Introduction to Clustering Large and High-Dimensional Data *by Jacob Kogan.* Cambridge Univ Press 2007. $29.99. xiii+222 pp., paperback. ISBN 978-0-521-61793-2.

Daniel Boley.
University of Minnesota.

This book is devoted to clustering methods designed specifically for very large datasets in which the data items are represented by vectors of numerical attributes. All the methods are designed to take advantage of sparsity in the data and are particularly well suited for text data represented by a bag-of-words model. Only clustering, as an unupervised method, is discussed. Experimental results to illustrate the behavior of the methods is presented. One of the virtues is that all the methods are tried on the same dataset of text data so that one can compare the behavior of each method with another, though the dataset itself is decidedly small by the standards today, having 3891 documents and 4099 words.

Evaluating the performance quality of a clustering method is always problematic. Often it appears each author chooses his/her own method based on convenience and it is often difficult to compare the performance of a method in one paper with that in another paper, even if used on similar data, because the performance measures used are different. The book includes a description of a whole variety of measures, giving the reader an overview of the variety of methods one can use. In a separate chapter on evaluation, the author describes so-called internal criteria, depending only the same attributes used by the clustering algorithm itself, and so-called external criteria which essentially involves comparing the clustering results with a previously fixed gold standard labelling.

One internal criterion is the sum of all the distances between data samples within each individual cluster using an appropriate "distance" function. In the case of Euclidean distance or closely related distance measure, this is also the criterion that K-means tries to optimize. So one ends up with measuring performance with a criterion that the clustering algorithm itself is trying to optimize. A second criterion mentioned is the "silhouette coefficient" that has been commonly used in Statistics. This criterion has the advantage of comparing the distances among samples within each cluster against the smallest distance from each sample to a data item in a different cluster, and reducing this to a ratio in the interval [-1,1]. This gives some idea on how well the clusters are separated from each other.

When an external gold standard is available, then many methods exist to compare the computed results with the so-called "true" results. The book has a nice summary of the principal ones used in the literature.

- The confusion matrix (known among statisticians as a contingency table).
- The "entropy" (actually a relative entropy treating the gold standard as the "true" result).

- A purity measure (for each cluster, what label occurs the most, add up their respective portions, each weighted by the cluster size).
- From the Information Retrieval community: the precision, recall and F-measure. These measures are typically used to measure how well queries into a database return relevant documents, and is only indirectly useful in measuring the quality of a clustering of a whole data set.
- From Information theory, the mutual information. This is actually closely related to the entropy measure previously mentioned, though the mutual information is symmetric between the two labellings being compared (here the computed vs the gold standard).
- The Rand Statistic and Jaccard coefficients, which are both based on counting the number of pairs of data samples in which both the computed clustering and the gold standard are consistent in placing the the two samples together in the same cluster or not.

All the external methods are based simply on comparing cluster memberships, ignoring any intrinsic distance measure among the data items. Hence they can all be derived from the confusion matrix, which is then a very useful precursor to all the other measures. In the body of the book, the experimental results include the confusion matrix and some simple accuracy measures. Hence it would be theoretically possible for the reader to compute all these other external measures. By sticking to the confusion matrix, however, it is a lot easier to give the reader some intuition on how well the methods are doing. Whenever the performance of a clustering method is reduced to a single number, it is easy to see if one clustering method is doing better than another, but it hard to get a sense of whether the performance is good enough in an intuitive sense to capture the "essence" of the data well enough to be useful in an application.

The table of contents shows which clustering methods are discussed.

K-means - quadratic and spherical. The K-means method is by now a classical method which is very popular. It is a simple local optimization method, based on trying the reduce the distances among the data items within each cluster, which the author observes can get stuck at a local minimum. The author promotes the use of an "incremental" procedure whenever the standard k-means method has stabilized (every data sample is closer to its own center than to any other center and hence no data sample is moved). This consists of doing a "first variation" by attempting to move a data sample from one cluster to a neighboring cluster and seeing if the objective function is reduced. The author also observes that the k-means algorithm is also useful as a postprocessing step after the use of the other algorithms.

BIRCH (Balanced Iterative Reducing and Clustering) - including a k-means variant. BIRCH [3] is an incremental algorithm designed to produce a clustering based on one pass through the data in an incremental

manner. As each data item is processed, it is either assigned to an existing cluster or acts as a seed for a new cluster. At the end one has just the clusters and their centers, but not the original data.

PDDP (Principal Direction Divisive Partitioning) - original and a spherical version. PDDP [1] is a hierarchical top-down clustering method. The original method project all the data onto a single line chosen to maximize the spread (specifically the variance) among the projections, and then use this one-dimensional projection to split the dataset in two. This process is repeated recursively on each half, yielding a tree whose leaves are the clusters. For text datasets, the author observes that a spherical variant actually performs much better, and post-processing by K-means also improves the results.

The rest of the book is devoted to variations to these methods obtained by varying the measures of distance between the data samples, or other modifications to the objective function, to achieve either higher quality clusterings or faster convergence. The main items treated are

Smoothing. K-means is a simple local optimization method which can easily land at an undesireable local minimum, and a chapter is devoted to a smoothing method that is designed to promote convergence to the global minimum. Though the original idea comes from deterministic annealing [2], this is not the intuition used to derive the method in the book. Within the K-means algorithm, one must compute the minimum distance from each data sample to a cluster center. In smoothing methods, such a discrete minimum is replaced by a parametrized smooth function of these distances which converge to the discrete minimum as the parameter goes to infinity. For example, suppose $\mathbf{z} = (z_j)_{j=1}^k$ is a vector of distances from a given data sample to the j-th cluster center. Instead of computing $\min_j z_j = \max_j -z_j$, we compute

$$f(s) = -\log\left(\sum_j \exp(-z_j/s)\right)$$

and observe that

$$\lim_{s\to\infty} f(s) = \max_j(-z_j),$$

analogous to the definition of the l_∞ norm of a vector. For s small enough, $f(s)$ is convex leading to an easy optimization problem. The parameter s can then be gradually increased leading in the end to the "best" solution of the original minimization algorithm, though the details on how to choose s are left open.

Information Theoretic Measures. Two chapters are devoted to distance measures derived from Information theory: the Kullback-Liebler distance, and the more general Bregman divergences. These are generally much better suited to bag-of-words text data than using simple Euclidean distance

measures. Actually, the book discusses a special variant of the Bregman divergence, namely the Csiszar divergence. A Csiszar divergence starts with a base function $\phi(t)$ which is

- ϕ has two continuous derivatives on $(0, +\infty)$.
- ϕ is strictly convex on $(0, +\infty)$.
- $\lim_{t \to 0^+} \phi(t) = +\infty$.
- $\phi(1) = \phi'(1) = 0$ and $\phi''(1) > 0$.

Then a Csiszar divergence is the pseudo-distance function

$$d_\phi(s, t) = t\phi(s/t).$$

It turns out that $d_\phi(s, t)$ is convex in both s and t. If ϕ satisfies only the first two conditions above, then we have a Bregman divergence, in which case $d_\phi(s, t)$ is convex only in s. The Kullback-Liebler distance and many other information theoretic distances are special cases of Bregman or Csiszar divergences using a special choice for ϕ. For vectors, one applies these divergences elementwise and adds up the results. Even the Euclidean norm is a Bregman divergence, though in this case, the base function $\phi(t) = t^2$ has domain consisting of $(-\infty, +\infty)$. For a Kullback-Liebler-like distance, convex in both parameters, the base function is $\phi(t) = t \log t - t + 1$.

Extensive discussion is carried out on how to adapt the K-means algorithm to use these information theoretic measures. Even a BIRCH-like method is presented.

The book is a useful reference for the methods presented and has an extensive bibliography, including short informative bibliographic notes at the end of each chapter and pointers to many related methods not treated within this book. The methods are presented in clear and precise manner, though the reader must get used to some non-standard notation used in the algorithms and formulas, and in some cases must hunt around the book to find the definitions for some of the notation.

References

[1] D.L. Boley. Principal Direction Divisive Partitioning. *Data Mining and Knowledge Discovery*, 2:325–344, 1998.

[2] K. Rose, E. Gurwitz, and C.G.Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

[3] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. *J. of Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.