Gaussian Motion Data

Introduction

In the Evolving Clustering literature there is currently a lack of publicly available data sets (synthetic or real) for evaluating algorithms in concept drift situations. The synthetic database presented here can be used to test the ability of the algorithms to follow data drift.

The fact that the data is simulated provides detailed knowledge about the structure underlying the data, enabling a more thorough evaluation of the results. This is particularly important in online clustering for evaluating not only the final result provided by the algorithm, but also the intermediate results.

By having detailed knowledge about the models that generated the data it is possible to accurate assess the performance of the algorithms through all the intermediate states.

The database is composed by several data sets with concept drift which also contains information about the temporal evolution of the models that generated the data.

The data sets have been generated by Gaussian distributions whose mean and/or covariance change over time. In the database both the simulated data and the Gaussians that generated it are provided, hence enabling an accurate evaluation of the partitions through time.

Naming Convention

The data sets are named using the following naming convention: first the number of clusters followed by the letter C, then the number of dimensions of the data set followed by the letter D, then the number of samples (k is used for thousands) and finally a final word roughly describing clusters' movement.

For example, 3C2D2400Spiral is a data set with 3 clusters in 2 dimensions where spiral alike movements are present.

Data set name	Short name
1C2D1kLinear	a
4C2D800Linear	b
4C2D3200Linear	c
3C2D2400Spiral	d
4C3D20kLinear	e
5C5D1kLinear	f
2C3D4kHelix	g
2C2D200kHelix	h
4C2D4kStatic	i

Data Format

Each data set is made up ofthree files.

- SamplesFile_{short name}_{long name}.csv
- MeanFile_{short name}_{longname}.csv
- VarsFile_{short name}_{long name}.csv

The first file contains in each row one of the samples of the data set, being the last column the cluster number. The other files contain in the corresponding rows the mean and covariance matrix that generated that sample.

Reference

```
@article{MARQUEZ201816,
title = "A novel and simple strategy for evolving prototype based clustering",
journal = "Pattern Recognition",
volume = "82",
pages = "16 - 30",
year = "2018",
issn = "0031-3203",
doi = "https://doi.org/10.1016/j.patcog.2018.04.020",
url = "http://www.sciencedirect.com/science/article/pii/S0031320318301547",
author = "David G. Márquez and Abraham Otero and Paulo Félix and Constantino A. García",
keywords = "Evolving clustering, Data stream, Concept drift, Gaussian mixture models, K-means, Cluster evolution"
}
```

License

Maintained by David Gonzalez Marquez.

INFORMACIÓN

Investigadores
Paulo Félix Lamas
David González Márquez
Abraham Otero Quintana
Constantino Antonio García Martínez

DESCARGAR

Repositorio Gitlab

Descargar de Gitlab

PUBLICACIÓNS

A novel and simple strategy for evolving prototype based clustering
Pattern Recognition, 2018