**ĐẠI HỌC THÁI NGUYÊN**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**MA THỊ HỒNG THU**

# HỌC NỬA GIÁM SÁT DỰA TRÊN ĐỒ THỊ VÀ ỨNG DỤNG

**Chuyên ngành: Khoa học máy tính**
**Mã số: 60 48 01 01**

## LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

**NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. ĐOÀN VĂN BAN**

**THÁI NGUYÊN - 2015**

# LỜI CẢM ƠN

Trong quá trình làm luận văn "Học nửa giám sát dựa trên đồ thị và ứng dụng" tôi đã nhận được sự giúp đỡ tận tình của các cá nhân và tập thể.

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc đến thầy giáo PGS.TS Đoàn Văn Ban, người đã tận tình hướng dẫn, chỉ bảo cho tôi trong suốt quá trình thực hiện luận văn.

Xin cùng bày tỏ lòng biết ơn chân thành tới các thầy, cô giáo trong Viện Công nghệ Thông tin cũng như các thầy, cô giáo trong Trường Đại học Công nghệ Thông tin & Truyền thông Thái Nguyên, đã đem lại cho tôi những kiến thức vô cùng có ích trong những năm học tập tại trường.

Cuối cùng tôi xin gửi lời cám ơn đến gia đình, bạn bè, đồng nghiệp những người đã luôn bên cạnh, động viên và khuyến khích tôi trong quá trình thực hiện đề tài nghiên cứu của mình.

*Tôi xin chân thành cảm ơn!*

*Thái Nguyên, ngày 10 tháng 4 năm 2015*

# MỤC LỤC

## DANH MỤC CÁC THUẬT NGỮ VÀ TỪ VIẾT TẮT

| Thuật ngữ | Viết tắt | Ý nghĩa |
|---|---|---|
| Concept | Concept | Khái niệm |
| Self-training | Self-training | Tự huấn luyện |
| Co-training | Co-training | Đồng huấn luyện |
| Machine learning | Machine learning | Học máy |
| Supervised learning | Supervised learning | Học có giám sát |
| Unsupervised learning | Unsupervised learning | Học không giám sát |
| Reinforcement learning | Reinforcement learning | Học tăng cường |
| Semi-supervised learning | Semi-supervised learning | Học nửa giám sát |
| Support vector machine | SVM | Máy véc tơ hỗ trợ |
| Transductive support vector machine | TSVM | Máy véc tơ hỗ trợ truyền dẫn |
| Labeled Propagation | Labeled Propagation | Lan truyền nhãn |
| Graph-based | Graph-based | Dựa trên đồ thị |

## DANH MỤC HÌNH VẼ

# LỜI MỞ ĐẦU

## 1. Lý do chọn đề tài

Học máy (Machine learning) là một ngành khoa học nghiên cứu các kĩ thuật, các phương pháp cho phép các máy tính có khả năng "học" giống như con người. Hay nói một cách khác cụ thể hơn, học máy là một phương pháp để tạo ra các chương trình máy tính bằng việc phân tích các tập dữ liệu, qua đó máy tính có khả năng tích lũy được tri thức thông qua việc học được các khái niệm để có thể ra quyết định trong các trường hợp tương tự.

Lĩnh vực học máy truyền thống thường được chia thành bốn lĩnh vực con, bao gồm: Học có giám sát (Supervised learning), Học không giám sát (Unsupervised learning), Học nửa giám sát (Semi-Supervised learning) và Học tăng cường (Reinforcement learning).

Học nửa giám sát sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện - điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn. Học nửa giám sát đứng giữa học không giám sát (không có bất kì dữ liệu có nhãn nào) và có giám sát (toàn bộ dữ liệu đều được gán nhãn). Để gán nhãn dữ liệu cho một bài toán học máy thường đòi hỏi một phân loại bằng tay các ví dụ huấn luyện. Chi phí cho quy trình này khiến tập dữ liệu được gán nhãn hoàn toàn trở nên không khả thi, trong khi dữ liệu không gán nhãn thường có chi phí thấp. Trong tình huống đó, học nửa giám sát có giá trị thực tiễn lớn lao. Chính vì vậy, học nửa giám sát là sự kết hợp một số lượng lớn các dữ liệu chưa được gán nhãn cùng với các dữ liệu đã được gán nhãn để xây dựng các bộ phân lớp tốt hơn.

Một số phương pháp điển hình trong lĩnh vực này được kể đến như: Phương pháp EM với mô hình sinh hỗn hợp (EM with generative mixture models), phương pháp Tự huấn luyện (Self-training), phương pháp Đồng huấn luyện (Co-training), phương pháp máy véc tơ hỗ trợ (Transductive support vector machines) và phương pháp Dựa trên đồ thị (Graph-based). Trong đó phương pháp học nửa giám sát dựa trên đồ thị (Graph-based) đang là hướng nghiên cứu mở và đem lại hiệu quả lớn.

Với những lý do trên, tác giả đã chọn đề tài "**Học nửa giám sát dựa trên đồ**

**thị và ứng dụng**" làm đề tài nghiên cứu luận văn tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính.

## 2. Mục đích nghiên cứu

Nghiên cứu tổng quan về học nửa giám sát và một số phương pháp học nửa giám sát.

Nghiên cứu phương pháp học nửa giám sát dựa trên đồ thị

Cài đặt thử nghiệm thuật toán lan truyền nhãn trên đồ thị và thuật toán tự huấn luyện.

## 3. Đối tượng và phạm vi nghiên cứu

*Đối tượng nghiên cứu*: Học nửa giám sát.

*Phạm vi nghiên cứu:*

- Nghiên cứu tổng quan về học có giám sát, học không giám sát và học nửa giám sát.
- Các phương pháp học nửa giám sát phổ biến.
- Phương pháp học nửa giám sát dựa trên đồ thị (Graph-based) và một số thuật toán.
- Cài đặt thử nghiệm thuật toán lan truyền nhãn trong phương pháp học nửa giám sát dựa trên đồ thị và thuật toán tự huấn luyện.

## 4. Tóm tắt luận điểm cơ bản

Các luận điểm chính mà luận văn đã thể hiện được:

Nghiên cứu tổng quan và đánh giá các phương pháp học nửa giám sát, tập trung vào phương pháp học nửa giám sát dựa trên đồ thị.

Tập trung tìm hiểu một số thuật toán trong lĩnh vực học nửa giám sát như: Phương pháp EM với mô hình sinh hỗn hợp, phương pháp Tự huấn luyện, phương pháp Đồng huấn luyện và phương pháp máy véc tơ hỗ trợ. Đồng thời tập trung nghiên cứu chi tiết phương pháp dựa trên đồ thị.

Cài đặt phần mềm thử nghiệm mô phỏng thuật toán lan truyền nhãn và thuật toán tự huấn luyện, đánh giá độ phức tạp của hai thuật toán này.

## 5. Phương pháp nghiên cứu

- Đọc tài liệu, phân tích, tổng hợp.
- Thống kê, phân tích dữ liệu.
- Thực nghiệm và đánh giá kết quả.
- Kết hợp nghiên cứu lý thuyết, tìm hiểu tình hình ứng dụng, đánh giá khả năng ứng dụng và đề xuất giải pháp.

## 6. Nội dung luận văn

Nội dung luận văn gồm 03 chương:

- Chương 1: Tổng quan về các phương pháp học máy

  Chương này trình bày tổng quan về các phương pháp học máy gồm phương pháp Học có giám sát (Supervised learning), Học không giám sát (Unsupervised learning), Học nửa giám sát (Semi-Supervised learning).

- Chương 2: Phương pháp học nửa giám sát dựa trên đồ thị

  Tập trung tìm hiểu một số thuật toán trong lĩnh vực học nửa giám sát như: Phương pháp EM với mô hình sinh hỗn hợp, phương pháp Tự huấn luyện, phương pháp Đồng huấn luyện và phương pháp máy véc tơ hỗ trợ. Đồng thời tập trung nghiên cứu chi tiết phương pháp dựa trên đồ thị.

- Chương 3: Cài đặt và thử nghiệm thuật toán

  Cài đặt thử nghiệm thuật toán tự huấn luyện và lan truyền nhãn dựa trên đồ thị, đánh giá độ phức tạp của hai thuật toán này.

# CHƯƠNG 1:
# TỔNG QUAN VỀ CÁC PHƯƠNG PHÁP HỌC MÁY

## 1.1. Giới thiệu về học máy

**Học máy** (*Machine Learning*) là một ngành khoa học nghiên cứu các ***thuật toán*** cho phép máy tính có thể học được các khái niệm (*concept*)[7].

Có hai loại phương pháp học máy chính:

- Phương pháp quy nạp: Máy học/phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụng được nguồn dữ liệu rất nhiều và sẵn có.
- Phương pháp suy diễn: Máy học/phân biệt các khái niệm dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ máy tính.

Hiện nay, các thuật toán đều cố gắng tận dụng được ưu điểm của hai phương pháp này.

Các ngành khoa học liên quan đến lĩnh vực học máy điển hình là:

- Lý thuyết thống kê: các kết quả trong xác suất thống kê là tiền đề cho rất nhiều phương pháp học máy. Đặc biệt, lý thuyết thống kê cho phép ước lượng sai số của các phương pháp học máy.
- Các phương pháp tính: các thuật toán học máy thường sử dụng các tính toán số thực/số nguyên trên dữ liệu rất lớn. Trong đó, các bài toán như: tối ưu có/không ràng buộc, giải phương trình tuyến tính v.v… được sử dụng rất phổ biến.
- Khoa học máy tính: là cơ sở để thiết kế các thuật toán, đồng thời đánh giá thời gian chạy, bộ nhớ của các thuật toán học máy.
- Lĩnh vực học máy truyền thống thường được chia thành bốn lĩnh vực con:
- Học có giám sát: Máy tính được xem một số mẫu gồm đầu vào và đầu ra tương ứng trước. Sau khi học xong các mẫu này, máy tính quan sát một đầu vào mới và cho ra kết quả.
- Học không giám sát: Máy tính chỉ được xem các mẫu không có đầu ra, sau đó máy tính phải tự tìm cách phân loại các mẫu này và các mẫu mới.