1 Abbreviations

—ABRE—Description— ——-————— —QK-Means—Quantum K-Means —qubit—Quantum bit —PCA—Principal Component Analysis —PC—Principal Component —SVD—Singular Value Decomposition —GPGPU—General-Purpose Computing on Graphics Processing Units

Contents

3 Context

3.1 Clustering with EAC

3.1.1 Clustering

Advances in technology allow for the collection and storage unprecedented amount and variety of data. Since data is mostly stored electronically, it presents a potential for automatic analysis and thus creation of information and knowledge. A growing body of statistical methods aiming to model, structure and/or classify data already exist, e.g. linear regression, principal component analysis, cluster analysis, support vector machines, neural networks. Many of these methods fall into the realm of machine learning, which is usually divided into 2 major groups: *supervised* and *unsupervised* learning. Supervised learning deals with labelled data, i.e. data for which ground truth is known, and tries to solve the problem of classification. Unsupervised learning deals with unlabelled data and tries to solve the problem of clustering.

Cluster analysis is the backbone of the present work. The goal of data clustering, as defined by [2], is the discovery of the *natural grouping(s)* of a set of patterns, points or objects. In other words, the goal of data clustering is to discover structure on data, structured or not. And the methodology used is to group patterns that are similar by some metric (e.g. euclidean distance, Pearson correlation) and separate those that are dissimilar.

Clustering is used in a wide variety of fields to solve numerous problems, e.g.:

- image segmentation in the field of image processing;
- generation of hierarchical structure for easy access and retrieval of information systems;
- recommender systems by grouping a users by their behavious and/or preferences;
- clustering customers for targeted marketing in
- clustering gene expression data in biology;
- grouping of

3.1.2 Ensemble Clustering

Ensemble clustering Data from real world problems appear in different configurations regarding shape, size, sparsity, etc. Different clustering algorithms are appropriate for different data configurations, e.g. K-Means using euclidean distance as metric tends to group patterns in hyperspheres so it is more appropriate for data whose structure is formed by hypershere like clusters.If the true structure of the data at hand is heterogeneous in configuration, a single clustering algorithm might perform well for some part of the data while other performs better for some other part. The underlying idea behind ensemble clustering is to use multiple clusterings from one or more clustering algorithms and combine them in such a way that the final clustering is better than any of the individual ones.

Formulation Some notation and nomenclature, adopted from [1], should be defined since it will be used throughout the remainder of the present work. The term *data* refers to a set X of n objects or patterns $X = \{\mathrm{x}_1, ..., \mathrm{x}_n\}$, and may be represented by $\chi = \{x_1, ..., x_n\}$, such that $x_i \in \mathbb{R}^d$. A clustering algorithm takes χ as input and returns k groups or *clusters* C of some part of the data, which form a *partition* P. A clustering *ensemble* $\mathbb{P}$ is group of such partitions. This means that:

$$\mathbb{P} = \left\{P^1, P^2, ...P^N\right\} \; P^j = \left\{\mathrm{x}_1^j, \mathrm{x}_2^j, ...\mathrm{x}_{k_j}^j\right\} \; C_k^j = \{x_a, x_b, ..., x_z\}$$

overview of EAC The Evidence Accumulation Clustering (EAC) makes no assumption on the number of clusters in each data partition. Its approach is divided in 3 steps:

1. Produce a clustering ensemble $\mathbb{P}$ (the evidence)
2. Combine the evidence
3. Recover natural clusters

A clustering ensemble, according to [1]., can be produced from (1) different data representations, e.g. choice of preprocessing, feature extraction, sampling; or (2) different partitions of the data, e.g. output of different algorithms, varying the initialization parameters.

The ensemble of partitions is combined in the second step, where a non-linear transformations turns the ensemble into a co-association matrix, i.e. a matrix describing the association between any two data patterns. The association between any pair of patterns is given by the number of times those two patterns appear clustered together in any cluster of any partition of the ensemble. The rationale is that pairs that are frequently clustered together are more likely to be representative of a true link between the patters, revealing the underlying structure of the data. The construction of this matrix is at the very core of this method.

The co-association matrix itself doesn't output a clustering partition. Instead, it is used as input to other methods to obtain the final partition. Since this matrix is a similarity matrix it's appropriate to use in algorithms take this type of matrices as input, e.g. hierarchical algorithms such as Single-Link, K-Medoids. Typically, algorithms use a distance as the similarity, which means that they minimize the values of similarity to obtain the highest similarity between objects. However, a low value on the co-association matrix translates in a low similarity between a pair of objects, which means that the co-association matrix requires prior transformation for accurate clustering results, e.g. replace every similarity value n_{ij} between every pair of object (i, j) by $max\{C\} - n_{ij}$.

examples of applications

advantages

disadvantages quadratic space and time complexities because of the nxn co-association matrix

3.2 The Big Data paradigm

examples of success application

characteristics and challenges

5 Methodology

7 Discussion?

7.1 References

References

[1] Ana N L Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[2] Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.