ACTIVE SEEDS SELECTION WITH A K-NEAREST NEIGHBORS GRAPH

Vũ Việt Vũ, Nicolas Labroche, Violaine Antoine, Lê Bá Dũng

TÓM TẮT:

Active learning allows semi-supervised clustering algorithms to sol-
licit domain experts to retrieve a few set of class labels (or seeds) to improve their
efficiency or the relevance of their results. However, some recent studies show
that, even in the case of a good answer from the domain expert, semi-supervised
clustering can see their performances drop with badly chosen seeds. Until now,
only few works address the problem of determining the best queries for a cluster-
ing algorithm in an active learning context, and most of these studies are limited
because of their hypothesis on the size and the shape of expected clusters. In this
paper, we propose a new active seed selection algorithm that makes no hypothe-
sis on the underlying data distribution. Experiments conducted on real data sets
show the efficiency of this new approach compared to existing ones.
Active learning allows semi-supervised clustering algorithms to sol-licit domain experts to retrieve a few set of class labels (or seeds) to improve theirefficiency or the relevance of their results. However, some recent studies showthat, even in the case of a good answer from the domain expert, semi-supervisedclustering can see their performances drop with badly chosen seeds. Until now,only few works address the problem of determining the best queries for a cluster-ing algorithm in an active learning context, and most of these studies are limitedbecause of their hypothesis on the size and the shape of expected clusters. In thispaper, we propose a new active seed selection algorithm that makes no hypothe-sis on the underlying data distribution. Experiments conducted on real data setsshow the efficiency of this new approach compared to existing ones.