ thuộc hàm dạng bằng nhau

Cho bảng dữ liệu r trên $R = (A_1, A_2, ...A_n)$. Giả sử với A_p, A_q nào đó thuộc R, mà với mọi $t_i \in r$ ta có: $t_i(A_p) = t_i(A_q)$. Khi đó ta dễ thấy có phụ thuộc hàm: $A_p \rightarrow A_q$ (cũng đồng thời có $A_q \rightarrow A_p$). Người ta gọi các phụ thuộc hàm dạng này là các phụ thuộc hàm dạng bằng nhau [2].

Các phụ thuộc hàm dạng bằng nhau có trong các bảng dữ liệu được sinh ra trong trường hợp chúng ta kết nối hai hoặc nhiều bảng dữ liệu với nhau.

1.2.2. Dạng phụ thuộc hàm dạng tỉ lệ

Cho r là một bảng dữ liệu trên tập thuộc tính R. Giả sử có các thuộc tính số: A_s, A_{s1}, A_{s2}, ...$A_{sk} \in R$ và các số thực: p_1, p_2, ...p_k với $p_j \leq 1$; j = 1..k

Và $\sum_{j=1}^{k} p_j = 1$; với mọi $t_i \in r$ sao cho:

$t_i(A_{s1}) = p_1 * t_i(A_s)$

$t_i(A_{s2}) = p_2 * t_i(A_s)$

.....

$t_i(A_{sk}) = p_k * t_i(A_s)$

Trong trường hợp này ta có phụ thuộc hàm:

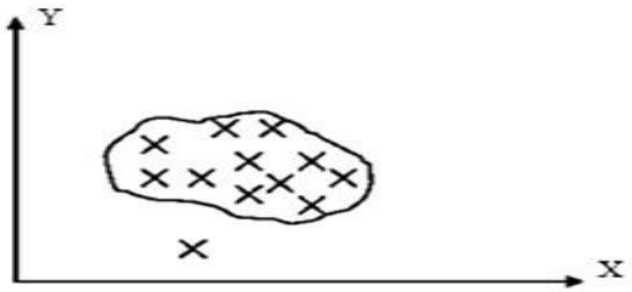
$A_s \rightarrow A_{s1}.A_{s2}$..... Ta gọi phụ thuộc hàm dạng này là phụ thuộc hàm dạng tỉ lệ. Gọi p_j là tỉ lệ đối với phụ thuộc tính A_{sj} (j = 1...k) [2].

Trong thực tế chúng ta gặp loại phụ thuộc hàm loại này trong các trường hợp tạo các bảng kê về khối lượng một loại sản phẩm được sản xuất ra cùng với các thành phần dùng để tạo nên sản phẩm đó (theo định mức qui định) [4].

1.3 Phần tử ngoại lai và mối quan hệ giữa chúng với khai phá dữ liệu

1.3.1 Khái niệm về phần tử ngoại lai

Một cách hình thức người ta có thể định nghĩa phần tử ngoại lại (outliers) của một tập dữ liệu là các phần tử mà theo một cách nhìn nào đó có các đặc tình không giống với tập hợp đa số còn lại của tập dữ liệu [1]. Chẳng hạn trong hình dưới đây cho thấy một phần tử ngoại lai theo vị trí hình học:



Hình 1 Phần tử ngoại lai trong tập điểm có tọa độ (x,y) trên mặt phẳng có giá trị tung độ y nhỏ hơn hẳn các phần tử khác của tập hợp

Các khái niệm về ngoại lai đầu tiên có nguồn gốc từ lĩnh vực thống kê. Barnett và Lewis định nghĩa: một phần tử ngoại lai là một quan trắc hoặc một tập con các quan trắc mà sự xuất hiện của chúng trái ngược với những quan trắc còn lại. Phần tử ngoại lai cũng có thể được hiểu như một quan trắc mà giá trị của nó khác biệt quá nhiều so với những quan trắc khác gây cho người người ta nghi ngờ rằng nó đã được thực hiện bằng một kỹ thuật khác.

Có nhiều cách định nghĩa và hiểu khác nhau về phần tử ngoại lai. Tuy nhiên chúng có điểm chung là: phần tử ngoại lai của một file dữ liệu là những phần tử của file dữ liệu có sự khác biệt đáng kể đối với những phần tử còn lại. Và khi tiến hành xác định phần tử ngoại lai, trước hết người ta đưa ra định nghĩa, sau đó sẽ xây dựng phương pháp để xác định.

Vai trò của phần tử ngoại lai trong mô hình CSDL quan hệ:

Cho một bảng dữ liệu *r* trên một tập thuộc tính R. Kí hiệu T là tập các qui tắc, ràng buộc (gọi là các luật) mà các phần tử của *r* phải tuân theo. Phần tử ngoại lai của *r* là những phần tử của bảng dữ liệu này không tuân theo một trong các qui tắc, ràng buộc đó [4].

Một phần tử của bảng dữ liệu được hiểu là một bộ các giá trị của các thuộc tính.

Các qui tắc, ràng buộc được đề cập bao gồm những ràng buộc về cấu của CSDL (khóa, phụ thuộc hàm, các dạng chuẩn phải tuân theo đối với một quan hệ và các ràng buộc theo ngữ nghĩa phụ thuộc vào yêu cầu, ý nghĩa của ứng dụng mà trong đó CSDL được sử dụng).

Phần tử ngoại lai giữ một vai trò đặc biệt quan trọng trong mô hình CSDL quan hệ, đặc biệt là đối với phụ thuộc hàm của bảng dữ liệu r.

Cho r là một bảng dữ liệu trên sơ đồ quan hệ (R,F). Giải thiết r là một quan hệ. Ta gọi một cặp bộ $t_1, t_2 \in r$ không thỏa mãn điều kiện phụ thuộc hàm của F là cặp phần tử ngoại lai đối với phụ thuộc hàm của bảng dữ liệu r.

Người ta biểu diễn một cách hình thức như sau:

Giả sử $X \rightarrow Y$ là một phụ thuộc hàm thuộc F. Khi đó cặp $t_1, t_2 \in r$ là cặp phần tử ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$ nếu:

$t_1(X) = t_2(X)$ và $t_1(Y) \neq t_2(Y)$.

Khái niệm phần tử ngoại lai đi cùng với mô hình CSDL quan hệ ngày nay đã ngày càng đi sâu vào mọi mặt trong đời sống kinh tế -xã hội [1]. Chúng dùng để:

- Phát hiện xâm nhập (phát hiện các hoạt động nguy hiểm (phá vỡ thâm nhập và các hình thức khác của máy tính lạm dụng) trong một hệ thống máy tính liên quan từ một vấn đề bảo mật. Khác với hệ thống hành vi bình thường, phát hiện xâm nhập là một ứng cử viên hoàn hảo cho việc áp dụng các kỹ thuật phát hiện ngoại lai).

-Phát hiện gian lận (liên quan đến hoạt động tội phạm xảy ra trong các tổ chức thương mại, các tổ chức như ngân hàng, các công ty thẻ tín dụng, cơ quan bảo hiểm, các công ty điện thoại di động, thị trường chứng khoán,… Người sử dụng độc hại có thể là khách hàng thực tế của tổ chức hoặc phải dùng đến hành vi trộm cắp danh tính (giả làm khách hàng). Các hoạt động phát hiện nhằm mục đích

phát hiện tiêu thụ trái phép các nguồn tài nguyên được cung cấp bởi tổ chức để ngăn chặn thiệt hại kinh tế).

- Phát hiện bảo hiểm yêu cầu bồi thường gian lận (ví dụ xe hơi gian lận bảo hiểm. Các cá nhân và tổ chức bên yêu sách và các nhà cung cấp thao tác yêu cầu bồi thường hệ thống xử lý cho các tuyên bố trái phép và bất hợp pháp. Các dữ liệu trong lĩnh vực này để phát hiện gian lận đến từ các văn bản trình của các bên tranh chấp).

- Phát hiện gian lận trong y tế công cộng (Dữ liệu có thể có giá trị ngoại lai do một số lý do như tình trạng bệnh nhân bất thường hoặc thiết bị đo đạc lỗi hoặc lỗi ghi âm. Hầu hết các ngoại lai hiện tại kỹ thuật phát hiện trong này nhằm mục đích phát hiện tại miền hồ sơ bất thường (ngoại lai điểm)).

- Phát hiện thiệt hại công nghiệp (đơn vị công nghiệp bị thiệt hại do liên tục sử dụng và hao mòn thông thường, thiệt hại như vậy cần phải được phát hiện sớm để ngăn chặn sự leo thang hơn nữa và gây tổn thất dẫn đến thiệt hại. Các dữ liệu trong phạm vi này thường là cảm biến dữ liệu được ghi bằng các cảm biến khác nhau và thu thập cho phân tích. Ví dụ như Phát hiện lỗi trong đơn vị cơ khí và thành phần như động cơ, tua-bin, dầu chảy trong đường ống,... Các vết nứt trong dầm, thủng trong khung máy bay, dữ liệu không lường trước được sử dụng cho lỗi phát hiện ở các đơn vị cơ khí,…).

- Phát hiện gian lận trong Xử lý hình ảnh (Phát hiện ngoại lai ở đây nhằm phát hiện những thay đổi trong một hình ảnh theo thời gian (phát hiện chuyển động) hoặc trong các khu vực mà xuất hiện bất thường trên hình ảnh tĩnh. Tên miền này bao gồm các hình ảnh vệ tinh, công nhận chữ số, quang phổ, hình ảnh X quang vú, và giám sát video,... Các yếu tố được gây ra bởi chuyển động hoặc chèn đối tượng hoặc thiết bị lỗi. Các dữ liệu có không gian cũng như đặc điểm thời gian. Mỗi điểm dữ liệu có một vài các thuộc tính liên tục như màu sắc, kết cấu,…Các

giá trị ngoại lai thú vị là những điểm hoặc bất thường hoặc khu vực trong ảnh (điểm và sự chênh lệch theo ngữ cảnh)).

- Phát hiện sai sót trong mạng cảm biến (ngoại lai trong dữ liệu thu thập hoặc có thể bao hàm một hoặc nhiều cảm biến bị lỗi (Các ứng dụng cảm biến phát hiện lỗi), hoặc các cảm biến sự kiện phát hiện (ứng dụng phát hiện xâm nhập)).

1.3.2 Mối quan hệ giữa phần tử ngoại lai với khai phá dữ liệu

Trước khi các kỹ thuật khai thác dữ liệu ra đời, thông tin hữu ích phục vụ cho người dùng chỉ được khai thác hiệu quả trên các tập dữ liệu có cỡ và số chiều nhỏ. Do vậy, để có thể khai thác dữ liệu một cách hiệu quả với khối lượng thông tin dữ liệu lớn thì cần phải có các công dụng khai thác dữ liệu tốt, các thuật toán khai thác dữ liệu thông minh, tự động, thời gian thực hiện nhanh [1].

Trong thực tế, các chương trình ứng dụng khai thác dữ liệu thường phải khai thác dữ liệu trên các tập dữ liệu rất lớn với khối lượng thông tin khổng lồ, không phù hợp với bộ nhớ chính, dữ liệu đó được nằm ở bộ nhớ ngoài (Disk-resident Data).

Trong khoá luận này vấn đề được quan tâm lớn nhất là tìm hiểu các thuật toán khai thác, tìm kiếm các phần tử ngoại lai trong các tập dữ liệu lớn, nhiều chiều. Hiện nay, một số các kỹ thuật phát hiện phần tử ngoại lai nhằm các mục đích sau:

- Cung cấp một số thông tin về mối quan hệ giữa các phần tử ngoại lai.
- Cung cấp một số giải thích hoặc mô tả về không gian dữ liệu mà trong đó xuất hiện phần tử ngoại lai.

Và một vấn đề khác chúng ta cần quan tâm đó là việc liên quan tới ý nghĩa của các phần tử ngoại lai. Cho đến nay, chưa có một định nghĩa nào có thể định

nghĩa một cách đầy đủ và chính xác về phần tử ngoại lai, việc xác định các phần tử ngoại lai trong mỗi lĩnh vực là khác nhau, bởi vì ý nghĩa ngoại lai của các phần tử ngoại lai mang tính chất và đặc trưng của từng lĩnh vực áp dụng (có thể nhiễu của người này nhưng lại là tín hiệu tốt của người khác), nên rất khó có thể đưa ra được một định nghĩa hoàn chỉnh và chính xác về phần tử ngoại lai.

1.4. Mô hình phát hiện các phần tử ngoại lai trong dữ liệu và trong CSDL quan hệ

1.4.1 Định nghĩa

Cho một sơ đồ quan hệ (R,F), với tập thuộc tính R = {A_1, A_2, ...A_n} và tập các phụ thuộc hàm F đúng trên R. Gọi F^+ là bao đóng của F (theo Hệ tiên đề Amstrong). Giả sử *r* là một bảng dữ liệu có các trường (thuộc tính) và miền giá trị trùng với quan hệ trên sơ đồ quan hệ (R,F). Ta gọi *r* là bảng dữ liệu trên R. Bảng dữ liệu này có thể chứa những bộ trùng nhau. Kí hiệu T là tập các ràng buộc và qui tắc mà các phần tử của R thỏa mãn.

Định nghĩa:

Cho một bảng dữ liệu *r* trên một tập thuộc tính R. Kí hiệu T là tập các qui tắc, ràng buộc (gọi là các luật) mà các phần tử của *r* phải tuân theo. Phần tử ngoại lai của *r* là những phần tử của bảng dữ liệu này không tuân theo một trong các qui tắc, ràng buộc đó.

Một phần tử của bảng dữ liệu được hiểu là một bộ các giá trị của các thuộc tính.

Các qui tắc, ràng buộc được đề cập bao gồm những ràng buộc về cấu của CSDL (khóa, phụ thuộc hàm, các dạng chuẩn phải tuân theo đối với một quan hệ và các ràng buộc theo ngữ nghĩa phụ thuộc vào yêu cầu, ý nghĩa của ứng dụng mà trong đó CSDL được sử dụng) [2].

1.4.2 Phân loại các phần tử ngoại lai trong CSDL quan hệ

Tùy theo các loại ràng buộc đối với các phần tử trong một quan hệ của CSDL quan hệ ta cũng có những loại phần tử ngoại lai đối với từng trường hợp đó (phần tử vi phạm các ràng buộc tương ứng). Tùy theo ngữ cảnh và yêu cầu của bài toán thực tế mà các khái niệm, định nghĩa, phương pháp xác định phần tử ngoai lai sẽ được đưa ra. Trong phạm vi tìm hiểu của khoá luận, em chỉ đề cập tới hai dạng phần tử ngoại lai khá đơn giản đối với CSDL quan hệ và có ý nghĩa trong công việc ứng dụng vào một số bài toán hỗ trợ xếp loại học lực và danh hiệu cho học sinh THPT. Các phần tử ngoại lai trong CSDL quan hệ được đề cập tới trong khoá luận chỉ bao gồm phần tử ngoại lai đối với phụ thuộc hàm [1].

1.4.3. Ứng dụng của các phần tử ngoại lai

Đối với một số ứng dụng khám phá tri thức, các sự kiện hiếm có thường được quan tâm hơn các sự kiện thông thường, chẳng hạn việc phát hiện các thẻ tín dụng giả, theo dõi các hoạt động tội phạm trong thương mại điện tử.

Sau sự tấn công các trang Web năm 2000 và đặc biệt sự kiên khủng bố tấn công nước Mỹ ngày 11/9/2001, người ta quan tâm nhiều đến việc bảo mật máy tính, bao gồm cả phần cứng, phần mềm và cả hệ thống mạng (ví dụ: phát hiện sự xâm nhập). Bảo mật hệ thống mạng bao gồm tần suất của các tấn công dịch vụ mà một sự kiện bên ngoài được phát hiện trong gói dữ liệu hệ thống mạng (ví dụ: Số lượng lớn không bình thường các gói dữ liệu từ một nguồn nặc danh). Công cụ thống kê có thể được dùng để tìm ra một thói quen là ngoại lệ tương ứng với một lịch sử đã biết (ví dụ: Những thói quen điển hình theo đăng nhập, sử dụng CPU và truy xuất dữ liệu) [1].

Đối với các hệ thống thanh toán điện tử bao gồm các ứng dụng thẻ tín dụng, thẻ điện thoại và thẻ thông minh, chúng ta quan tâm tới việc phát hiện thẻ giả.

Thêm một ứng dụng nữa trong việc phát hiện phần tử ngoại lai là ứng dụng để nghiên cứu cổ phiếu, chứng khoán. Nhiều cá nhân và công ty đã từng thử dự đoán giá trị các cổ phiếu được niêm yết dựa trên việc tìm kiếm các phần tử ngoại lai (ví dụ: Giả sử phần lớn giá các cổ phiếu ở một ngành đang lên cao ở một thị trường ảo và có các thị trường khác (trong cùng một ngành) mà giá cổ phiếu biến động đột ngột, các phần tử ngoại lai như thế nên được xác định và sau đó các nhà phân tích có thể dựa vào các nguyên nhân để giải thích sự quá nóng hoặc quá lạnh của thị trường, để xác định khuynh hướng của cố phiếu có thể mua vào hay bán ra hoặc tích lũy). Sự có mặt của các phần tử ngoại lai trong các cổ phiếu của các quỹ chung, có thể giúp làm đa dạng hóa bảng niêm yết cổ phiếu trên sàn chứng khoán trong cùng một loại.

Trên các thị trường chứng khoán thế giới, các giao dịch được thực hiện mỗi ngày lên đến con số hàng triệu giao dịch, các nhà quản lý bảng niêm yết, các doanh nhân và các nhà tư vấn đầu tư tìm mua các cổ phiếu xuống thấp và có thể có tín hiệu tốt trong tương lai để kiếm lời. Các hãng kinh doanh ở phố Wall là một trong những nhà chứng khoán sành sỏi nhất thế giới về phần cứng và phần mềm máy tính tiên tiến ứng dụng trong lĩnh vực chứng khoán như phân tích, dự báo, thống kê, (ví dụ công ty Insighful nhà sản xuất phần mềm thống kê S-PLUS bán các máy công cụ thống kê và khai thác dữ liệu siêu việt cho rất nhiều khách hàng trong lĩnh vực đầu tư).

Trong thể thao chuyên nghiệp các ông bầu ai cũng muốn xây dựng cho mình một đội hình mạnh nhất, nhưng chi phí rẻ, hợp lý với nguồn tài chính cố định và một đội hình tài năng, đa dạng được xác định bằng những sự thống kê

hiệu suất và sự trình diễn kỹ thuật của các cầu thủ, ví dụ một ông bầu có thể chấp nhận các cầu thủ chưa nổi danh (có thể mua tương đổi rẻ) hay những cầu thủ có phong cách thi đấu tốt và ấn tượng, tùy vào việc huấn luyện viên có nhu cầu từng vị trí.

Việc phát hiện các phần tử ngoại lai đối với CSDL quan hệ có nhiều ứng dụng trong thực tế: làm sạch dữ liệu (Data cleaning) trong lĩnh vực khai thác dữ liệu (Data mining), phát hiện gian lận và sai sót trong lĩnh vực kiểm toán, thương mại điện tử. Lý thuyết nghiên cứu về phần tử ngoại lai trong các CSDL quan hệ đặt cơ sở cho việc phát triển các phần mềm phát hiện tự động các phần tử có dấu hiệu khác biệt để cho các chuyên gia xác định xem cần loại bỏ nó khỏi CSDL hay cần xử lý đặc biệt với các đối tượng liên quan tới các dữ liệu này

Chương 2

Phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong CSDL quan hệ

2.1 Phần tử ngoại lai đối với phụ thuộc hàm

2.1.1 Khái niệm phần tử ngoại lai đối với phụ thuộc hàm

Cho r là một bảng dữ liệu trên sơ đồ quan hệ (R,F). Giả thiết r là một quan hệ. Người ta gọi một cặp bộ $t_1, t_2 \in r$ không thỏa mãn điều kiện phụ thuộc hàm của F là cặp phần tử ngoại lai đối với phụ thuộc hàm của bảng dữ liệu r.

Có thể biểu diễn một cách hình thức như sau [2]:

Giả sử $X \rightarrow Y$ là một phụ thuộc hàm thuộc F. Khi đó cặp $t_1, t_2 \in r$ là cặp phần tử ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$ nếu $t_1(X) = t_2(X)$ và $t_1(Y) \neq t_2(Y)$.

2.1.2 Định lý nhận biết cặp ngoại lai đối với phụ thuộc hàm

Định nghĩa Hệ bằng nhau:

Giả sử $r = \{t_1, t_2, ...t_m\}$ là bảng dữ liệu trên cơ đồ quan hệ (R,F). Tập E_r được xác định như sau [2]:

$E_r = \{E_{i,j}: 1 \leq i < j \leq m$ và $E_{i,j} = \{a \in R; t_i(a) = t_j(a)\}\}$.

Gọi E_r là hệ bằng nhau của r.

Định lý nhận biết cặp ngoại lai:

Cho r là một bảng dữ liệu trên sơ đồ quan hệ (R,F); E_r là hệ bằng nhau của r; $X \rightarrow Y$ là một phụ thuộc hàm được giả thiết đúng trên r. Cặp phần tử (t_i,t_j) với $t_i, t_j \in r$ là ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$ khi và chỉ khi $E_{i,j} \in E_r$ mà $X \subseteq E_{i,j}$ nhưng $Y \not\subset E_{i,j}$.

Chứng minh:

Thật vậy giả sử (t_i, t_j) là cặp ngoại lai đối với phụ thuộc hàm X $\rightarrow$ Y, khi đó ta có: $t_1(X) = t_2(X)$ nhưng $t_1(Y) \neq t_2(Y)$. Từ định nghĩa $E_{i,j}$ ta có X $\subseteq$ $E_{i,j}$ nhưng Y $\not\subset$ $E_{i,j}$, vì nếu Y $\subseteq$ $E_{i,j}$ thì $t_i(Y) = t_j(Y)$ trái với giải thiết.

Ngược lại: Nếu có $E_{i,j} \in E_r$ (xác định theo t_i, t_j) mà X $\subseteq E_{i,j}$ nhưng Y $\not\subset$ $E_{i,j}$ thì cũng theo cách xác định $E_{i,j}$ ta có: $t_i(a) = t_j(a)$ với a $\in E_{i,j}$ do (Y $\not\subset$ $E_{i,j}$), do vậy $t_i(X) = t_j(X)$. Cũng do Y $\not\subset$ $E_{i,j}$ nên $t_i(Y) \neq t_j(Y)$. Theo định nghĩa 2.1.1 thì (t_i, t_j) là cặp ngoại lai. Điều phải chứng minh.

Ví dụ:

Cho bảng quan hệ r sau

A	B	C	D	E
0	1	2	1	3
1	2	2	1	3
0	1	2	1	4
3	1	2	1	1
3	2	1	3	1

Bảng 2

Hàng trên cùng biểu diễn các thuộc tính A, B, C, D, E. Các hàng tiếp theo được coi là các phần tử của quan hệ và được đánh số lần lượt 1, 2, 3, 4, 5, 6.

Xét phụ thuộc hàm: A $\rightarrow$ BE. Tìm các cặp ngoại lai đối với phụ thuộc hàm này.

Tính E_r: $E_{1,2}$ = CDE; $E_{1,3}$ = ABCD; $E_{1,4}$ = BCD; $E_{1,5} = \varnothing$; $E_{2,3}$ = CD; $E_{2,4}$ = CD; $E_{2,5}$ = B; $E_{3,4}$ = BCD; $E_{4,5}$ = AE.

Ta thấy trong $E_{4,5}$ có chứa A, nhưng không chứa BE. Như vậy cặp (t_4, t_5) là cặp ngoại lai đối với phụ thuộc hàm $A \rightarrow BE$. Tương tự với $E_{1,3}$.

Từ định lý trên ta có thuật toán xác định các cặp ngoại lai đối với phụ thuộc hàm như ở dưới đây [2].

Thuật toán xác định các cặp ngoại lai đối với tập các phụ thuộc hàm:

Input**:** Tập thuộc tính $R = \{A_1, A_2 ....A_n\}$, bảng dữ liệu $r = \{t_1, t_2,...t_m\}$ trên R,

Tập các phụ thuộc hàm $F = \{X_1 \rightarrow Y_1, X_2 \rightarrow Y_2, X_m \rightarrow Y_s\}$

Output: OUTLI – tập các cặp ngoại lai đối với phụ thuộc hàm

Bước 1: Tính hệ bằng nhau $E_r = \{E_{i,j}: 1 \leq i < j \leq m, E_{i,j} = \{ a \in R; t_i(a) = t_j(a)\}\}$.

Bước 2: Với mỗi phụ thuộc hàm $X_i \rightarrow Y_i \in F$ và mọi $E_{i,j} \in E_r$, kiểm tra điều kiện $X_i \subseteq E_{k,j}$ và $Y_i \not\subset E_{k,j}$. Nếu đúng, lưu cặp (t_k, t_j) vào tập OUTLI. Nếu không kiểm tra tiếp các phụ thuộc hàm khác.

Tập OUTLI là tập các cặp ngoại lai đối với phụ thuộc hàm của r.

2.2 Phần tử ngoại lai đối với một số dạng phụ thuộc hàm đặc biệt

Chúng ta thấy rằng trong trường hợp đối với một phụ thuộc hàm nói chung thì các thuật toán nếu trên chỉ có thể tìm được các cặp phần tử mà trong đó có ít nhất một phần tử là ngoại lại đối với phụ thuộc hàm. Trong một số trường hợp đặc biệt của phụ thuộc hàm trong các CSDL thực tế như phụ thuộc hàm dạng bằng nhau, phụ thuộc hàm dạng tỉ lệ,... chúng ta có thể có thuật toán riêng để xác định chính xác phần tử ngoại lai đối với các phụ thuộc hàm loại này. Phần dưới đây khoá luận sẽ trình bày một số trường hợp đó.

2.2.1. Phần tử ngoại lai đối với phụ thuộc hàm dạng bằng nhau

Chúng ta đã biết khái niệm về phụ thuộc hàm bằng nhau:

Cho bảng dữ liệu r trên R = (A_1, A_2, ...A_n). Giả sử với A_p, A_q nào đó thuộc R, mà với mọi $t_i \in r$ ta có: $t_i(A_p) = t_i(A_q)$. Khi đó nếu có phụ thuộc hàm $A_p \rightarrow A_q$ mà đồng thời cũng có $A_q \rightarrow A_p$ thì đó là phụ thuộc hàm có dạng bằng nhau.

Trong trường hợp này, để xác định các cặp phần tử ngoại lai t_i, t_j ta có thể so sánh: $t_i(A_p)$ với $t_i(A_q)$ (hoặc $t_j(A_p)$ với $t_j(A_q)$). Nếu $t_i(A_p) \neq t_i(A_p)$ (hoặc $t_j(A_p) \neq t_j(A_p)$) thì khi đó mọi cặp (t_i, t_k) (hoặc (t_j, t_k)) với $t_k \in r$ đều là cặp phần tử ngoại lai [4]. Trong trường hợp này ta có thể xác định chính xác phần tử ngoại lai:

Định nghĩa:

Cho r là một bảng dữ liệu trên sơ đồ quan hệ (R,F) với R = (A_1, A_2, ...A_n); cho $A_p \rightarrow A_q$ là phụ thuộc hàm dạng bằng nhau đúng trên r ($t_i(A_p) = t_i(A_q)$ với mọi $t_i \in r$). Phần tử ngoại lai đối với $A_p \rightarrow A_q$ là phần tử $t_k \in r$ mà $t_k(A_p) \neq t_k(A_q)$.

Dựa trên định nghĩa này, người ta đưa ra thuật toán để phát hiện những phần tử ngoại lai đối với phụ thuộc hàm dạng này. Thuật toán dựa trên việc kiểm tra giá trị thuộc tính của vế trái và vế phải phụ thuộc hàm.

Thuật toán:

Input: r là bảng dữ liệu trên R; $A_p \rightarrow A_q$ là phụ thuộc hàm dạng bằng nhau:

Output: Tập các phần tử ngoại lai của r: OTL

Begin

Với mỗi $t_i \in r$ thực hiện kiểm tra: $t_i(A_p) \neq t_i(A_q)$, nếu đúng lưu t_i vào tập OTL.

End.

Ví dụ [4]:

Cho bảng kê các hợp đồng nhập khẩu hàng hóa của một công ty (đã được kết nối với bảng thuế suất qui định theo MA_HANG)

Trong đó cột TL_THUE là thuế suất được công ty kê khai, TL_THUE_QĐ là thuế suất do Nhà nước qui định theo từng loại hàng hóa.

NGAY	SO HD	MA HANG	TL_THUE	THANH TIEN	TIEN THUE	TL_ THUE_QĐ
16/4/2015	HD80000	M1001	*35%*	305,415	106,895	*35%*
17/4/2015	HD80100	M1011	*50%*	11,360	5,680	*50%*
17/4/2015	HD80100	M1020	*20%*	76,000	15,200	*35%*
17/4/2015	HD80100	M1014	*50%*	683,090	341,545	*50%*
17/4/2015	HD80100	M1001	*35%*	162,888	57,011	*35%*
28/4/2015	HD80500	M1002	*10%*	17,600	1,760	*25%*
28/4/2015	HD80500	M1005	*20%*	90,000	18,000	*40%*
28/4/2015	HD80700	M1006	*15%*	271,200	40,680	*100%*
28/4/2015	HD80800	M1007	*40%*	50,000	20,000	*55%*
29/4/2015	HD80900	M1001	*35%*	203,610	71,264	*35%*

Bảng 3 Bảng kê các hợp đồng nhập khẩu hàng hóa của một công ty

Ta thấy có phụ thuộc hàm TL_THUE $\rightarrow$ THUE_QĐ có dạng bằng nhau. Áp dụng thuật toán trên chúng ta sẽ thấy các hóa đơn ứng với loại hàng có mã số: M1020, M1002, M1005, M1006, M1007 có sự kê khai thuế suất thấp hơn thuế suất qui định.

2.2.2 Phần tử ngoại lai đối với phụ thuộc hàm dạng tỉ lệ

Định nghĩa:

Cho r là một bảng dữ liệu trên sơ đồ quan hệ (R,F). Giả sử có các thuộc tính số: A_s, A_{s1}, A_{s2}, ...$A_{sk} \in R$ và các số thực: p_1, p_2, ...p_k với $p_j \leq 1$; j = 1..k

Và $\sum_{j=1}^{k} p_j = 1$; với mọi $t_i \in r$ sao cho:

$t_i(A_{s1}) = p_1 * t_i(A_s)$

$t_i(A_{s2}) = p_2 * t_i(A_s)$

.....

$t_i(A_{sk}) = p_k * t_i(A_s)$

Trong trường hợp này ta có phụ thuộc hàm:

$A_s \rightarrow A_{s1}.A_{s2}$..... Ta gọi phụ thuộc hàm dạng này là phụ thuộc hàm dạng tỉ lệ. Gọi p_j là tỉ lệ đối với phụ thuộc tính A_{sj} (j = 1...k) [2].

Trong thực tế chúng ta gặp loại phụ thuộc hàm loại này trong các trường hợp tạo các bảng kê về khối lượng một loại sản phẩm được sản xuất ra cùng với các thành phần dùng để tạo nên sản phẩm đó (theo định mức qui định). Đồng thời trong thực tế các tỉ lệ này có thể được chấp nhận với một giá trị sai số nào đó vì vậy đối với trường hợp này ta đưa ra định nghĩa về phần tử ngoại lai như sau.

Định nghĩa:

Cho r là một bảng dữ liệu trên R. Với một số δ cho trước. Phần tử $t_i \in r$ sẽ là ngoại lai đối với phụ thuộc hàm dạng tỉ lệ $A_s \rightarrow A_{s1}.A_{s2}.....A_{sk}$, nếu tồn tại một $A_{sj} \in \{A_{s1}, A_{s2},.....A_{sk}\}$ để sao cho:

$$\left| \frac{t_i(A_{sj})}{t_i(A_s)} - p_j \right| > \delta$$

Trong thực tế đối với từng thuộc tính người ta có thể chọn giá trị δ khác nhau tùy theo yêu cầu độ chính xác.

Thuật toán phát hiện phần tử ngoại lai đối với phụ thuộc hàm dạng tỉ lệ [4]:

- Thuật toán:

Input: r là bảng dữ liệu trên R; $A_s \rightarrow A_{s1}.A_{s2}.....A_{sk}$ là phụ thuộc hàm dạng tỉ lệ. Cho các dạng thuộc tính $\{A_{s1}, A_{s2},.....A_{sk}\}$; các tỉ lệ $\{p_1, p_2....p_k\}$; δ là sai số cho phép.

Output: OTL tập các phần tử ngoại lai của r đối với phụ thuộc hàm này.

Begin

Đối với mỗi $t_i \in r$ và đối với mỗi $A_{sj} \in \{A_{s1}, A_{s2},.....A_{sk}\}$, kiểm tra điều kiện:

$$\left| \frac{t_i(A_{sj})}{t_i(A_s)} - p_j \right| > \delta$$

Nếu đúng lưu t_i vào OTL.

End.

Ví dụ:

Cho bảng dữ liệu

AS	A1	A2	A3	A4
120	24	42	30	24
75	15	26.25	18.75	15
120	24	36	30	30
80	16	28	20	16
90	28	22.5	22.5	27
80	16	28	20	16

Bảng 4 Bảng dữ iệu

Với tỉ lệ của các thuộc tính A1, A2, A3, A4 so với AS như sau:

Tỉ lệ	A1	A2	A3	A4
	0.2	0.35	0.25	0.2

Áp dụng thuật toán trên ta có bảng tính tỉ lệ thực tế như sau:

A1	A2	A3	A4
0.2	0.35	0.25	0.2
0.2	0.35	0.25	0.2
0.2	0.3	0.25	0.25
0.2	0.35	0.25	0.2
0.2	0.25	0.25	0.3
0.2	0.35	0.25	0.2

Bảng 5 Bảng tính tỉ lệ thực tế

Nếu chọn δ_1 = 0.01 thì chúng ta sẽ có các bộ tương ứng với các hàng 3, hàng 5 là ngoại lai (có tỉ lệ giữa giá trị A_3 với A_s và giữa A_5 với A_s sai khác với tỉ lệ qui định vượt quá 1%) [1].

Chương 3
Ứng dụng tìm phần tử ngoại lai để kiểm tra xếp loại học lực và danh hiệu cho học sinh trường THPT Kiến Thụy ở Hải Phòng

3.1 Bài toán đặt ra và mục tiêu chương trình

3.1.1 Bài toán đặt ra

Khi kết thúc học kỳ hay cả năm, giáo viên chủ nhiệm phải nộp lại file chứa xếp loại học lực, hạnh kiểm và danh hiệu của học sinh cho ban giám hiệu nhà trường để quản lý và theo dõi làm căn cứ báo cáo Sở giáo dục và đào tạo. Trong quá trình kết xuất file để nộp giáo viên chủ nhiệm không để ý đến học lực và danh hiệu của học sinh hay học sinh đã nghỉ học mà một số môn vẫn có điểm để tính xếp loại học lực nhưng có một số môn không tính xếp loại, hay lúc nộp file bị tác động của các yếu tố chủ quan hay khách quan làm ảnh hưởng đến kết quả xếp loại và danh hiệu.

3.1.2. Mục tiêu chương trình

Bài toán: Kiểm tra kết quả xếp loại học lực, danh hiệu do giáo viên chủ nhiệm cung cấp có đúng với qui định của Bộ giáo dục và đào tạo, cụ thể:

+ Đối với xếp loại:

1. Loại giỏi, nếu có đủ các tiêu chuẩn sau đây:

a) Điểm trung bình các môn học từ 8,0 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 8,0 trở lên; riêng đối với học sinh lớp chuyên của

trường THPT chuyên phải thêm điều kiện điểm trung bình môn chuyên từ 8,0 trở lên;

b) Không có môn học nào điểm trung bình dưới 6,5;

c) Các môn học đánh giá bằng nhận xét đạt loại Đ.

2. Loại khá, nếu có đủ các tiêu chuẩn sau đây:

a) Điểm trung bình các môn học từ 6,5 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 6,5 trở lên; riêng đối với học sinh lớp chuyên của trường THPT chuyên phải thêm điều kiện điểm trung bình môn chuyên từ 6,5 trở lên;

b) Không có môn học nào điểm trung bình dưới 5,0;

c) Các môn học đánh giá bằng nhận xét đạt loại Đ.

3. Loại trung bình, nếu có đủ các tiêu chuẩn sau đây:

a) Điểm trung bình các môn học từ 5,0 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 5,0 trở lên; riêng đối với học sinh lớp chuyên của trường THPT chuyên phải thêm điều kiện điểm trung bình môn chuyên từ 5,0 trở lên;

b) Không có môn học nào điểm trung bình dưới 3,5;

c) Các môn học đánh giá bằng nhận xét đạt loại Đ.

4. Loại yếu: Điểm trung bình các môn học từ 3,5 trở lên, không có môn học nào điểm trung bình dưới 2,0.

5. Loại kém: Các trường hợp còn lại.

6. Nếu $ĐTB_{hk}$ hoặc $ĐTB_{cn}$ đạt mức của từng loại quy định tại các Khoản 1, 2 nhưng do kết quả của một môn học nào đó thấp hơn mức quy định cho loại đó nên học lực bị xếp thấp xuống thì được điều chỉnh như sau:

a) Nếu $ĐTB_{hk}$ hoặc $ĐTB_{cn}$ đạt mức loại G nhưng do kết quả của một môn học nào đó mà phải xuống loại Tb thì được điều chỉnh xếp loại K.

b) Nếu $ĐTB_{hk}$ hoặc $ĐTB_{cn}$ đạt mức loại G nhưng do kết quả của một môn học nào đó mà phải xuống loại Y thì được điều chỉnh xếp loại Tb.

c) Nếu $ĐTB_{hk}$ hoặc $ĐTB_{cn}$ đạt mức loại K nhưng do kết quả của một môn học nào đó mà phải xuống loại Y thì được điều chỉnh xếp loại Tb.

d) Nếu $ĐTB_{hk}$ hoặc $ĐTB_{cn}$ đạt mức loại K nhưng do kết quả của một môn học nào đó mà phải xuống loại Kém thì được điều chỉnh xếp loại Y.

+ Đối với danh hiệu:

1. Công nhận đạt danh hiệu học sinh giỏi học kỳ hoặc cả năm học, nếu đạt hạnh kiểm loại tốt và học lực loại giỏi.

2. Công nhận đạt danh hiệu học sinh tiên tiến học kỳ hoặc cả năm học, nếu đạt hạnh kiểm từ loại khá trở lên và học lực từ loại khá trở lên.

3.2 Chọn thuật toán thử nghiệm

Với bài toán nêu trên, chúng ta hình dung cần phải tính lại các tiêu chí như xếp loại học lực, danh hiệu có đúng như báo cáo của giáo viên chủ nhiệm không. Vậy, chúng ta cần tính lại điểm trung bình và cho hiện ra hai cột giá trị tính được bên cạnh 2 cột mà giáo viên chủ nhiệm đã báo cáo. Để tránh nhầm lẫn, chúng ta gọi các cột "xếp loại" và "danh hiệu" tương ứng trên báo cáo của giáo viên chủ nhiệm có thêm chữ báo cáo hoặc ngầm hiểu là báo cáo. Còn cột "xếp loại" và

“danh hiệu” do máy tính tính được gọi là “xếp loại TT (tính toán)” hay “danh hiệu TT”.

Trên cơ sở đó, tiến hành so sánh các giá trị theo hàng, nếu giá trị 2 cột đó không bằng nhau thì hàng đó chính là “phần tử ngoại lai “. Như vậy, thực chất dẫn tới bài toán thử nghiệm sử dụng thuật toán phát hiện phần tử ngoại lai đối với phụ thuộc hàm dạng bằng nhau.

STT	Tên học sinh	Ngày sinh	Toán	…	Điểm trung bình	Xếp loại học lực báo cáo	Xếp loại học lực tính toán	Hạnh kiểm	Danh hiệu báo cáo (BC)	Danh hiệu tính toán (TT)
						p	q			

Bài toán đòi hỏi: $p = q, p \rightarrow q$

Từ đây hình thành bài toán thử nghiệm gồm 2 công đoạn:

1 - Sử dụng thuật toán phát hiện phần tử ngoại lai đối với phụ thuộc hàm dạng bằng nhau đối với 2 cột “xếp loại học lực tính toán” và “xếp loại học lực báo cáo”.

STT	Tên học sinh	Ngày sinh	Toán	…	Điểm trung bình	Xếp loại học lực	Xếp loại học lực

						báo cáo	tính toán

2 - Sử dụng thuật toán phát hiện phần tử ngoại lai đối với phụ thuộc hàm dạng bằng nhau đối với 2 cột xác định "danh hiệu tính toán" và "danh hiệu báo cáo".

STT	Tên học sinh	Ngày sinh	Toán	…	Điểm trung bình	Xếp loại học lực báo cáo	Xếp loại học lực tính toán	Hạnh kiểm	Danh hiệu báo cáo (BC)	Danh hiệu tính toán (TT)

Thuật toán dựa trên việc kiểm tra giá trị thuộc tính của vế trái và vế phải phụ thuộc hàm.

Thuật toán toán phát hiện phần tử ngoại lai đối với phụ thuộc hàm dạng bằng nhau [2]:

Input: r là bảng dữ liệu trên R; $A_p \rightarrow A_q$ là phụ thuộc hàm dạng bằng nhau:

Output: Tập các phần tử ngoại lai của r: OTL

Begin

Với mỗi $t_i \in r$ thực hiện kiểm tra: $t_i(A_p) \neq t_i(A_q)$, nếu đúng lưu t_i vào tập OTL.

End.

+ Áp dụng vào thực tế vào bài toán ta có thuật toán phát hiện phần tử ngoại lai xếp loại kết quả học tập như sau:

Bước 1: Nhập dữ liệu; N;

Bước 2: i ← 1 ; loi ← 0; Học lực_TT_i ← "Kém";

Bước 3:

If (ĐiểmTB$_i$>=7.95) and (Văn$_i$>=8 or Toán$_i$>=8) and (Thể dục$_i$ ="Đ") and (MIN(các môn$_i$)>=6.5)) then Học lực_TT_i ← "G"

Else

If (ĐiểmTB$_i$>=6.45) and (Văn$_i$>=6.5 or Toán$_i$>=6.5) and (Thể dục$_i$ ="Đ") and (MIN(các môn$_i$)>=5.0)) then Học lực_TT_i ← "K"

Else

If (ĐiểmTB$_i$>=4.95) and (Văn$_i$>=5.0 or Toán$_i$>=5.0) and (MIN(các môn$_i$)>=3.5)) then Học lực_TT_i ← "TB"

Else

If (ĐiểmTB$_i$>=3.45) and (MIN(các môn$_i$)>=2.0)) then Học lực_TT_i ← "Y";

Bước 4: Nếu Học lực_TT_i <> Học lực_BC_i thì loi ← loi + 1;

Bước 5: i ← i + 1;

Bước 6: Nếu i > N thì thông báo số lỗi rồi kết thúc.

Bước 7: Quay về bước 3;

+ Áp dụng vào thực tế vào bài toán ta có thuật toán phát hiện phần tử ngoại lai xếp loại danh hiệu sau:

Bước 1: Nhập dữ liệu; N;

Bước 2: i ← 1 ; loi ← 0; Danh hiệu_TT_i ← " ";

Bước 3:

If (Học lực$_k$ ="G") and (Hạnh kiểm $_k$ ="T") then

Danh hiệu_TT_i ← "HSG"

Else

If (Học lực$_k$ ="G") and (Hạnh kiểm $_k$ ="K") then

Danh hiệu_TT_i ← "HSTT"

Else

If (Học lực$_k$ ="K") and (Hạnh kiểm $_k$ ="T") then

Danh hiệu_TT_i ← "HSTT"

Else

If (Học lực$_k$ ="K") and (Hạnh kiểm $_k$ ="K") then

Danh hiệu_TT_i ← "HSTT";

Bước 5: Nếu Danh hiệu_TT_i <> Danh hiệu_BC_i thì loi ← loi + 1;

Bước 6: i ← i + 1;

Bước 7: Nếu i > N thì thông báo số lỗi rồi kết thúc.

Bước 8: Quay về bước 3;

3.3 Dữ liệu vào và yêu cầu kết quả ra

+ Dữ liệu vào: File Excel chứa dữ liệu điểm trung bình các môn, kết quả xếp loại học lực (báo cáo) và danh hiệu (báo cáo) của học sinh một lớp trong trường THPT.

+ Kết quả ra: Thể hiện số lượng và đánh dấu các phần tử ngoại lai dưới dạng lỗi, bao gồm 2 dạng thử nghiệm:

a-không tìm thấy phần tử ngoại lai;

b-có tìm thấy phần tử ngoại lai.

3.4 Môi trường thử nghiệm và quy trình thực hiện

3.4.1 Hệ quản trị dữ liệu

Hệ quản trị cơ sở dữ liệu MySQL.

3.4.2 Ngôn ngữ lập trình

Ngôn ngữ lập trình Php.

3.4.3. Quy trình thực hiện

Quy trình thực hiện bài toán có thể tiến hành theo các bước sau:

-Bước 1: Nhập tệp dữ liệu (tệp Input1) vào máy từ tệp trong Excel do giáo viên chủ nhiệm gửi tới.

Tệp Input1 (tệp dữ liệu đầu vào, được lưu giữ trong excel):

STT	Họ tên	Ngày sinh	Toán	Vật lý	Hoá học	Sinh học	Tin học	Ngữ văn	Lịch sử	Địa lý	Tiếng anh	GDCD	Công nghệ	Thể dục	GDQP-A
1	Nguyễn Trường Giang	1/10/2000	9.0	7.0	9.0	5.7	9.0	8	8.5	8.4	9	8.2	9.1	Đ	7.6
2	Nguyễn Chí Nguyên	1/11/2000	6.0	7.1	7.2	7.5	9.0	8	8.5	5.6	6.2	8.6	8.6	Đ	7.8
3	Trần Văn Tân	1/12/2000	6.6	7.0	7.0	6.1	9.0	8	8.5	6.4	5.5	8.4	8.7	Đ	7.8
4	Nguyễn Văn Minh	1/13/2000	8.0	6.8	6.0	6.1	9.0	8	8.5	7.1	5.6	7.6	8.5	Đ	8.1
5	Nguyễn Tuấn Minh	1/14/2000	8.0	5.5	6.9	6.6	9.0	8	8.5	7.5	6.3	7.5	8.5	Đ	8.2
6	Nguyễn Văn Ninh	1/15/2000	9.0	7.1	7.1	7.0	9.0	8	8.2	9.1	9.5	6.6	7.6	Đ	8.4
7	Phạm Thị Hoa	1/16/2000	7.0	6.5	7.2	7.7	9.0	8	8.5	8.2	7.1	6.9	8.4	Đ	7.5
8	Ninh Văn Khánh	1/17/2000	8.0	6.0	6.1	5.4	9.0	8	6.9	7.4	8.1	8.1	8.6	Đ	8.1
9	Đỗ Thị Minh	1/18/2000	9.0	6.6	5.0	5.0	9.0	8	8.3	5.4	8.7	7.5	7.4	Đ	8.1
10	Đào Thị Quỳnh	1/19/2000	8.0	7.0	6.4	7.0	9.0	8	8.7	8.0	5.9	8.6	6.9	Đ	8.1
11	Vũ Thị Hải Yến	1/20/2000	8.0	7.9	6.4	6.8	9.0	8	7.9	6.0	8.3	8	6	Đ	7.9
12	Văn Đức Vũ	1/21/2000	9.0	6.3	7.6	6.2	9.0	8	8.2	8.1	6.4	8.8	8.5	Đ	7.8
13	Đỗ Minh Tiến	1/22/2000	8.0	6.1	5.6	6.6	9.0	8	8	7.5	5.8	7.4	7.4	Đ	7.9
14	Đỗ Minh Tuấn	1/23/2000	8.0	6.5	6.5	6.6	9.0	8	4.9	6.3	5.6	7.7	6.7	Đ	7.7
15	Nguyễn Tiến Minh	1/24/2000	7.0	6.1	5.9	5.0	9.0	8	7.6	8.4	6.7	6.4	9	Đ	9.0
16	Trần Thị Nhài	1/25/2000	8.5	7.1	6.7	8.6	9.0	8	7	7.0	7	6.9	8	Đ	8.5
17	Nguyễn Thị Thu Hà	2/2/2000	9.5	8.0	8.0	9.0	9.0	8	9	7.0	8	8	9	Đ	8.0

Hình 2 Giao diện tệp đầu vào

-Bước 2: Tạo thêm tệp input2 trong Excel

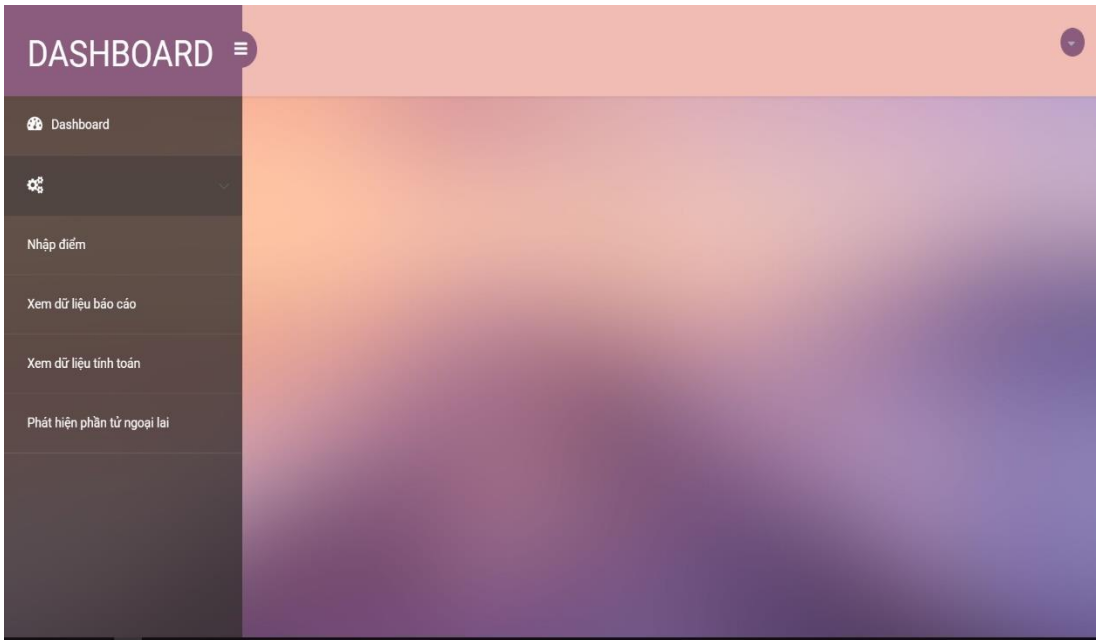
Tệp input2 (tệp dữ liệu tính toán trung gian) là tệp trong Excel được tính từ tệp 1 bằng cách tính lại Điểm TB (theo công thức tính điểm của nhà trường), có thể gọi là Điểm TBTT, thay cột Học lực bằng Học lực TT theo điểm TBTT vừa tính được, thay cột Danh hiệu bằng Danh hiệu TT theo kết quả Học lực TT vừa mới tính được). Sau đó bỏ bớt các cột điểm môn học cho nhẹ tính toán sau này.

-Bước 3: Tạo tệp 3 (tệp dữ liệu tính toán trung gian trong máy) được tính từ tệp 1 và tệp 2 (lấy từ Excel) nhờ phép toán xử lý tệp **Join.**

-Bước 4: Tìm phần tử ngoại lai trên tệp 3 theo thuật toán đã diễn tả trong chương 2.

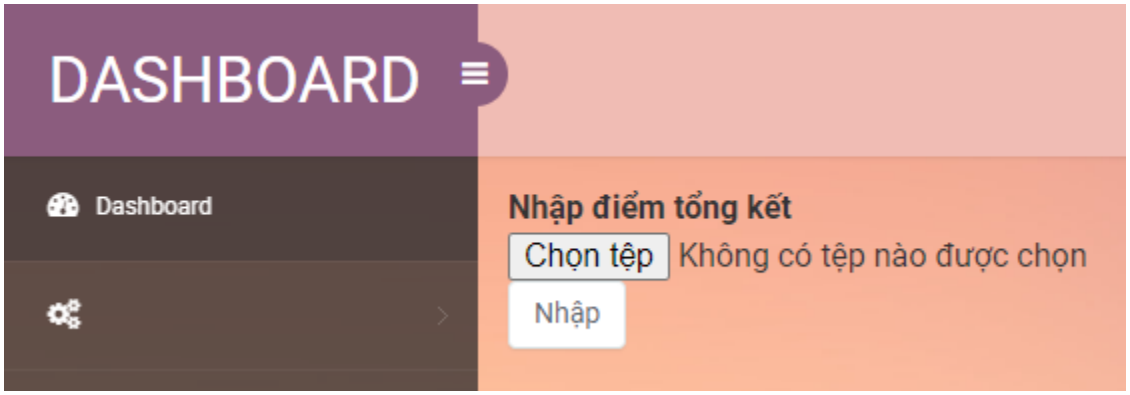
3.5 Một số giao diện chính

3.5.1 Giao diện trang chủ:



Hình 3 Giao diện trang chủ

3.5.2 Giao diện nhập liệu vào hệ thống :



Hình 4 Giao diện nhập File Excel

3.5.3 Giao diện xem dữ liệu báo cáo

Quay lại

Bảng điểm báo cáo

Lấy dữ liệu thành công

Xlsx

Export

STT	Họ tên	Ngày sinh	Toán	Vật lý	Hóa học	Sinh học	Tin học	Ngữ văn	Lịch sử	Địa lý	Tiếng Anh	GDCD	Công nghệ	Thể dục	GDQP-AN	Điểm TB	Học lực	Hạnh kiểm	Danh hiệu
1	Nguyễn Trường Giang	1/10/2000	9	7	9	5.7	9	8	8.5	8.4	9	8.2	9.1	Đ	7.6	8.2	K	Tốt	Tiên tiến
2	Nguyễn Chí Nguyên	1/11/2000	6	7.1	7.2	7.5	9	8	8.5	5.6	6.2	8.6	8.6	Đ	7.8	7.5	K	Tốt	Tiên tiến
3	Trần Văn Tân	1/12/2000	6.6	7	7	6.1	9	8	8.5	6.4	5.5	8.4	8.7	Đ	7.8	7.4	K	Tốt	Tiên tiến
4	Nguyễn Văn Minh	1/13/2000	8	6.8	6	6.1	9	8	8.5	7.1	5.6	7.6	8.5	Đ	8.1	7.4	K	Tốt	Tiên tiến
5	Nguyễn Tuấn Minh	1/14/2000	8	5.5	6.9	6.6	9	8	8.5	7.5	6.3	7.5	8.5	Đ	8.2	7.5	K	Khá	Tiên tiến
6	Nguyễn Văn Ninh	1/15/2000	9	7.1	7.1	7	9	8	8.2	9.1	9.5	6.6	7.6	Đ	8.4	8	G	Khá	Tiên tiến
7	Phạm Thị Hoa	1/16/2000	7	6.5	7.2	7.7	9	8	8.5	8.2	7.1	6.9	8.4	Đ	7.5	7.7	K	Khá	Tiên tiến
8	Ninh Văn Khánh	1/17/2000	8	6	6.1	5.4	9	8	6.9	7.4	8.1	8.1	8.6	Đ	8.1	7.5	K	Tốt	Tiên tiến
9	Đỗ Thị Minh	1/18/2000	9	6.6	5	5	9	8	8.3	5.4	8.7	7.5	7.4	Đ	8.1	7.3	K	Tốt	Tiên tiến
10	Đào Thị Quỳnh	1/19/2000	8	7	6.4	7	9	8	8.7	8	5.9	8.6	6.9	Đ	8.1	7.6	K	Tốt	Tiên tiến
11	Vũ Thị Hải Yến	1/20/2000	8	7.9	6.4	6.8	9	8	7.9	6	8.3	8	6	Đ	7.9	7.5	K	Khá	Tiên tiến
12	Văn Đức Vũ	1/21/2000	9	6.3	7.6	6.2	9	8	8.2	8.1	6.4	8.8	8.5	Đ	7.8	7.8	K	Tốt	Tiên tiến
13	Đỗ Minh Tiến	1/22/2000	8	6.1	5.6	6.6	9	8	8	7.5	5.8	7.4	7.4	Đ	7.9	7.3	K	Tốt	Tiên tiến
14	Đỗ Minh Tuấn	1/23/2000	8	6.5	6.5	6.6	9	8	4.9	6.3	5.6	7.7	6.7	Đ	7.7	7	TB	Tốt	Trung bình
15	Nguyễn Tiến Minh	1/24/2000	7	6.1	5.9	5	9	8	7.6	8.4	6.7	6.4	9	Đ	9	7.3	K	Trung bình	Trung bình
16	Trần Thị Nhài	1/25/2000	8.5	7.1	6.7	8.6	9	8	7	7	7	6.9	8	Đ	8.5	7.7	K	Khá	Tiên tiến
17	Nguyễn Thị Thu Hà	2/2/2000	9.5	8	8	9	9	8	9	7	8	8	9	Đ	8	8.4	G	Tốt	Giỏi

Hình 5 Giao diện xem dữ liệu báo cáo

3.5.4 Giao diện tính toán trung gian (tệp 3):

Quay lại

Bảng điểm tính toán

Tính toán thành công

Xlsx

Export

STT	Họ tên	Ngày sinh	Toán	Vật lý	Hóa học	Sinh học	Tin học	Ngữ văn	Lịch sử	Địa lý	Tiếng Anh	GDCD	Công nghệ	Thể dục	GDQP-AN	Điểm TB	Học lực TT	Hạnh kiểm	Danh hiệu TT
1	Nguyễn Trường Giang	1/10/2000	9	7	9	5.7	9	9	8.5	8.4	9	8.2	9.1	Đ	7.6	8.2	K	Tốt	Tiên tiến
2	Nguyễn Chí Nguyên	1/11/2000	6	7.1	7.2	7.5	9	9	8.5	5.6	6.2	8.6	8.6	Đ	7.8	7.5	K	Tốt	Tiên tiến
3	Trần Văn Tân	1/12/2000	6.6	7	7	6.1	9	9	8.5	6.4	5.5	8.4	8.7	Đ	7.8	7.4	K	Tốt	Tiên tiến
4	Nguyễn Văn Minh	1/13/2000	8	6.8	6	6.1	9	9	8.5	7.1	5.6	7.6	8.5	Đ	8.1	7.4	K	Tốt	Tiên tiến
5	Nguyễn Tuấn Minh	1/14/2000	8	5.5	6.9	6.6	9	9	8.5	7.5	6.3	7.5	8.5	Đ	8.2	7.5	K	Khá	Tiên tiến
6	Nguyễn Văn Ninh	1/15/2000	9	7.1	7.1	7	9	9	8.2	9.1	9.5	6.6	7.6	Đ	8.4	8	G	Khá	Tiên tiến
7	Phạm Thị Hoa	1/16/2000	7	6.5	7.2	7.7	9	9	8.5	8.2	7.1	6.9	8.4	Đ	7.5	7.7	K	Khá	Tiên tiến
8	Ninh Văn Khánh	1/17/2000	8	6	6.1	5.4	9	9	6.9	7.4	8.1	8.1	8.6	Đ	8.1	7.5	K	Tốt	Tiên tiến
9	Đỗ Thị Minh	1/18/2000	9	6.6	5	5	9	9	8.3	5.4	8.7	7.5	7.4	Đ	8.1	7.3	K	Tốt	Tiên tiến
10	Đào Thị Quỳnh	1/19/2000	8	7	6.4	7	9	9	8.7	8	5.9	8.6	6.9	Đ	8.1	7.6	K	Tốt	Tiên tiến
11	Vũ Thị Hải Yến	1/20/2000	8	7.9	6.4	6.8	9	9	7.9	6	8.3	8	6	Đ	7.9	7.5	K	Khá	Tiên tiến
12	Văn Đức Vũ	1/21/2000	9	6.3	7.6	6.2	9	9	8.2	8.1	6.4	8.8	8.5	Đ	7.8	7.8	K	Tốt	Tiên tiến
13	Đỗ Minh Tiến	1/22/2000	8	6.1	5.6	6.6	9	9	8	7.5	5.8	7.4	7.4	Đ	7.9	7.3	K	Tốt	Tiên tiến
14	Đỗ Minh Tuấn	1/23/2000	8	6.5	6.5	6.6	9	9	4.9	6.3	5.6	7.7	6.7	Đ	7.7	7	TB	Tốt	Trung bình
15	Nguyễn Tiến Minh	1/24/2000	7	6.1	5.9	5	9	9	7.6	8.4	6.7	6.4	9	Đ	9	7.3	K	Trung bình	Trung bình
16	Trần Thị Nhài	1/25/2000	8.5	7.1	6.7	8.6	9	9	7	7	7	6.9	8	Đ	8.5	7.7	K	Khá	Tiên tiến
17	Nguyễn Thị Thu Hà	2/2/2000	9.5	8	8	9	9	9	9	7	8	8	9	Đ	8	8.4	G	Tốt	Giỏi

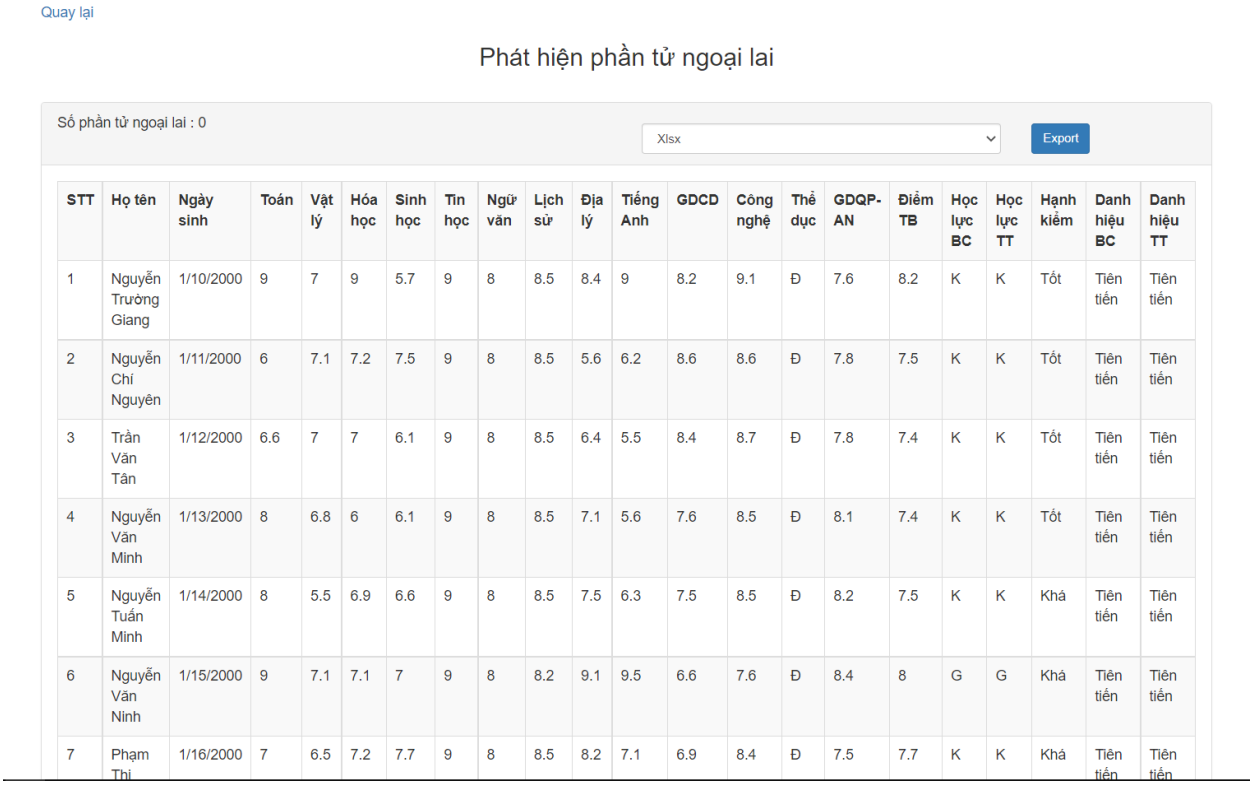
Hình 6 Giao diện tính toán trung gian

Đây là bảng cơ sở để xác định phần tử ngoại lai.

3.5.5 Giao diện phát hiện phần tử ngoại lai:

a/ Trường hợp không phát hiện được phần tử ngoại lai:

So sánh bằng máy giá trị 2 cột Điểm TB với điểm TBTT, hoặc Học lực BC với Học lực TT, hoặc Danh hiệu BC với Danh hiệu TT nếu không bằng nhau thì dòng đó là phần tử ngoại lai. Ở đây ta chọn tệp dữ liệu đúng (Input1) để thử nghiệm trường hợp thứ nhất là không có phần tử ngoại lai.



STT	Họ tên	Ngày sinh	Toán	Vật lý	Hóa học	Sinh học	Tin học	Ngữ văn	Lịch sử	Địa lý	Tiếng Anh	GDCD	Công nghệ	Thể dục	GDQP-AN	Điểm TB	Học lực BC	Học lực TT	Hạnh kiểm	Danh hiệu BC	Danh hiệu TT
1	Nguyễn Trường Giang	1/10/2000	9	7	9	5.7	9	8	8.5	8.4	9	8.2	9.1	Đ	7.6	8.2	K	K	Tốt	Tiên tiến	Tiên tiến
2	Nguyễn Chí Nguyên	1/11/2000	6	7.1	7.2	7.5	9	8	8.5	5.6	6.2	8.6	8.6	Đ	7.8	7.5	K	K	Tốt	Tiên tiến	Tiên tiến
3	Trần Văn Tân	1/12/2000	6.6	7	7	6.1	9	8	8.5	6.4	5.5	8.4	8.7	Đ	7.8	7.4	K	K	Tốt	Tiên tiến	Tiên tiến
4	Nguyễn Văn Minh	1/13/2000	8	6.8	6	6.1	9	8	8.5	7.1	5.6	7.6	8.5	Đ	8.1	7.4	K	K	Tốt	Tiên tiến	Tiên tiến
5	Nguyễn Tuấn Minh	1/14/2000	8	5.5	6.9	6.6	9	8	8.5	7.5	6.3	7.5	8.5	Đ	8.2	7.5	K	K	Khá	Tiên tiến	Tiên tiến
6	Nguyễn Văn Ninh	1/15/2000	9	7.1	7.1	7	9	8	8.2	9.1	9.5	6.6	7.6	Đ	8.4	8	G	G	Khá	Tiên tiến	Tiên tiến
7	Phạm Thị	1/16/2000	7	6.5	7.2	7.7	9	8	8.5	8.2	7.1	6.9	8.4	Đ	7.5	7.7	K	K	Khá	Tiên tiến	Tiên tiến

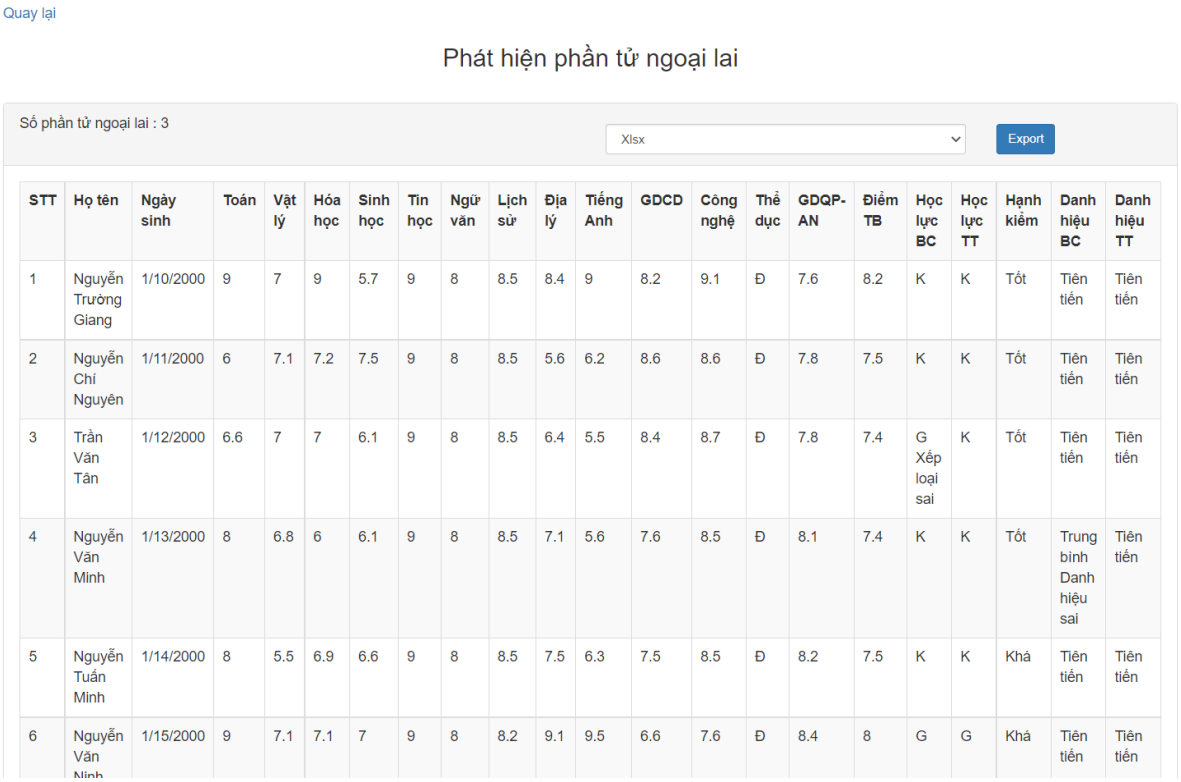
Hình 7 Trường hợp không phát hiện phần tử ngoại lai

b/ Trường hợp phát hiện phần tử ngoại lai:

Để thử nghiệm trường hợp thứ 2 có phần tử ngoại lai, chúng ta tạo tệp **Input1b** mới trong Excel (tệp có sai sót), sau đó nhập vào chương trình và thực

hiện theo 4 bước giống như quy trình thử nghiệm ở quá trình trên, chúng ta sẽ thu được kết qủa có phần tử ngoại lai.

Giao diện kết quả có chứa phần tử ngoại lai:


Quay lại

Phát hiện phần tử ngoại lai

Số phần tử ngoại lai : 3

Xlsx

Export

STT	Họ tên	Ngày sinh	Toán	Vật lý	Hóa học	Sinh học	Tin học	Ngữ văn	Lịch sử	Địa lý	Tiếng Anh	GDCD	Công nghệ	Thể dục	GDQP-AN	Điểm TB	Học lực BC	Học lực TT	Hạnh kiểm	Danh hiệu BC	Danh hiệu TT
1	Nguyễn Trường Giang	1/10/2000	9	7	9	5.7	9	8	8.5	8.4	9	8.2	9.1	Đ	7.6	8.2	K	K	Tốt	Tiên tiến	Tiên tiến
2	Nguyễn Chí Nguyên	1/11/2000	6	7.1	7.2	7.5	9	8	8.5	5.6	6.2	8.6	8.6	Đ	7.8	7.5	K	K	Tốt	Tiên tiến	Tiên tiến
3	Trần Văn Tân	1/12/2000	6.6	7	7	6.1	9	8	8.5	6.4	5.5	8.4	8.7	Đ	7.8	7.4	G Xếp loại sai	K	Tốt	Tiên tiến	Tiên tiến
4	Nguyễn Văn Minh	1/13/2000	8	6.8	6	6.1	9	8	8.5	7.1	5.6	7.6	8.5	Đ	8.1	7.4	K	K	Tốt	Trung bình Danh hiệu sai	Tiên tiến
5	Nguyễn Tuấn Minh	1/14/2000	8	5.5	6.9	6.6	9	8	8.5	7.5	6.3	7.5	8.5	Đ	8.2	7.5	K	K	Khá	Tiên tiến	Tiên tiến
6	Nguyễn Văn Ninh	1/15/2000	9	7.1	7.1	7	9	8	8.2	9.1	9.5	6.6	7.6	Đ	8.4	8	G	G	Khá	Tiên tiến	Tiên tiến

Hình 8 Trường hợp phát hiện phần tử ngoại lai

Các giao diện kết quả ngoại lai có thể được lưu ra tệp Excel để báo cáo!

Từ kết quả này, chúng ta phản hồi lại cho giáo viên chủ nhiệm để điều chỉnh lại dòng điểm của học sinh (ứng với phần tử ngoại lai) trong báo cáo. Có 2 khả năng có thể xảy ra: một là có thể là điểm môn học của học sinh được nhập vào chưa đúng, hai là có thể là điểm được nhập vào đúng nhưng điểm trung bình tính sai hoặc kết luận về học lực hoặc danh hiệu sai (điều này chỉ có giáo viên chủ nhiệm mới có thể sửa được).

3.6. Đánh giá kết quả và hướng mở rộng

-Đánh giá kết quả:

Kết quả thử nghiệm đã đáp ứng tốt yêu cầu thử nghiệm một ứng dụng cụ thể khá đơn giản về phát hiện sai sót (phần tử ngoại lai) trong trường học. Tương tự như mô hình này đồ án có thể ứng dụng cho nhiều bài toán khác trong trường học cũng như trong đời sống kinh tế -xã hội để phát hiện các sai sót trong các bảng dữ liệu.

-Hướng mở rộng:

Vì thời gian và năng lực còn hạn chế nên trong thời gian sau này cho phép, đồ án có thể tự động hoá việc lập bảng Input2 (chứ không lập trong Excel) cũng như tự động hoá đưa kết quả phát hiện có phần tử ngoại lai ra Excel để phản hồi lại cho giáo viên chủ nhiệm (vì người lập trình không thể tự ý sửa được).

Phát triển ứng dụng lý thuyết về phần tử ngoại lai sẽ được nghiên cứu sâu hơn cho cả nội dung phụ thuộc hàm dạng đặc biệt loại phức tạp hơn như phụ thuộc hàm xấp xỉ, phụ thuộc hàm mờ,...

KẾT LUẬN ĐỒ ÁN TỐT NGHIỆP

Đề tài tốt nghiệp "Xác định phần tử ngoại lai dựa vào phụ thuộc hàm đặc biệt trong cơ sở dữ liệu quan hệ và ứng dụng" đã có ý nghĩa khoa học và tính thực tế nhất định. Việc ứng dụng chủ đề về phần tử ngoại lai trong đời sống kinh tế - xã hội – giáo dục đang ngày được mở rộng và phát triển.

Nội dung đồ án về cơ bản đã đáp ứng đầy đủ các yêu cầu cần giải quyết đề ra trong nhiệm vụ đề tài tốt nghiệp:

- Tổng quan được những vấn đề chung nhất về phụ thuộc hàm và phần tử ngoại lai trong cơ sở dữ liệu quan hệ.
- Tìm hiểu rõ về phương pháp phát hiện phần tử ngoại lai đối với phụ thuộc hàm trong cơ sở dữ liệu quan hệ.
- Biết ứng dụng lý thuyết phần tử ngoại lai để kiểm tra đúng sai kết luận xếp loại học lực và danh hiệu cho học sinh một trường THPT ở Hải Phòng.
- Phần thử nghiệm chương trình tuy còn đơn giản nhưng đã chứng tỏ được khả năng nhận thức và khả năng vận dụng kiến thức lý thuyết học được trong nhà trường vào giải quyết những bài toán thực tế.

Kết quả đạt được trong đồ án là sự cố gắng của bản thân em và chất lượng giáo dục của nhà trường. Em xin chân thành cám ơn các Thầy Cô Trường Đại học Quản lý và Công nghệ Hải Phòng, đặc biệt là Thầy Lê Văn Phùng người đã trực tiếp hướng dẫn em thực hiện đồ án này./.

TÀI LIỆU THAM KHẢO

[1] Lê Thanh Hà (2016), *Nghiên cứu phụ thuộc hàm và khoá trong CSDL quan hệ*, Luận văn Thạc sĩ, Đại học Sư phạm 2 Hà Nội.

[2] Lê Văn Phùng, Quách Xuân Trưởng (2017), *Khai phá dữ liệu*, Tái bản lần 1, Nhà xuất bản Thông tin và Truyền thông.

[3] Vũ Đức Thi (1997), *Cơ sở dữ liệu-Kiến thức và thực hành*, Nhà xuất bản Thống kê Hà Nội.

[4] Phạm Hạ Thủy (2005), *Xác định phần tử ngoại lai trong cơ sở dữ liệu quan hệ*, Hội thảo khoa học "Một số vấn đề chọn lọc của CNTT", Hải Phòng, tháng 8, Viện Công nghệ Thông tin, Viện Hàn Lâm Khoa học Việt Nam.

PHỤ LỤC

Phụ lục 1- Phép nối 2 file dữ liệu (Join)

Giả sử r là file dữ liệu n cột

t là file dữ liệu m cột và có q cột có tên trùng nhau với r : $A_1,...,A_q$,

ta ký pháp phép nối r với t như sau: $r \rhd\lhd t$

Phép nối được thực hiện như sau:

1. Tính r x t
2. Thực hiện chọn trong tích Đề-các lấy ra các bản ghi thoả mãn điều kiện các cột có tên trùng nhau thì có cùng giá trị.
3. Dùng phép chiếu loại bỏ q cột (tên giống nhau), mỗi cặp giống nhau chỉ giữ lại một

Như vậy số cột của phép nối là n+m-q

$$r \rhd\lhd t = \Pi\, i_1 i_2 ... i_{n+m-q}(\delta_{rA1=tA1 \wedge ... \wedge rAq=tAq}(r \times t))$$

Ví dụ:

Bảng 1- file r:

A	B	C
0	*1*	*0*
1	*1*	*1*
0	*1*	*1*

Bảng 2- file t:

D	A	E
0	*0*	*0*
1	*1*	*0*

Bảng 3 (file r x t):

rA	B	C	D	tA	E
0	*1*	*0*	*0*	*0*	*0*
0	*1*	*0*	*1*	*1*	*0*
1	*1*	*1*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*0*
0	*1*	*1*	*0*	*0*	*0*
0	*1*	*1*	*1*	*1*	*0*

Bảng 4:

$(\delta_{rA1 = tA1 \wedge ... \wedge rAq=t.Aq}) = \Delta$

rA	*B*	*C*	*D*	*tA*	*E*
0	*1*	*0*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*0*
0	*1*	*1*	*0*	*0*	*0*

Bảng 5:

$\Pi(\Delta) =$

rA	*B*	*C*	*D*	*E*
0	*1*	*0*	*0*	*0*
1	*1*	*1*	*1*	*0*
0	*1*	*1*	*0*	*0*