Universiteit
Leiden
The Netherlands

Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data

Kampert, M.M.D.

Citation
Kampert, M. M. D. (2019, July 3). *Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data*. Retrieved from https://hdl.handle.net/1887/74690

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/74690

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/74690 holds various files of this Leiden University dissertation.

Author: Kampert, M.M.D.
Title: Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data
Issue Date: 2019-07-03

Summary

The research in this monograph is focused on clustering of objects in high-dimensional data, given the assumption that the objects do not cluster on all the attributes, or not even on a single subset of attributes, but often on different subsets of attributes in the data. With the objective to reveal such a clustering structure, Friedman and Meulman (2004) proposed a framework and a specific algorithm, called COSA. Instead of producing clusters directly, COSA gives a representative distance matrix that can subsequently be analyzed by a variety of distance-based analysis methods, such as hierarchical clustering or multidimensional scaling. In this monograph we study the behavior of COSA to demonstrate its usefulness, but also a number of weaknesses, and we propose improvements to address the latter.

Chapter 1

The important notions of the context in which COSA is embedded are cluster analysis, regularized attribute weighting in high-dimensional data settings, and distances. To gain insight into data, and to be able to generate hypotheses, or detect anomalies and salient features one can perform a cluster analysis. Cluster analysis is the search for natural groupings in a data set of objects that are measured on attributes, while conducted in absence of the group labels of each object. When successful, these natural groupings are 'tightly' knit and preferably distinct from each other on the attributes.

The cluster analysis should involve an attribute weighting strategy, since in high-dimensional data settings it is assumed that there are many attributes irrelevant for the clustering of a group. Possibly, each cluster could have its own subset of relevant attributes. The main challenge for these attribute weighting strategies in cluster analysis is to produce stable results. By limiting the solution space for the values of the attribute weights, stability can be increased

For any type of cluster analysis, there is always a notion of (dis)similarity between the objects. In COSA, the representative distance function has the ability to reveal clusters that are formed in attribute subspaces of high dimensional datasets. Multi-dimensional scaling analysis (MDS) configurations, as well as the dendrogram that results from hierarchical clustering, are fruitful visualizations of the COSA distances with which these particular clustering structures can be revealed.

Chapter 2

The COSA algorithm that is the starting point in this monograph is the Friedman and Meulman algorithm that is based on a K nearest neighbors strategy, and will be referred to as COSA-KNN. The main purpose of COSA-KNN is to produce a matrix that contains the distances between N objects. A distance between object i and j, denoted by D_{ij}, is constructed from a weighted sum of P attribute distances. These P attribute distances, denoted by $\{d_{ijk}\}_{k=1}^{P}$, are in turn constructed from measurements collected in a $N \times P$ data set. The data set is assumed to have an underlying clustering structure in which the objects are clustered on cluster-specific subsets of attributes. However, COSA-KNN circumvents the need to specify the expected number of clusters in the data by the use of a K nearest neighbors method.

The attribute distances are weighted with a λ-regulated constraint on the negative entropy of the attribute weights. This particular constraint ensures that COSA-KNN can produce sensible solutions for the attribute weights of each cluster in high-dimensional data settings where $P \gg N$. The definition of the COSA distance matrix itself ensures that the distances for the object pairs within a cluster will always be smaller than the distances for the between-cluster object pairs. We will refer to this property of the distances as 'majorizing'.

The 'majorizing' distances are obtained from an iterative algorithm that heuristically minimizes a the COSA-KNN criterion. We start with an initial set of attribute weights from which a first distance matrix can be derived. Then, based on a K nearest neighbors method, new attribute weights are calculated, from which in turn a new distance matrix is derived, and so on. This iterative procedure is continued until a particular convergence criterion is reached. Within this iterative process COSA-KNN uses a strategy based on the homotopy between the criterion and an approximate criterion for COSA-KNN such that the algorithm can elegantly avoid inferior local minima.

Chapter 3

The choice of the values for two tuning parameters in COSA is crucial to detect a subtle clustering structure in the data. These two tuning parameter are λ and K. The definition of λ is related to the range of the values that defines a grouping of objects on an attribute, and it directly determines the weights of the attributes. The definition of K is the the size of the neighborhood for each object in a cluster.

The number of successful values of the tuning parameters can be increased when we apply a robust version of COSA-KNN. In the robust version of COSA-KNN, we use median-based estimates for the attribute weights, compared to the originally proposed mean-based estimates in COSA-KNN. The robust version of COSA-KNN is faster and more successful in revealing clustering structures in the data.

To automatically find successful combinations for the values of the tuning parameters, the so-called Gap statistic procedure is used (based on Tibshirani et al., 2001). The Gap statistic procedure is a permutation approach in which we select that particular combination of tuning parameters that provides the largest gap between

the value of the robust COSA-KNN criterion obtained for a particular data set, and (an approximation of) the expected value obtained for a null-reference model, i.e. a comparable data set with no clustering structure.

While the non-robust COSA-KNN criterion seems to be a concave smooth surface over a grid of values for both λ and K, the robust version of the criterion shows a zigzag pattern over the odd and even values for K. This particular zigzag pattern is also propogated in the Gap statistic values, leading towards the preference of even values for K, over odd values for K. Since the estimates based on even values for K are in general a little bit more stable, we suggest to use a grid with preferably only even values for K.

Chapter 4

We make COSA-KNN more powerful by implementing four improvements:

i. by a change of notation in the definition, we can generalize the clustering problem from Chapter 2, to settings where the assumption of mutually exclusive clusters does not need to hold;

ii. instead of letting λ regulate the negative entropy of the attribute weights for each neighborhood, we let λ regulate the Kullback-Leibler divergence between the attribute weights and a pre-specified set of attribute weights. This way, we can incorporate pre-specificied attribute weights that indicate the importance of each attribute in the clustering;

iii. we reformulate the criterion to allow for zero-valued attribute weights, also driven by λ;

iv. we change the COSA distance such that it is more succesful in separating pairs of objects from different clusters.

While COSA-KNN would not be able to detect clusters that mainly differ in their mean on attributes where the within-cluster variances are equal to those of the noise objects, these improvements make COSA-KNN to do so. Moreover, they make COSA-KNN more flexible and more powerful.

In COSA-KNN each object is assigned the same number of K nearest neighbors. Instead of setting the size of each neighborhood to a fixed value for K, we can let the size of each neighborhood be driven by the tuning parameter λ, making K superfluous. This approach and its associated algorithm will be called COSA-λNN.

COSA-λNN performs equally well as the improved version of COSA-KNN, if not better. The advantage of COSA-λNN over COSA-KNN is not only that each neighborhood can be of a different size, but it also comes with a large reduction in computing costs since the value for K does not need to be tuned anymore. While COSA-KNN needs tuning for all value combinations of λ and K, in COSA-λNN only the value for λ needs to be tuned.

Chapter 5

Although the main output of COSA is the distance matrix that we have labeled as "cluster-happy", it is not straightforward to extract L groups from this matrix. COSA by itself is not equipped to be compared directly with L-groups clustering algorithms. To extract L clusters from the COSA distances we propose to apply a new algorithm called MVPIN, since it uses a Minimum Variance strategy to Partition In Neighborhoods.

MVPIN applies a restricted form of Ward's method; the restriction is based on the popular Jarvis and Patrick (1973) shared nearest neighbors similarity. While using Ward's Minimum Variance method, clusters are only allowed to merge if the similarity measure based on the *shared nearest neighbors* between the two clusters is the highest among the similarities that pass a certain threshold.

When compared to competing clustering algorithms such as Partitioning Around Medoids (PAM; Kaufman & Rousseeuw, 1987), hierarchical clustering with average linkage, or Ward's method, a first examination shows that MVPIN is more successful at detecting the clustering structure for prototypical COSA data. In particular, MVPIN performs well in detecting small homogeneous clusters that are similar to each other, because of a similar clustering pattern on an overlapping subset of overlapping attributes. In combination with the Gap statistic procedure, MVPIN shows better overall results than those obtained by PAM and Ward's method on the (optimized) COSA-λNN distances for 12 benchmark data sets.

Chapter 6

Like some of the current 'state-of-the-art' L-groups clustering algorithms that include a form of regularized weighting of the attributes, COSA is largely motivated by the analysis of omics data. State-of-the-art L-groups clustering algorithms that were either inspired by, or compared themselves with COSA (Friedman & Meulman, 2004) are Entropy Weighted K-means clustering (EWKM; Jing, Ng, & Huang, 2007), Sparse clustering (SPARCL; Witten & Tibshirani, 2010), Simple Approach to Sparse clustering (SAS; Arias-Castro & Pu, 2017). When we replicate the results of these algorithms, and compare them with the results of MVPIN applied to the original COSA-KNN distances (Chapter 3) and COSA-λNN distances (Chapter 4), we find that COSA-λNN and SAS are the overall winners. The comparison is based on the analysis of 11 omics benchmark data sets.

Instead of framing COSA-λNN and the other algorithms as competitors of each other, we can also use them for validation of each others results. Given a clustering structure, we can compute an optimal COSA distance and optimal cluster attribute weights based on a self-learned cluster-specific λ_l. Then, the self-learned cluster attribute weights can be used together with attribute importance measures to filter for those attributes on which a cluster is homogeneous (the attribute on which the cluster has a low variance). The validation of the clusters based on the COSA framework can be formulated as a permutation test.

Chapter 7

This monograph ends with a discussion chapter in which a number of topics are being reviewed. These are the limitations in the monograph related to the choice of the simulation examples, the study of the computational costs of COSA-KNN and COSA-λNN, convergence properties, and how missing data are dealt with within the COSA framework. Furthermore, various aspects of possible future improvements and extensions are discussed. For possible improvements, we discuss a different regularization strategy for the weighting of the attributes, the COSA distances, and the distances at the attribute level. We also discuss how the COSA framework can be extended to a framework for Information Retrievial, Self-Organizing Maps (Kohonen, 1980) or Point of View Analysis (Tucker & Messick, 1963; Meulman & Verboon, 1993). These shortcomings, improvements, and extensions together serve as a plea for further study of COSA.