ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRƯƠNG ĐỨC CƯỜNG

PHÂN CỤM DỮ LIỆU SỬ DỤNG GIẢI THUẬT DI TRUYỀN VÀ MẠNG NƠ RON

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

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

TRƯƠNG ĐỨC CƯỜNG

PHÂN CỤM DỮ LIỆU SỬ DỤNG GIẢI THUẬT DI TRUYỀN VÀ MẠNG NƠ RON

Chuyên ngành **: *Khoa học máy tính***
Mã số **: *60.48.01***

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. Vũ Mạnh Xuân

Thái Nguyên - 2012

LỜI CẢM ƠN

Em xin bày tỏ lòng biết ơn sâu sắc tới TS. Vũ Mạnh Xuân, thầy đã hướng dẫn, chỉ dạy tận tình để em hoàn thành luận văn này. Em xin chân thành cảm ơn các thầy, cô giáo Trường Đại học Công nghệ Thông tin & Truyền thông - Đại học Thái Nguyên, cùng các thầy, cô giáo Viện Công nghệ Thông tin - Viện Khoa học và Công nghệ Việt Nam đã truyền thụ kiến thức cho em trong suốt quá trình học tập vừa qua.

Tôi cũng xin cảm ơn cơ quan, bạn bè đồng nghiệp, gia đình và những người thân đã cùng chia sẻ, giúp đỡ, động viên, tạo mọi điều kiện thuận lợi để tôi có thể học tập và hoàn thành bản luận văn này.

Tuy đã có những cố gắng nhất định nhưng do thời gian và trình độ có hạn nên chắc chắn luận văn còn nhiều thiếu sót và hạn chế nhất định. Rất mong nhận được sự góp ý của thầy cô và các bạn.

Thái Nguyên, ngày 27 tháng 06 năm 2012

Học viên

Trương Đức Cường

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm nghiên cứu, tìm hiểu của riêng cá nhân tôi. Trong toàn bộ nội dung của luận văn, những điều được trình bày hoặc là của cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Thái Nguyên, ngày 27 tháng 06 năm 2012

Học viên

Trương Đức Cường

MỤC LỤC

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG BIỂU

DANH SÁCH TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
KPDL	Khai phá dữ liệu
PCDL	Phân cụm dữ liệu
CSDL	Cơ sở dữ liệu
GA	Giải thuật di truyền
NST	Nhiễm sắc thể

MỞ ĐẦU

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin và ngành công nghiệp phần cứng đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Bên cạnh đó việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo cho chúng ta một hệ thống cơ sở dữ liệu khổng lồ. Hệ thống này đã đem lại những lợi ích vô cùng to lớn cho con người trong việc lưu trữ, tìm kiếm và thống kê. Tuy vậy, sự bùng nổ này đã dẫn tới một nhu cầu mới là phát hiện tri thức từ kho dữ liệu khổng lồ đó. Đây là một vấn đề rất phức tạp, cần phải có những công cụ và kỹ thuật xử lý linh hoạt như suy nghĩ của con người.

Trong ngành khoa học máy tính, tìm kiếm lời giải tối ưu cho các bài toán là vấn đề được các nhà khoa học máy tính đặc biệt rất quan tâm. Mục đích chính của các thuật toán là tìm kiếm thuật giải chất lượng cao và sử dụng kỹ thuật trí tuệ nhân tạo đặc biệt rất cần thiết khi giải quyết các bài toán có không gian tìm kiếm lớn.

Giải thuật di truyền (Genetic Algorithm - GA) là một trong những kỹ thuật tìm kiếm lời giải tối ưu đã đáp ứng được yêu cầu của nhiều bài toán và ứng dụng. Hiện nay, thuật toán di truyền cùng với mạng nơ ron được ứng dụng rất rộng rãi trong các lĩnh vực phức tạp. Thuật toán di truyền kết hợp với mạng nơ ron chứng tỏ được hiệu quả của nó trong các vấn đề khó có thể giải quyết bằng các phương pháp thông thường hay các phương pháp cổ điển, nhất là trong các bài toán cần có sự lượng giá, đánh giá sự tối ưu của kết quả thu được.

Chính vì vậy, trong phạm vi đề tài này, tôi chọn hướng kết hợp giữa mạng nơ ron và giải thuật di truyền áp dụng vào bài toán phân cụm dữ liệu,