

**RNN-DBSCAN: A Density-based Clustering Algorithm using Reverse Nearest Neighbor Density Estimates**

**Abstract:**

A new density-based clustering algorithm, RNN-DBSCAN, is presented which uses reverse nearest neighbor counts as an estimate of observation density. Clustering is performed using a DBSCAN-like approach based on k nearest neighbor graph traversals through dense observations. RNN-DBSCAN is preferable to the popular density-based clustering algorithm DBSCAN in two aspects. First, problem complexity is reduced to the use of a single parameter (choice of k nearest neighbors), and second, an improved ability for handling large variations in cluster density (heterogeneous density). The superiority of RNN-DBSCAN is demonstrated on several artificial and real-world datasets with respect to prior work on reverse nearest neighbor based clustering approaches (RECORD, IS-DBSCAN, and ISB-DBSCAN) along with DBSCAN and OPTICS. Each of these clustering approaches is described by a common graph-based interpretation wherein clusters of dense observations are defined as connected components, along with a discussion on their computational complexity. Heuristics for RNN-DBSCAN parameter selection are presented, and the effects of k on RNN-DBSCAN clusterings discussed. Additionally, with respect to scalability, an approximate version of RNN-DBSCAN is presented leveraging an existing approximate k nearest neighbor technique.

**Existing System:**

In density-based clustering, such as the popular DBSCAN [1], the definition of clustering is refined as identifying dense regions in feature space which are separated by regions of low density. With respect to clustering, a dense region is defined by the group of observations lying within it.

Several desirable properties of density-based clustering include an ability to handle and identify noise, discover clusters with arbitrary shapes, and automatic discovery of the number of clusters. Solutions for the density-based clustering problem can be broken up into procedures for performing two tasks.

First, a procedure for estimating the density of each observation is defined and applied to identify observations lying within dense regions (core observations). Second, a region growing procedure is defined that identifies the group of observations which are reachable from some core observation.

Key to the region growing procedure is the definition of observation reachability which must be restricted to the use of dense regions (i.e., no two observations should be reachable through a region of low density).

**Proposed System:**

First, there is a reduction in problem complexity as they only require the use of a single parameter, k, whereas DBSCAN requires two parameters, minpts and eps. Second, DBSCAN’s use of the distancebased threshold eps, under certain conditions, leads to the algorithm’s inability to distinguish amongst clusters with large variations in density. Additionally, to ensure deterministic properties of DBSCAN clustering results, a metric (symmetric) distance measure must be used, whereas the reverse nearest neighbor approaches have no such restriction.

With respect to the second benefit, in DBSCAN eps and minpts are used to define density which is applied globally throughout the data. In contrast, with respect to DBSCAN’s definition of density, density in the reverse nearest neighbor approaches is determined locally (i.e., with respect to observations local neighborhood which is defined by k). This allows these methods to identify dense regions of space, again using DBSCAN’s definition of density, that can vary greatly in density.