

Preface

Both the amount and the complexity of the data gathered by current scientific and productive enterprises are increasing at an exponential rate, in the most diverse knowledge areas, such as biology, physics, medicine, astronomy, climate forecasting, etc. To find patterns and trends in these data is increasingly important and challenging for decision making. As a consequence, the analysis and management of *Big Data* is currently a central challenge in Computer Science, especially with regards to complex datasets. Finding clusters in large complex datasets is one of the most important tasks in data analyses. For example, given a satellite image database containing several tens of Terabytes, how can we find regions aiming at identifying native rainforests, deforestation, or reforestation? Can it be made automatically? Based on the results of the work discussed in this book, the answers to both questions are a sound “yes”, and the results can be obtained in just minutes. In fact, results that used to require days or weeks of hard work from human specialists can now be obtained in minutes with high precision.

Clustering complex data is a computationally expensive task, and the best existing algorithms have a super-linear complexity regarding both the data set cardinality (number of data elements) and dimensionality (number of attributes describing each element). Therefore, those algorithms do not scale well, precluding being efficient to process large data sets. Focused on the analysis of *Big Data*, this book discusses new algorithms created to perform clustering in moderate-to-high dimensional data involving many billions of elements stored in several Terabytes of data, such as features extracted from large sets of complex objects, but that can nonetheless be quickly executed, in just a few minutes.

To achieve that performance, it was taken into consideration that high-dimensional data have the clusters bounded to a few dimensions each, thus existing only in subspaces of the original high-dimensional space, although each cluster can have correlations among dimensions distinct from those dimensions correlated in the other clusters. The novel techniques were developed to perform both hard and soft clustering (that is, assuming that each element can participate in just one or in several clusters that overlap in the space) that can be executed by

serial or by parallel processing. Moreover, their applications are shown in several practical test cases.

Distinctly from most of the existing algorithms (and from all of the fastest ones), the new clustering techniques do not require the previous definition of the number of expected clusters, rather, it is inferred from the data and returned to the user. Besides, due to the assumption that each cluster exists because of correlations existing in a subset of the space dimensions, the new techniques not only find clusters with high quality and speed, but also spot the most significant dimensions for each cluster, a benefit that the previous algorithms only achieve at the expenses of costly processing.

The methodology to develop the techniques discussed in this book was based on the extension of hierarchical data structures, multidimensional multi-scaling analysis of the spatial data distribution based on a convolution process using Laplacian filters, on the evaluation of alternative cluster entropies, and on new cost functions that enable to evaluate the best strategies before executing them, allowing to perform a dynamic dataflow optimization of the parallel processing.

The new algorithms were compared with at least nine of the most efficient existing ones, and it was shown that the performance improvement is over at least one magnitude order, although always having its quality equivalent to the best achieved by the competing techniques. In extreme situations, it took just *two seconds* to obtain clusters from real data that the best competing techniques required *two days*, with equivalent accuracy. In one of the real cases evaluated, the new techniques described were able to find correct tags for every image from a data set containing several tens of thousands of images, performing soft clustering (thus assigning one or more tags to each image), using as guidelines the labeling performed by a user in not more than five images for each tag (that is, in at most 0.001 % of the image set). Experiments reported in the book show that the novel techniques achieved excellent results in real data from high impact applications, such as breast cancer diagnosis, region classification in satellite images, assistance to climate change forecast, recommendation systems for the Web, and social networks.

In summary, the work described here takes steps forward from traditional data mining (especially for clustering) by considering large, complex data sets. Note that, usually, current works focus on one aspect, either size or data complexity. The work described in this book considers both: it enables mining complex data from high impact applications; the data are large in the Terabyte-scale, not in Giga as usual; and very accurate results are found in just minutes. Thus, it provides a crucial and well-timed contribution for allowing the creation of *real time* applications that deal with *Big Data of high complexity* in which mining on the fly can make an immeasurable difference, like in cancer diagnosis or deforestation detection.

São Carlos, October 2012

Robson L. F. Cordeiro
Christos Faloutsos
Caetano Traina Júnior

Data Mining in Large Sets of Complex Data

Ferreira Cordeiro, R.L.; Faloutsos, C.; Traina Júnior, C.

2013, XI, 116 p. 37 illus., 25 illus. in color., Softcover

ISBN: 978-1-4471-4889-0