available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cosrev

Book review

Review of “Introduction to clustering large and high-dimensional data” by J. Kogan

Dieter Mitsche

Institut für Theoretische Informatik, ETH Zürich, CH-8092 Zürich, Switzerland

1. Overview

1.1. *Short summary*

Roughly speaking, clustering is a data analysis task to group a set of items into different categories so that items within one category are similar and items between different categories are dissimilar, where similar and dissimilar depend on the definition of distance between items. Although known for many decades, recently clustering has gained a lot of importance due to the exponential growth of digital libraries and the World Wide Web and the thus resulting need to find and extract information. Motivated by these Information Retrieval (IR) applications, which are usually characterized by large, sparse and high-dimensional data, “Introduction to Clustering Large and High-Dimensional Data” by J. Kogan is a textbook that tries to focus on a few clustering techniques that are very common in IR. In particular, it focuses on the k-means algorithm, which is by far the most popular one in IR, including many of its variations, among them incremental k-means, spherical k-means, quadratic k-means, k-means with divergences and others.

1.2. *Formal aspects*

The book is designed primarily for an audience of advanced undergraduate students or graduate students (in computer science or statistics), and it requires only minimal mathematical and programming prerequisites (in fact, some necessary linear algebra and optimization prerequisites are provided in the appendix). Most of the chapters contain a list of problems which vary in difficulty and a few programming projects which support the use of the book in a one-semester introductory course on clustering. Each chapter has a clear structure including numerical experiments of the particular algorithm discussed and is accompanied by a large number of figures to make the understanding easier. Mathematical notation is used when necessary, without overwhelming the reader. Finally, the further interested reader can find a large amount of bibliography given.

1.3. *Comparison with existing literature*

Machine learning/data mining textbooks usually contain a chapter about the standard k-means algorithm applied to the case when the distances between items are measured by Euclidean distances, see e.g. [1] or [2]. These books, however, cover a much wider range of clustering/learning techniques and therefore do not go into detail of k-means when similarities between items are measured by specific distances or “distance-like functions” as in this book. There are books which explain (usually on a higher, more intuitive level) the use of k-means in specific applications, see e.g. [3] for applications in bioinformatics or [4] for applications in quantitative psychology, but to the best of my knowledge, no (text)book exists covering details of variations of k-means algorithms as discussed in the present one.

2. Detailed summaries for each chapter

Chapter 1 is devoted to motivate the need of clustering in Information Retrieval applications. At first the task of clustering in this area is explained, and then possible

E-mail address: dmitsche@inf.ethz.ch.

1574-0137/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cosrev.2008.02.002

representations of the data to facilitate the clustering task are explained. In particular, possible representations of documents as vectors whose *j*-th coordinate consists of the number of occurrences of the *j*-th word (of a given dictionary) in that document are stressed. Next, the clustering techniques described in this book are outlined, and finally quite some bibliographical references for clustering in general are given.

Chapter 2 introduces the classical batch k-means algorithm: starting with an arbitrary *k*-partition[1] of a set of data (usually vectors in $\mathbb{R}^n$) one has to find *k* centroids for each of the *k* partition classes of the vectors: centroids are those points such that the sum of the distances of all vectors belonging to this partition class to the corresponding centroid is as small as possible. It is pointed out that in general (depending on the distance[2]) this requires solving a nontrivial optimization problem. Once having found these centroids, a new *k*-partition of the data set is defined by assigning each vector to the centroid closest to it, and the whole procedure iterates until the difference of the quality of two consecutive partitions (where quality is defined as the sum of the distances between all data vectors and its closest centroids) is below a certain tolerance threshold. The way the algorithm works when the distance is the squared Euclidean distance is illustrated with some figures. Also, the deficiencies of the algorithm (e.g. that in some cases it fails to produce partitions of "good" quality and that the right number of partition classes *k* has to be supplied initially) are presented. Motivated by these examples where the classical batch k-means algorithm fails the author then presents an incremental k-means algorithm which in addition to the classical algorithm contains a third step in each iteration which guarantees that the current solution is a "local maximum": a given set of centroids fulfils this condition, if there is no vector whose swap from one partition class to another one would give a solution of a better quality. Next, the author provides numerical experiments supporting evidence that this new algorithm performs substantially better than the classical one. Also, it is outlined that the new algorithm can be applied when the data are other geometrical objects than vectors in $\mathbb{R}^n$, e.g. lines. Since in general the optimal *k*-partition is not available, the author briefly explains how spectral methods can be used to derive lower bounds on the quality of the partitions.

In Chapter 3 the author extends the incremental k-means algorithm given in the previous chapter by a step to reduce the size of the dataset: since usually in IR applications the dataset is very large and does not fit into the available memory, there is a need to reduce the problem of clustering of the original dataset into a much smaller set which keeps most of the features of the original data set. This idea, so called BIRCH (Balanced Iterative Reducing and Clustering Algorithm), is explained in detail. The second part of this chapter then shows how this idea can be combined with the incremental k-means algorithm. Moreover, it is proven that the quality of the original partitions does not decrease in consecutive iterations when using BIRCH k-means.

The structure of Chapter 4 is similar to the one of Chapter 2: in this chapter the k-means algorithm with spherical distances (the distance between two vectors is measured by their scalar product) is presented. At first the classical spherical k-means algorithm is explained, together with an example that shows that it does not necessarily produce an optimal partition. Then it is shown, how an optimal two cluster partition on the unit circle can be found. The author also provides an example which points out that this task is not as straightforward as in the case of scalars (one-dimensional vectors) where the optimal two cluster partition is given by a vertical line with some scalars on its left and some on its right side. As in Chapter 2, a spherical incremental algorithm is then presented, and the author considers also some questions of computational complexity. Finally, the spherical k-means algorithm is compared with the k-means algorithm with quadratic distances, and it is explained that when constrained to the unit sphere, centroids for the quadratic and spherical k-means coincide.

In Chapter 5 the use of linear algebra techniques for clustering is presented. In particular, the idea of PDDP, Principal Direction Divisive Partitioning, is described: a given set of vectors is divided into two clusters according to the values of the projection of the data onto the line corresponding to the largest eigenvector of a suitably shifted covariance matrix of the original data. Moreover, the combination of the idea of PDDP with the spherical optimal two cluster partition on the circle of the previous chapter is discussed (sPDDP) — this corresponds to projecting the data onto the plane spanned by the two largest eigenvectors of the covariance matrix if the best two-dimensional approximation maximizes variance of the projections. Then numerical experiments are given for the different cases when applying PDDP only, applying sPDDP only, applying first PDDP and then the quadratic k-means algorithm as well as applying first sPDDP and then the quadratic k-means algorithm — it turns out that in any case a combination with k-means substantially improves the quality of the partitions, and in general sPDDP performs substantially better. Finally, the power method for computing the largest eigenvector together with an application of this procedure for computing the hub and authority value of a directed (web) graph is reviewed.

Chapter 6 deals with k-means clustering when the distance between data sets (usually probability distributions which are interpreted as unit vectors with all coordinates being nonnegative) has an information theoretic interpretation. In particular, for relative entropy, also called Kullback–Leibler divergence between two probability distributions, the classical batch k-means algorithm as well as the incremental k-means algorithm is discussed. As in Chapters 2 and 4, an example is given where the classical batch k-means algorithm does not produce an optimal partition. Also, as before, numerical experiments for this algorithm are given. In addition to this, in Chapter 6.4 a short overview of a "natural" distance

[1] In this review we also use the term "partition" instead of "clustering". What the author calls "cluster point", is here called "partition class".

[2] In this review, for simplicity, we always write distance for some functions which are only "distance-like", i.e. they do not satisfy all three properties of a distance metric. The author of the book, however, is here very careful and does not use the terminus "distance" for functions which do not fulfil all properties.

function between a pair of partitions, the so-called mutual information between two partitions, is given.

In Chapter 7 the focus is on the optimization step of the k-means algorithm to search a set of centroids. In general, this step leads to the problem of optimizing a nonsmooth objective function which is difficult to perform. Here a smooth approximation approach is given where the original objective function is replaced by a family of smooth functions depending on a scalar smoothing parameter. For a particular family of these functions, a smoothed k-means algorithm is presented, together with an example where this algorithm fails to produce an optimal partition. Moreover, it is proved that this family of smooth functions converges uniformly to the original objective function. A quite large survey of numerical experiments of this smoothed k-means algorithm compared with the incremental k-means algorithm is given. It turns out that while producing partitions of similar quality, the number of iterations needed in the smoothed version is significantly smaller.

The structure of Chapter 8 is similar to the one of Chapter 2 or 4: here the application of the k-means algorithm is discussed when the distance functions are Bregman distances or belong to the class of Ψ-divergence measures. It is shown exemplarily that for Bregman distances the computational cost of computing centroids in consecutive iterations is relatively cheap. Moreover, the author describes that for both distance functions the only knowledge needed for clustering is the clusters' centroids, its sizes and its qualities which makes a "BIRCH" type approach to k-means feasible and yields significant memory savings. Numerical experiments for vectors where different distance functions are combined (weighted squared Euclidean distance and Kullback–Leibler divergence) then give evidence that a combination of different k-means algorithms yields significantly improved results. Finally, it is shown how the smoothed k-means algorithm described in the previous chapter can also be combined with the techniques described here, and, as before, empirically it is observed that the quality of the resulting partitions is similar to a nonsmooth approach while running only a fraction of the iterations.

Chapter 9 considers in general the problem of assessing the quality of a clustering procedure: on the one hand, there are **internal criteria** to assess the quality of a clustering algorithm: these are known to the algorithm and turn the clustering into an optimization problem; one of them is exemplarily described here. On the other hand, **external criteria** are not known to the clustering algorithm and can be used for learning purposes only. It is assumed that the optimal partition is available, and the results of the clustering are compared with this partition. The author surveys several different external criteria, among them confusion matrices with suitably defined "misclassifications" and entropy measures.

The appendix consists of Chapter 10, which collects some background in linear algebra, Lagrange multipliers and convex analysis and of Chapter 11, which contains detailed master solutions to a selected list of problems given throughout the different chapters.

3. Typos, mistakes and suggestions

This section contains a small list of errors/suggestions found during reading.

- p. 6, when making the reference to Figure 1.1: "...a picture of two very different objects" (and not object).
- p. 17, statement of Problem 2.1.6: it should be $a_\ell \le c(\mathcal{A}) \le a_{\ell+1}$ (and not x_ℓ and $x_{\ell+1}$)
- p. 76, line 7: "a two-dimensional plain" should be "a two-dimensional plane"
- p. 79, p. 174: there is a typographical problem: instead of a line representing fractions only dots are printed.
- p. 112, second last line: "if x and y are two distinct limit points ...." (and not limit point)
- p. 127: in the statement of Problem 8.1.2 there are two mistakes: in 1) it should be $D_\Psi(x, y) = e^x - e^y - e^y(x - y)$, in 2) there is a log-term missing: it should read $D_\Psi(x, y) = x \log\left(\frac{x}{y}\right) + (1 - x)\log\left(\frac{1-x}{1-y}\right)$.

4. Conclusion

The textbook "Introduction to Clustering Large and High-Dimensional Data" by J. Kogan is a very good reference for an introductory course on clustering, in particular k-means clustering. The clear structure of each chapter and the large amount of figures make it easily accessible and a good choice for using it as a textbook for advanced undergraduate students. There are quite a lot of problems which require some thinking and can serve as good exercises for students (especially those whose master solution is given in the book), and a few ideas for programming projects support these problems. Admittedly, some of the problems presented are rather tedious calculation exercises (repeating variations of proofs given) and hardly illuminating, but they still can serve as good exercises for training students how to write proofs. The large amount of numerical experiments given in the book makes the book also a good reference for practitioners wanting to implement one specific variation of k-means discussed in this book.

REFERENCES

[1] T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley, New York, 2005.

[2] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC, 2005.

[3] C.A. Orengo, Bioinformatics: Genes, Proteins and Computers, Oxford University Press, 2003.

[4] M.J. Brusco, S. Stahl, Branch-and-bound Applications in Combinatorial Data Analysis, Springer, New York, 2005.