Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Add links

Canopy clustering algorithm

From Wikipedia, the free encyclopedia

The **canopy clustering algorithm** is an unsupervised pre-clustering algorithm introduced by Andrew McCallum, Kamal Nigam and Lyle Ungar in 2000.[1] It is often used as preprocessing step for the K-means algorithm or the Hierarchical clustering algorithm. It is intended to speed up clustering operations on large data sets, where using another algorithm directly may be impractical due to the size of the data set.

The algorithm proceeds as follows, using two thresholds T_1 (the loose distance) and T_2 (the tight distance), where $T_1 > T_2$.[1][2]

1. Begin with the set of data points to be clustered.
2. Remove a point from the set, beginning a new 'canopy'.
3. For each point left in the set, assign it to the new canopy if the distance less than the loose distance T_1.
4. If the distance of the point is additionally less than the tight distance T_2, remove it from the original set.
5. Repeat from step 2 until there are no more data points in the set to cluster.
6. These relatively cheaply clustered canopies can be sub-clustered using a more expensive but accurate algorithm.

An important note is that individual data points may be part of several canopies. As an additional speed-up, an approximate and fast distance metric can be used for 3, where a more accurate and slow distance metric can be used for step 4.

Since the algorithm uses distance functions and requires the specification of distance thresholds, its applicability for high-dimensional data is limited by the curse of dimensionality. Only when a cheap and approximative – low-dimensional – distance function is available, the produced canopies will preserve the clusters produced by K-means.

Benefits [edit]

- The number of instances of training data that must be compared at each step is reduced
- There is some evidence that the resulting clusters are improved[3]

References [edit]

1. ^ *a* *b* McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123
2. ^ http://courses.cs.washington.edu/courses/cse590q/04au/slides/DannyMcCallumKDD00.ppt Retrieved 2014-09-06.
3. ^ Mahout description of Canopy-Clustering Retrieved 2011-04-02.

This algorithms or data structures-related article is a stub. You can help Wikipedia by expanding it.

Categories: Data clustering algorithms | Statistical algorithms | Algorithms and data structures stubs | Computer science stubs