Watershed on Vector Quantization for clustering of big data.

S.V. Mitsyn, G.A. Osokov (JINR, Dubna)
E-mail address: ososkov@jinr.ru, svm@jinr.ru

A method for clustering large amounts of data is presented which is a sequenced composition of two algorithms: the first accomplishes a partition of input space into Voronoi regions and the second partitions them into a set of clusters. A model of clusters as high-density regions in input space is proposed, then it is shown how a Voronoi tessellation and it's topological map (a) can be build and (b) used as a low complexity approximation of the input space. Mathematica computational software is used as an appropriate and effective tool to fulfil Delaunay triangulation and a corresponding Voronoi partitionings. During the (b) step, common clustering algorithms, as K-means and single linkage, are used. Results of application two stage algorithm for clustering simulated and real data are presented.

Двухэтапная кластеризация данных большого объема с применением программы Mathematica

С.В. Мицын (ОИЯИ, Дубна)
Г.А. Ососков (ОИЯИ, Дубна)
E-mail address: ivanov@ccas.ru, petrova@yandex.ru

Представлен метод кластеризации данных большого объёма в виде последовательной композиции двух алгоритмов: первый строит разбиение входного пространства на области Вороного, а второй кластеризует их. Вначале предложена модель кластеризации данных как областей большой плотности во входном пространстве, затем показано как разбиение Вороного и его топология могут быть (а) построены и (б) использованы как упрощённое приближение входного пространства. Программа Mathematica используется как удобный и эффективный инструмент для осуществления триангуляции Делоне и соответствующего разбиения Вороного. На этапе (б) применялись известные алгоритмы кластеризации К-средних и ближайшего соседа. Представлены результаты применения двухэтапного алгоритма для кластеризации модельных и реальных данных.