**TÓM TẮT LUẬN VĂN THẠC SĨ**

Đề tài: **Kỹ thuật phân loại dữ liệu sử dụng thuật toán MCAR trong data mining**..

Tác giả luận văn:.**Trần Hữu Minh**……................………Khóa:..**2015B**………………

Người hướng dẫn:..… **PGS.TS. Phạm Văn Hải** …………………………………........

**NỘI DUNG TÓM TẮT**

1. **Lý do chọn đề tài**

Sự phát triển của công nghệ thông tin trong mọi lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua đồng nghĩa với bùng nổ dữ liệu số. Khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền Công nghệ thông tin hiện nay. Một vấn đề được đặt ra là làm sao trích chọn được thông tin có ý nghĩa từ tập dữ liệu lớn để giải quyết các bài toán thực tế như trợ giúp ra quyết định và dự đoán. Bài toán phân lớp có nhiều giải thuật được đưa ra, phân lớp dựa vào luật phân lớp kết hợp đã được chứng minh là tốt hơn các phương pháp phân lớp dựa vào luật hiện có như cây quyết định, ILA…

1. **Mục đích nghiên cứu của luận văn, đối tượng, phạm vi nghiên cứu.**

Đồ án tập trung nghiên cứu mô hình phân lớp dựa trên giải thuật *Phân loại dựa trên nhiều luật kết hợp - CMAR* nhằm tạo ra một mô hình phân lớp tốt hơn so với các luật liên kết thông thường. Sau đó áp dụng mô hình phân lớp này để giải quyết bài toán chẩn đoán bệnh và dự báo tác động của biến môi trường đối với sức khỏe.

*Đối tượng nghiên cứu* là tập dữ liệu về bệnh vô sinh ở nam giới và bệnh tuyến giáp. *Phạm vi nghiên cứu* tập trung vào xây dựng mô hình phân lớp khai phá thuộc tính bệnh có giá trị và ứng dụng nó trong chuẩn đoán bệnh.

1. **Tóm tắt cô đọng các nội dung chính và đóng góp mới của tác giả**

Khai thác luật phân lớp kết hợp được đề xuất bởi Liu và các đồng sự vào năm 1998. Thuật toán CBA cũng đã được đề xuất trong công trình này. Phương pháp này thường cho độ chính xác cao hơn so với các phương pháp phân lớp dựa trên luật khác như cây quyết định, ILA, v.v. Hiện này đã có nhiều thuật toán được phát triển nhằm làm tăng độ chính xác, giảm thời gian khai thác như CMAR. CMAR đề xuất phương pháp dự đoán lớp của mẫu mới dựa vào đa luật nên thường có độ chính xác cao hơn so với CBA.

* Trình bảy giải thuật CMAR: xây dựng cây CR-tree dựa trên tập các mẫu lớn hơn một giá trị ngưỡng hỗ trợ, dựa trên đó cắt tỉa tạo ra cây FP-tree và khai thác được các luật phân loại.
* Kết hợp các luật phân loại phù hợp để xác định nhãn của mẫu test.
* Trình bày giải thuật Apriori-TFP dựa trên cơ sở giải thuật CMAR và lợi ích vượt trội hơn của giải thuật này so với sử dụng cây FP-tree.
* Xây dựng chương trình dự đoán bệnh dựa trên nguồn dữ liệu “Bệnh vô sinh ở nam giới” và “bệnh suy tuyến giáp”.

1. **Phương pháp nghiên cứu.**

Tiếp cận các bài báo đã nghiên cứu về phân loại và dữ liệu mở của các bệnh nhân , tác giả đề xuất xây dựng mô hình phân loại dựa trên thuật toán CMAR để ứng dụng cho chuẩn đoán xác suất mắc bệnh của người khám thông qua các cơ sở dữ liệu đã lưu lại của các đối tượng bệnh nhân khác trước đó.

Bên cạnh đó tác giả kết hợp và phân tích kết quả của một số phương pháp phân loại để đánh giá kết quả của đề tài.

1. **Kết luận**

Luận văn đã xây dựng được mô hình phân lớp sử dụng thuật toán CMAR kết hợp nhiều luật phân loại nhằm tạo ra một mô hình phân lớp tốt hơn so với các luật liên kết thông thường. Sau đó áp dụng mô hình phân lớp này để giải quyết bài toán chẩn đoán bệnh và dự báo tác động của biến môi trường đối với sức khỏe.