BEYOND THE CURSE OF MULTIDIMENSIONALITY: HIGH DIMENSIONAL CLUSTERING IN TEXT MINING

Simona Balbi[1]

Department of Mathematics and Statistics, University of Naples "Federico II", Naples, Italy

Abstract. *Data accumulate as time passes and there is a growing need for automated systems for partitioning data into groups, in order to describe, organise and retrieve information. With documental databases, one of the main aims is text categorisation, consisting of identifying documents with similar topics. In the usual vector space model, documents are represented as points in the high-dimensional space spanned by words. One obstacle to the efficient performance of algorithms is the curse of dimensionality: as dimensions increase, the space in which individuals are represented shrinks. Classical statistical methods lose their properties, and researchers' interest is devoted towards finding dense areas in lower dimensional spaces. The aim of this paper is to review the basic literature on the topic, focusing on dimensionality reduction and double clustering.*

Keywords: *Co-clustering, Dimensionality reduction, Distances, k-means algorithm, Text categorisation.*

1. INTRODUCTION

In statistical literature, we often refer to the curse of dimensionality: as dimensions increase, the space in which individuals are represented becomes smaller, and it is difficult to find interesting relations between variables. From a statistical perspective, classical methods lose their properties. It is also difficult to find clusters of individuals sharing common behaviour. From a computational viewpoint, increasing dimensionality makes existing algorithms inapplicable in real situations. Conversely, multidimensionality means richness of information, and developing methods for dealing with this situation can transform a problem into an opportunity.

The aim of this paper is to illustrate the problems related to one of the commonest tasks in text mining, i.e. text categorisation, focusing on clustering

[1] Simona Balbi, email: simona.balbi@unina.it

documents in high-dimensional spaces, in which the meaning of distance loses sense. After introducing the methodological framework in Sections 2 and 3, in Section 4 attention is addressed to the effectiveness of variuos techniques proposed in the literature for dimensionality reduction and double clustering for text categorisation tasks. A review of some common solutions and algorithms is presented. The paper ends by proposing some future research lines.

2. TEXT CATEGORISATION

The undertaking of finding documents which share common characteristics, usually in terms of a topic or a hierarchy of topics, originated in libraries. In the past, it was a human task, performed thanks to personal knowledge and abilities. The assigning of a new document (e.g., a book) to its category (its shelf) was a closely connected task, performed by librarians. With the diffusion of the World Wide Web, automatisation of both tasks became necessary.

The most common way of organising textual data is "bag-of-words" coding: documents are vectors in the high dimensional space spanned by words (vector space model).

Several research fields have been devoted to finding efficient solutions to classification problems, first of all pattern recognition and machine learning. Within this framework, the two tasks are called "unsupervised" and "supervised" classification, and are not only applied to documents but also to other objects, like images or sounds.

The main difference between the two approaches is the presence/absence of external information regarding the correct classification, on both previously known categories and the correspondence document/category. Given a training set, supervised classification procedures build the rules for assigning a new document (object) to a category. Two of the most commonly used approaches to supervised classification are Decision Trees, and Support Vector Machines (SVM), based on a kernel trick which transforms the input space into another higher-dimensional feature space, where boundary functions are easily identified.

Unsupervised classification algorithms do not need any background knowledge and search data to be partitioned in groups (clusters). The corresponding statistical methods are called "cluster analysis".

In the following, we review the basic concepts and some open questions in clustering high-dimensional data, such as textual databases.

3. CLUSTERING

Cluster analysis is one of the main classes of methods in multidimensional data analysis (see, e.g., Benzécri, 1973) and, more generally, in exploratory data analysis (Jain and Dubes, 1988). Finding groups in a set of objects answers a fundamental purpose in scientific thought. From the statistical perspective, cluster analysis methods enable good synthesis of available information by extracting an underlying structure.

The aim of clustering is to assign an object to a group (called cluster) whose other elements are in some senses more similar to each other than those belonging to other groups. In other words, we want to obtain groups as homogeneous as possible internally and as heterogeneous as possible externally, with respect to some measure. From a geometrical viewpoint, at the end of this process, a data point has a shorter distance to points within the cluster than those outside the cluster.

In low-dimensional spaces, two families of techniques are proposed, hierarchical and non-hierarchical. The latter produce a data set partition and the former a hierarchical structure of partitions, with a decreasing (/increasing) number of clusters. Statistical measures and graphical displays lead to the choice of the proper number of groups.

As hierarchical clustering techniques are not feasible in high dimensional spaces, due to their computational burden, we focus here on non-hierarchical methods.

3.1 BASIC CONCEPTS AND METHODOLOGICAL NOTES

For the sake of space, we can only make a brief survey of the ample literature devoted to cluster analysis (Everitt *et al.*, 2011). Here, our interest lies in the situations which characterise the finding of groups in high-dimensional data sets, typical of documental databases, where we may find thousands of documents and thousands of words.

3.2 THE *K- MEANS* ALGORITHM

In data mining and text mining, there is a simple algorithm which is often considered the parent of the most frequently proposed algorithms: the *k- means* algorithm (MacQueen, 1967), which partitions a set of n points lying in a d-dimensional space into k non-overlapping groups.

INPUT: (n, d) data matrix $\mathbf{X}$, number k of groups

STEP 1 – k individuals are randomly selected as starting centroids;

STEP 2 – the distances of all other ($n - k$) individuals from the k centroids are computed, and individuals are assigned to the nearest group;

STEP 3 – the centroids are recomputed for each group;

STEP 2 and STEP 3 are iterated until a stable solution is achieved, i.e., a further iteration leads to a (near) equal partition.

OUTPUT: k non-overlapping clusters.

In the basic formulation, distances are Euclidean distances. The Euclidean distance between point x_i and point x_l is given by:

$$d_2\left(x_i, x_l\right) = \sqrt{\sum_{j=1}^{d}\left(x_{ij} - x_{lj}\right)^2} \tag{1}$$

$x_i \in \Re^d$ and $x_l \in \Re^d$.
(1) is a special case of Minkowski's distance, with $p = 2$:

$$d_p\left(i,l\right) = \sqrt[p]{\sum_{j=1}^{d}\left|x_{ij} - x_{lj}\right|^p} \tag{2}$$

The *k-means* algorithm has a number of variations, for example, different ways of choosing the initial centroids. The success of the algorithm is due to its simplicity and speed, although it often converges to a local solution. However, some warnings on its performance have been made, even in low-dimensional spaces (Vattani, 2011).

One of the key points of this algorithm is choosing the number of clusters k, which is given as input.

3.3 NUMBER OF CLUSTERS

With real data, it is usually difficult to make reasonable hypotheses on the number of groups in a data set. In low-dimensional spaces, we can try inspecting data, but even on a plane it is often difficult and sometimes arbitrary to decide how many natural clusters there are. Graphical and computational methods (e.g., cross-validation) can be helpful in exploring a data set.

The choice of number of groups is closely connected with the admissibility of overlapping clusters. Clustering algorithms usually produce crisp groups, with the classical assumption that each individual belongs to one and only one cluster. Although there are a few algorithms admitting the possibility of not assigning an individual to a class (preferring a residual class with low inner homogeneity),

promising results have been obtained in the field of fuzzy reasoning, where overlapping solutions are accepted (Pal *et al*, 1996; Iezzi *et al*., 2012).

The choice of number of groups and the admissibility of overlapping are closely connected.

4. CLUSTERING HIGH-DIMENSIONAL DATA

In data mining, cluster analysis is used to identify non-trivial (i.e., useful and/or interesting) groups in a huge data set, and is also useful for organising data and retrieving information.

The curse of multidimensionality has some peculiar effects on clustering methods, i.e., on how distances (or dissimilarities) between points are computed. As pointed out by Steinbach *et al.* (2003), the behaviour of distances in high-dimensional data is critical.

Discussing the usefulness of the *nearest neighbour* algorithm, for certain distributions (e.g., when all the variables are i.i.d.), Beyer *et al.* (1999) showed that the relative difference of the distances of the closest (*MinDist*) and farthest (*MaxDist*) data points of an independently selected point tends to 0 as dimensionality increases:

$$\lim_{d\to\infty} \frac{MaxDist - MinDist}{MinDist} = 0 \tag{3}$$

In such cases, the notion of distance becomes meaningless. The result has been extended (Hinneburg *et al.*, 2000) to examine the absolute difference *MaxDist – MinDist*, by considering different metrics. In the case of Minkowski's distance in (3), it has been shown that, for d=1, *MaxDist – MinDist* increases with dimensionality, for d=2, *MaxDist – MinDist* remains relatively constant, and for $d\geq 3$, *MaxDist – MinDist* tends to 0 as dimensionality increases. Found for the *nearest neighbour* algorithm, the result reveals a general problem in clustering high-dimensional data, at least when the data distribution causes the usual distances between points to become relatively uniform.

4.1 THE *K-MEANS* ALGORITHM IN HIGH-DIMENSIONAL SPACES

The *k-means* algorithm aggregates points according to (1). This implicitly means assuming that the data lie in a Euclidean space. In order to overcome this limitation, the kernel-based clustering methods of Kim *et al.* (2005) are applied to compute distances by means of a non-linear kernel function, defined as:

$$d_2^k\left(x_i, x_l\right) = \sqrt{\sum_{j=1}^{d} k^2 x_{ij}^2 + k^2 x_{lj}^2 - 2k\left(x_{ij} - x_{lj}\right)^2} \quad (4)$$

In the previous notations $\kappa(.,.)$: $\Re^d \times \Re^d \rightarrow \Re$ is the kernel function, which is a *similarity function* having the property that, for any x_i, $K(x_i, x_i) = 1$, as the distance between x_i and x_j increases, $K(x_i, x_j)$ decreases. The kernel distance enables the clustering algorithm to capture the non-linear structure in the data. As the algorithm needs to store the (n, n) kernel matrix, alternative methods have been proposed (Chitta *et al.*, 2011) when dealing with very large data-bases.

Other approaches consider the problem from a different perspective, with the common objective of overcoming the bias induced by Euclidean metrics in clustering data by means of the *k-means* technique.

In data mining, multidimensional distributions are often used to describe and summarise various features of large datasets, and the problem of clustering is faced by measuring distances between distributions. Applegate *et al.* (2011) suggest an approximated algorithm for computing Wasserstein's distances (Earth Mover distances).

4.2 REDUCTION OF DIMENSIONALITY

Knowledge discovery in very large documental and numerical databases often includes procedures to reduce the dimensionality of the characteristics considered in the analysis.

An easy way of achieving this is "feature selection": external information is used in order to discard variables considered not relevant for the analysis. For example, in text categorisation, functional parts-of-speech, or terms belonging to a stop list are frequently discarded.

Another approach, known as "feature extraction", is also often adopted: data points are projected in a lower dimensional space. The basic idea is that these methods enable the structural information represented in relatively few dimensions to be preserved, removing noise.

Multidimensional data analysis techniques are often used for this task. Correspondence analysis (Lebart *et al.*, 1998) is one of the most common methods for analysing lexical tables, and for identifying and representing linguistic behaviour. Similarly, latent semantic analysis, as a development of Latent Semantic Indexing (Deerwester *et al*, 1990), achieves a reduction of dimensionality by building linear combinations of the variables. Both approaches rely on singular value decomposition, although with different choices in terms of centering, weightings and, mainly, distances: χ^2 for correspondence analysis and the usual Euclidean distances in

latent semantic analysis. Projection Pursuit is another multivariate statistical method useful for finding clusters, if it is assumed that the centroids of the clusters lie in the same unknown subspace (Huber, 1985).

In any case, some methodological statistical instruments developed within the framework of multivariate analysis may have problems when dealing with very large data sets, as in text mining.

The clustering problem must be seen from a different perspective, and many algorithms have been proposed on the assumption that clusters lie only on subsets of the full space.

4.3 CLUSTERING IN SUBSPACES

Clustering in subspaces may be viewed as a special kind of "local" feature selection. Algorithms usually refer to some kind of heuristic reasoning.

Showing the limitations of clustering in full dimensions, Agrawal *et al.* (1998) state that, in high-dimensional spaces, various points can be better clustered with respect to different subsets of dimensions. On the assumption that, if a *d*-dimensional space is dense, then a (*d*-1)-dimensional subspace is also dense ("downward closure"), they propose the CLIQUE algorithm, as follows:

STEP 1 – each dimension is partitioned into the same number of equal length intervals,

STEP 2 – a *d*-dimensional data space is partitioned into non-overlapping rectangular units,

STEP 3 – the subspaces containing dense units are identified,

STEP 4 – the extension of the subspaces representing the dense units is intersected to form a candidate search space in which dense units of higher dimensionality may exist,

STEP 5 – the dense units are then examined to determine clusters.

Units are obtained by the cross-product of *d* intervals, one for each dimension. A unit is considered "dense" if it contains a number of points greater than a predefined threshold. Both the number of intervals and the threshold are given as inputs.

It is interesting to note that CLIQUE is an effective way of finding regions of greater density in high-dimensional data, but it cannot discover clusters in their classical meaning, because the units are not partitioned into separate groups and overlaps are admitted. Some units may also belong to no class at all. This produces

some negative consequences, mainly when clusters must be interpreted.

In order to overcome the problem and obtain the usual clusters, Aggarwal *et al.* (1999) proposed the algorithm PROCLUS (PROjected CLUStering) as a different extension of "feature selection" methods. The starting points of PROCLUS are: clustering can refer to points or dimensions, and the *k-means* algorithm, in its variation known as *k-medoid* (Hand, 1981). As a consequence, the above authors using the Manhattan distance (d=1 in (2)) between points for finding neighbours. The output of PROCLUS is partitioning of the data points into clusters, together with the sets of dimensions on which the clusters were built. As a consequence, the interpretation and description of results usually cause no problems.

4.4 CO-CLUSTERING

The idea of double clustering (i.e., clustering variables and individuals) is not new in the statistical literature (and before with Bertin's graphical representations, de Falguerolle *et al*., 1997). Hartigan (1972) proposed what was later called a "sequential approach", which means clustering the rows and columns of starting matrix **X** successively and independently, in order to find clusters of similar elements possessing similar clusters of features. This is an *ante litteram* "local feature selection" method, and is considered the precursor of co-clustering.

Co-clustering (or biclustering, double clustering, two-way clustering) has been developed in various contexts, where the original matrix is not necessarily an (individuals $\times$ variables) one, and it is interesting to exploit the duality between rows and columns. In addition, as Dhillon *et al.* (2003) suggest, although we are interested in clustering along one way path of the data matrix, when dealing with sparse, high-dimensional data (as is always the case with word-document tables), co-clustering turns out to be useful.

Interesting surveys of the ample literature may be found in Van Mechelen *et al.* (2004) from a statistical perspective, and Berkhin (2006) from the data mining viewpoint.

The two most important fields for co-clustering are gene expression data analysis and text categorisation, when clusters of words rather than single words introduce content to the clustering of documents.

Unlike feature selection methods, co-clustering involves simultaneous clustering, and cannot be reduced to a simple concatenation or addition of constituent row-and-column clusters, and several simultaneous clustering methods imply the optimisation of an overall objective function. Simultaneous approaches also allow researchers to characterise the nature of the interaction or dependence structure between rows and columns, as implied by the data. As proposed by

Marcotorchino (1987), the problem is one of block-seriation and can be solved by integer linear programming, resulting in unique optimal solutions.

The literature contains three families of methods (van Mechelen *et al.*, 2004): a) row-and-column partitioning, b) nested row-and-column clustering, and c) overlapping row-and-column clustering. In addition, it distinguishes deterministic, stochastic and procedural approaches, according to the level of modelling and the criterion of optimisation.

Deterministic methods look for an optimal approximating block diagonal matrix, from permutation of the rows and columns of the starting matrix. In the special case of a (*I*, *J*) contingency tables C, Greenacre (1988) proposes maximising the association between the row and column classes, where the strength of association is captured by the classical χ^2 measure:

$$\chi^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(f_{ij} - f_{i.}f_{.j}\right)^2}{f_{i.}f_{.j}} \tag{5}$$

where f_{ij} is the relative frequency in the *ij* cell and $f_{i.}f_{.j}$ is the expected frequency under the independence hypothesis. The solution is obtained by computing block frequencies w_{rc}(r=1,…,R; c=1,…,C) for all possible data clusters, with the associated χ^2 value:

$$\chi^2 = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(w_{rc} - w_{r.}w_{.c}\right)^2}{w_{r.}w_{.c}} \tag{6}$$

The optimal partitioning maximises the χ^2 in (6).

Bock (2003a; 2003b) has shown that loss function (6) is a member of a broad family of loss functions involving a convex function of the centroids, and that various measures can be used instead of χ^2 as a criterion, for instance, the Kullback-Leibler information measure. Other measures have also been proposed. In text categorization, Goodman and Kruskal's τ_b in a genetic algorithm (Balbi *et al.*, 2010) has been proposed, in order to enhance the predictability of the task.

The choice of working on optimisation a loss function usually means reducing the response burden and, when compared with computing distances, is one of the main advantages of co-clustering.

5. CONCLUSIONS

The huge quantity of data now available in all fields makes it necessary to develop methods and techniques for mapping low-level data into more compact, more abstract and more useful forms. High-dimensional data clustering and related problems is a prolific research area in data mining and specifically in text mining, where text categorisation is one of the most important tasks, so that many algorithms and procedures have been proposed. Developing new methods and strategies for evaluating and comparing the performance of the various proposals can be a challenge for statisticians, to bring the problem to their own methodological perspective.

REFERENCES

Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S. and Park, J.S. (1999). Fast algorithms for projected clustering. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA: 61-72.

Agrawal, R., Gehrke, J., Gunopolos, D. and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seattle, WA: 94-105

Applegate, D., Dasu, T., Krishnan, S. and Urbanek, S. (2011). Unsupervised clustering of multidimensional distributions using Earth Mover Distance. In: *Proceedings KDD 2011* San Diego, CA: 636-644.

Balbi, S., Miele, R. and Scepi, G. (2010). Clustering of documents from a two-way viewpoint. In: Bolasco S., Chiari I. and Giuliano L. (Eds.), *Proceedings of JADT 2010*, LED, Roma: 27-36.

Benzécri, J. P. (1973). *L'analyse des données, 1. La taxinomie*. Dunod, Paris.

Berkhin, P. (2006). A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C. and Teboulle, M. (Eds.), *Grouping Multidimensional Data Recent Advances in Clustering*, Springer, Berlin-Heidelberg: 25-71.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1998). When is 'nearest neighbor' meaningful?. In: *Proceedings of 7th International Conference on Database Theory (ICDT-1999)*, Jerusalem: 217-235.

Bock, H.H. (2003a). Two-way clustering for contingency tables: maximizing a dependence measure. In: Schader, M, Gaul, W. and Vichi, M. (Eds.), *Between data science and applied data analysis. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Heidelberg: 143-154.

Bock, H.H. (2003b). Convexity based clustering criteria: theory, algorithm and applications in statistics. *Statistical Methods and Applications*, (12): 293-318.

Chitta, R., Jin, R., Havens, T.C. and Jain, A.K. (2011). Approximate kernel *k*-means : Solution to large scale kernel clustering. In: *Proceedings KDD 2011* San Diego, CA: 895-903.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the Society for Information Science*, (41): 391-407.

Dhillon, I.S., Mallela, S. and Modha, D.S. (2003). Information-theoretic co-clustering. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York: 89-98.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis*, 5th Edition, John Wiley & Sons, Chichester, UK.

de Falguerolle, A., Friedrich, F. and Sawitzki, G. (1997). A tribute to J.Bertin's graphical data analysis. In: Bandilla W. and Faulbaum F. (Eds.), *SoftStat '97, Advances in Statistical Software*, Lucius and Lucius, Stuttgart: 11-20.

Hand, D. (1981). *Order Statistics*, John Wiley and Sons, New York.

Hartigan, J.A. (1972). Direct clustering of a data matrix, *Journal of the American Statistical Association*, (67): 123-129.

Hinneburg, A., Aggarwal, C. and Keim, D.A. (2000). What is the nearest neighbor in high dimensional spaces? In: *Proceedings 26thConference on Very Large Data Bases*, Morgan Kaufmann: 505-515.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2): 435–475.

Iezzi, D.F. and Mastrangelo, M. (2012). Fuzzy c-means for web mining: The Italian tourist forum case. In: *Analysis and Modeling of complex data in behavioural and Social Science*. Anacapri (NA) - ITALY, September 3-4, 2012, CLEUP, PADOVA.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*, Prentice Hall, New Jersey.

Kim, D.W., Lee, K.Y., Lee D. and Lee K.H. (2005). Evaluation of the performance of clustering algorithms in kernel-induced feature space, *Pattern Recognition*, (38): 607–611.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*, Kluwer Academic Publisher, Boston.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press: 281–297.

Marcotorchino, F. (1987). Block seriation problems: A unified approach, *Applied Stochastic Models and Data Analysis*, (3): 73–91.

van Mechelen, I., Bock, H.H. and de Boeck, P. (2004). Two-mode clustering methods: A structured overview, *Statistical Methods in Medical Research*, (13): 363-394.

Pal, N.R., Bezdek, J.C. and Hathaway, R.J. (1996). Sequential Competitive learning and the fuzzy c-means clustering algorithm. *Neural Networks*, (5): 789-796.

Steinbach, M., Ertoz L. and Kumar, V. (2003). *The challenges of clustering high-dimensional data.* In: Wille, L.T. (Eds.) *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition*, Springer-Verlag: 273-307.

Vattani, A. (2011). k-means requires exponentially many iterations even in the plane, *Discrete & Computational Geometry*, (45): 596–616.