PHÂN CỤM NỬA GIÁM SÁT DỰA TRÊN ĐỒ THỊ

Vũ Việt Vũ, Vũ Việt Thắng, Nicolas Labroche, Bernadette Bouchon Meunier, Nguyễn Thị Thu Hiền

TÓM TẮT:

Thuật toán phân cụm nửa giám sát sử dụng một số lượng ít các dữ liệu đã gán nhãn (seeds) hoặc một số ràng buộc (must-link hoặc can-not link) giữa các dữ liệu nhằm mục đích cải tiến chất lượng của bài toán phân cụm. Trong bài báo này, chúng tôi mở rộng một thuật toán phân cụm nửa giám sát sử dụng các seed bằng cách thêm vào một kỹ thuật học tích cực (active learning) để thu thập các ràng buộc từ người sử dụng. Theo chúng tôi biết đây là thuật toán đầu tiên trên thế giới sử dụng đồng thời cả hai loại seed và constraint vào trong cùng một quá trình phân cụm. Kết quả thực nghiệm cho thấy thuật toán của chúng tôi cải tiến đáng kể chất lượng của quá trình phân cụm trên các tập dự liệu thực.

Abstract: Semi-supervised clustering algorithms relies on side information, either labeled data (seeds) or pairwise constraints (must-link or can-not link) between data objects, to improve the quality of clustering. In this paper, we propose to extend an exisiting seed-based clustering algorithm with an active learning mechanism to collect pairwise constraints. Thus, to the best of our knowledge, our new semi-supervised algorithm is the first to work with both seeds and constraints. Preliminary results on real data sets show the benefit of our algorithm when compared to the initial seed-based clustering algorithm.