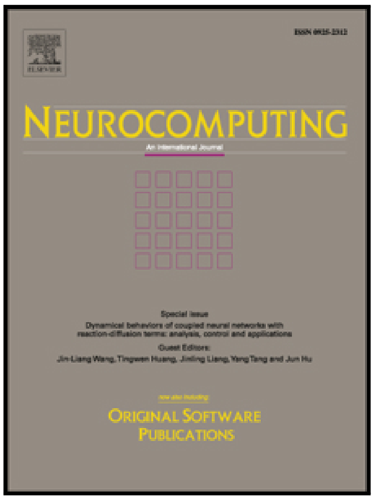
Accepted Manuscript



Quantum clustering in non-spherical data distributions: finding a suitable number of clusters

Raúl Casaña-Eslava, Ian H. Jarman, Paulo J.G. Lisboa, José D. Martín-Guerrero

PII: S0925-2312(17)30751-8
DOI: 10.1016/j.neucom.2017.01.102
Reference: NEUCOM 18375

To appear in: *Neurocomputing*

Received date: 8 July 2016
Revised date: 5 December 2016
Accepted date: 19 January 2017

Please cite this article as: Raúl Casaña-Eslava, Ian H. Jarman, Paulo J.G. Lisboa, José D. Martín-Guerrero, Quantum clustering in non-spherical data distributions: finding a suitable number of clusters, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.01.102

Quantum clustering in non-spherical data distributions: finding a suitable number of clusters

Raúl Casaña-Eslava[a], Ian H. Jarman[a], Paulo J. G. Lisboa[a], José D. Martín-Guerrero[b]

[a] *Liverpool John Moores University (LJMU), Liverpool, United Kingdom,*
E-Mail: `r.v.casanaeslava@ljmu.ac.uk`
[b] *Universitat de València (UV), València, Spain,*
E-Mail:`jose.d.martin@uv.es`

Abstract

Quantum Clustering (QC) provides an alternative approach to clustering algorithms, several of which are based on geometric relationships between data points. Instead, QC makes use of quantum mechanics concepts to find structures (clusters) in data sets by finding the minima of a quantum potential. The starting point of QC is a Parzen estimator with a fixed length scale, which significantly affects the final cluster allocation. This dependence on an adjustable parameter is common to other methods. We propose a framework to find suitable values of the length parameter σ by optimising twin measures of cluster separation and consistency for a given cluster number. This is an extension of the Separation and Concordance framework previously introduced for K-means clustering. Experimental results on two synthetic data sets and three challenging real-world data sets show that optimisation of cluster separation identifies QC solutions with consistently high Jaccard score measured against true-cluster labels while optimisation of cluster consistency provides insights into hierarchical cluster structure.

Keywords: Quantum clustering, Non-spherical data distributions, Number of clusters, Parameter optimization, Separation and Concordance

1. Introduction

As interest in knowledge extraction from data grows, this typically includes exploratory analysis especially when the data are unlabelled. A central step in exploratory data analysis is the discovery of different categories or profiles in the data. Clustering algorithms are efficient methods for unsupervised learning among which a frequently used algorithm is K-Means [1]. This method implements a hard partition of the data by identifying representative points, the prototypes, which minimize the sum of within cluster squared Euclidean distances as shown in Eqs. (1) and (2):