ACTIVE LEARNING FOR SEMI-SUPERVISED K-MEANS CLUSTERING

Vũ Việt Vũ, Nicolas Labroche, and Bernadette Bouchon-Meunier

TÓM TẮT:

K-Means algorithm is one of the most used clustering algorithm for Knowledge Discovery in Data Mining. SeedbasedK-Means is the integration of a small set of labeled data (called seeds) to the K-Means algorithm to improve itsperformances and overcome its sensitivity to initial centers. These centers are, most of the time, generated at random orthey are assumed to be available for each cluster. This paper introduces a new efficient algorithm for active seeds selectionwhich relies on a Min-Max approach that favors the coverage of the whole dataset. Experiments conducted on artificial and realdatasets show that, using our active seeds selection algorithm, each cluster contains at least one seed after a very smallnumber of queries and thus helps reducing the number of iterations until convergence which is crucial in many KDD applications.