Identifying Single Clusters in Large Data Sets

Frank Klawonn*
Department of Computer Science
University of Applied Sciences Braunschweig/Wolfenbuettel
Salzdahlumer Str. 46/48
D-38302 Wolfenbuettel
Germany
voice: (+49)(5331) 939-6111
fax: (+49)(5331) 939-6002
email: f.klawonn@fh-wolfenbuettel.de

Olga Georgieva
Institute of Control and System Research
Bulgarian Academy of Sciences
P.O. Box 79, 1113 Sofia
Bulgaria
voice: (+359 2) 979 20 52
fax: (+359 2) 870 33 61
email: ogeorgieva@icsr.bas.bg

(* Corresponding author)

Identifying Single Clusters in Large Data Sets

Frank Klawonn, University of Applied Sciences Braunschweig/Wolfenbuettel, Germany

Olga Georgieva, Institute of Control and System Research - Bulgarian Academy of Sciences, Bulgaria

INTRODUCTION

Most clustering methods have to face the problem of characterizing good clusters among noise data. The arbitrary noise points that just do not belong to any class being searched for are of a real concern. The outliers or noise data points are data that severely deviate from the pattern set by the majority of the data, and rounding and grouping errors result from the inherent inaccuracy in the collection and recording of data. In fact, a single outlier can completely spoil the least squares (LS) estimate and thus the results of most LS based clustering techniques such as the hard C-means (HCM) and the fuzzy C-means algorithm (FCM) (Bezdek, Keller, Krisnapuram, & Pal, 1999).

For these reasons, a family of robust clustering techniques has emerged. There are two major families of robust clustering methods. The first includes techniques, which are directly based on robust statistics. The second family, assuming a known number of clusters, is based on modifying the objective function of FCM in order to make the parameter estimates more resistant to the data noise. Among them one promising approach is the noise clustering (NC) technique (Dave, 1991; Klawonn, 2004). It maintains the principle of probabilistic clustering, but an additional noise cluster is introduced. NC was developed and investigated in the context of a variety of objective function-based clustering algorithms and it has demonstrated its reliable ability to detect clusters amongst noise data.

BACKGROUND

Objective function-based clustering aims at minimizing an objective function that indicates a kind of fitting error of the determined clusters to the given data. In this objective function, the number of clusters has to be fixed in advance. However, as the number of clusters is usually unknown, an additional scheme has to be applied to determine the number of clusters (Guo, Chen, & Lyu, 2002; Tao, 2002). The parameters to be optimized are the membership degrees, that are values of belonging of each data point to every cluster, and the parameters, characterizing the cluster, which finally determine the distance values. In the simplest case, a single vector named cluster centre (prototype) represents each cluster. The distance of a data point to a cluster is simply the Euclidean distance between the cluster centre and the corresponding data point. More generally, one can use the squared inner-product distance norm, in which by a norm inducing symmetric and positive matrix different variances in the directions of the coordinate axes of the data space are accounted for. If the norm inducing matrix is the identity matrix we obtain the standard Euclidean distance that form spherical clusters. Clustering approaches that use more complex cluster prototypes than only the cluster centres, leading to adaptive distance measures, are for instance the Gustafson-Kessel (GK) algorithm (Gustafson, & Kessel, 1979), the volume adaptation strategy (Höppner, Klawonn, Kruse, & Runkler, 1999; Keller, & Klawonn, 2003) and the Gath-Geva (GG) algorithm (Gath, & Geva, 1989). The latter one is not a proper objective function algorithm, but corresponds to a fuzzified expectation maximization strategy. No matter which kind cluster prototype is used, the assignment of the data to the clusters is based on the corresponding distance measure. In hard clustering, a data object is assigned to the closest cluster, whereas in fuzzy clustering a membership degree i.e. a value that belongs to the interval [0,1] is computed. The highest membership degree of a data corresponds to the closest cluster.

Noise clustering has a benefit of the collection of the noise points in one single cluster. A virtual noise prototype with no parameters to be adjusted is introduced that has always the same distance to all points in the data set. The remaining clusters are assumed to be the good clusters in the data set. The objective function that considers the noise cluster is defined in the same manner as the general scheme for the clustering minimization functional. The main problem of NC is the proper choice of the noise distance. If it is set too small, then most of the points will get classified as noise points, while for a large noise distance most of the points will be classified into clusters other than the noise cluster. A right selection of the distance will result in a classification where the points that are actually close enough to the good clusters will get correctly classified into a good cluster, while the points that are away from the good clusters will get classified into the noise cluster. The detection of noise and outliers is a serious problem and has been addressed in various approaches (Leski, 2003; Yang, & Wang, 2004; Zhang, & Leung, 2003). However, all these methods need complex additional computations.

MAIN THIRST OF THE CHAPTER

The purpose of most clustering algorithms are to partition a given data set into clusters. However, in data mining tasks partitioning is not always the main goal. Finding interesting patterns or substructures in a large data set in the form of one or a few clusters that do not cover or partition the data set completely is an important issue in data mining. For this purpose a new clustering algorithm named **noise clustering with one good cluster** based on the noise clustering technique and able to detect single clusters step by step in a given data set has been recently developed (Georgieva, & Klawonn, 2004). In addition to identifying clusters step by step, as a side-effect noise data are detected automatically.

The algorithm assesses the dynamics of the number of points that are assigned to only one good cluster of the data set by slightly decreasing the noise distance. Starting with some large enough noise distance it is decreased with a prescribed decrement till the reasonable smallest distance is reached. The number of data belonging to the good cluster is calculated for every noise distance using the formula for the hard membership values or fuzzy membership values, respectively. Note that in this scheme only one cluster centre have to be computed, which in is case of hard noise clustering is the mean value of the good cluster data points and in the case of fuzzy noise clustering is the weighted average of all data points. It is obvious that by decreasing the noise distance a process of ‘loosing’ data, i.e. separating them to the noise cluster will begin. Continuing to decrease the noise distance, we will start to separate points from the good cluster and add them to the noise cluster. A further reduction of the noise distance will lead to a decreasing amount of data in the good cluster until the cluster will be entirely empty as all data will be assigned to the noise cluster. The described dynamics can be illustrated in a curve viewing the number of data points assigned to the good cluster over the noise distance. In this curve a plateau will indicate that we are in a phase of assigning proper noise data to the noise cluster, whereas a strong slope means that we actually loose data belonging to the good cluster to the noise cluster.

Generally, a number of clusters with different shapes and densities exists in large data sets and thus a complicated dynamics of the data assigned to the single cluster will be observed. However, the smooth part of the considered curve corresponds to the situation that a relatively small amount of data is removed, which is usually caused by loosing noise data. When we loose data from a good cluster (with higher density than the noise data), a small decrease of the noise distance will lead to a large amount of data we loose to the noise cluster, so that we will see a strong slope in our curve instead of a plateau. Thus, a strong slope indicates that at least one

cluster is just removed and separated from the (single) good cluster we try to find. By this way the algorithm determines the number of clusters and detects noise data. It does not depend on the initialisation, so that the danger of converging into local optima is further reduced compared to standard fuzzy clustering (Höppner, & Klawonn, 2003).

The described procedure is implemented in a cluster identification algorithm that assesses the dynamics of the quantity of the points assigned to the good cluster or equivalently assigned to the noise cluster through the slight decrease of the noise distance. By detecting the strong slopes the algorithm separates one cluster at every algorithm pass. A significant reduction of the noise is achieved even in the first algorithm pass. The clustering procedure is repeated by proceeding with a smaller data set as the original one is reduced by the identified noise data and data belonging to the already identified cluster(s).

The curve that is determined by the dynamics of the data assigned to the noise cluster is smoother in case of fuzzy noise clustering compared to the hard clustering case due to the fuzzily defined membership values. Also local minima of this curve could be observed due to the given freedom of the points to belong to both good and noise cluster simultaneously. Fuzzy clustering can deal better with complex data sets than hard clustering due to the given relative degree of membership of a point to the good cluster. However, for the same reason the amount of the identified noise points is less than in the hard clustering case.

FUTURE TRENDS

Whereas the standard clustering partitions the whole data set, the main goal of the noise clustering with one good cluster is to identify single clusters even in the case when a large part of the data does not have any kind of group structure at all. This will have a large benefit in some application areas of cluster analysis like, for instance, gene expression data and astrophysics data

analysis, where the ultimate goal is not to partition the data, but to find some well-defined clusters that only cover a small fraction of the data. By removing a large amount of the noise data the obtained clusters are used to find some interesting substructures in the data set.

One future improvement will lead to an extended procedure that first finds the location of the centres of the interesting clusters using the standard Euclidean distance. Then, the algorithm could be started again with this cluster centres but using more sophisticated distance measures as GK or volume adaptation strategy that can better adapt to the shape of the identified cluster. For extremely large data sets the algorithm can be combined with speed-up techniques as the one proposed by Höppner (2002).

Another further application consists in incorporation of the proposed algorithm as an initialization step for other more sophisticated clustering algorithm providing the number of clusters and their approximate location.

CONCLUSION

A clustering algorithm based on the principles of noise clustering enable to find good clusters with variable shape and density step by step is proposed. It can be applied to find just a few substructures, but also as an iterative method to partition a data set including the identification of the number of clusters and noise data. The algorithm is applicable in terms of both hard and fuzzy clustering. It automatically determines the number of clusters, while in the standard objective function-based clustering additional strategies have to be applied in order to define the number of clusters.

The identification algorithm is independent of the dimensionality of the considered problem. It does not require comprehensive or costly computations and the computation time increases linearly only with the number of data points.

REFERENCES

Bezdek, J.C., Keller, J., Krisnapuram, R., & Pal, N.R. (1999). *Fuzzy models and algorithms for pattern recognition and image processing*. Kluwer Academic Publishers.

Dave, R.N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters,* 12, 657-664.

Gath, I., & Geva, A.B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 7, 773-781.

Georgieva, O., & Klawonn, F. (2004). A cluster identification algorithm based on noise clustering. *Transactions on SystemsMan and Cybernetics – Part B* (submitted).

Guo, P., Chen, C.L.P., & Lyu, M.R. (2002). Cluster number selection for a small set of samples using the Bayesian Ying-Yang model. *IEEE Transactions on Neural Networks,* 13, 757-763.

Gustafson, D., & Kessel, W. (1979). Fuzzy clustering with a fuzzy covariance matrix. *Advances in fuzzy set theory and applications*, North-Holland, 605-620.

Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy Cluster Analysis*. John Wiley & Sons, Chichester.

Höppner, F. (2002). Speeding up fuzzy c-means: Using a hierarchical data organisation to control the precision of membership calculation. *Fuzzy Sets and Systems,* 128(3), 365-378.

Höppner, F., & Klawonn, F. (2003). A contribution to convergence theory of fuzzy c-means and derivatives. *IEEE Transactions on Fuzzy Systems,* 11, 682-694.

Keller, A., & Klawonn, F. (2003). Adaptation of cluster sizes in objective function based fuzzy clustering. In: C.T. Leondes (ed.): *Intelligent Systems: Technology and Applications. IV: Database and Learning Systems*. CRC Press, Boca Raton, 181-199.

Klawonn, F. (2004). Noise clustering with a fixed fraction of noise. In: A. Lotfi, J.M. Garibaldi: *Applications and science in soft computing*. Springer, Berlin, 133-138.

Leski, J.M. (2003). Generalized weighted conditional fuzzy clustering. *IEEE Transactions on Fuzzy Systems,* 11, 709-715.

Tao, C.W. (2002). Unsupervised fuzzy clustering with multi-center clusters. *Fuzzy Sets and Systems 128*, 305-322.

Yang, T.-N., & Wang, S.-D. (2004). Competitive algorithms for the clustering of noisy data. *Fuzzy Sets and Systems*, 141, 281-299.

Zhang, J.-S., & Leung, Y.-W. (20003). Robust clustering by pruning outliers. *IEEE Transactions on Systems, Man and Cybernetics – Part B*, 33, 983-999.

TERMS AND THEIR DEFINITION

Cluster analysis: Partition a given data set into clusters where data assigned to the same cluster should be similar, whereas data from different clusters should be dissimilar.

Objective function-based clustering: Cluster analysis is carried on minimizing an objective function that indicates a fitting error of the clusters to the data.

Hard clustering: Cluster analysis where each data object is assigned to a unique cluster.

Fuzzy clustering: Cluster analysis where a data object can have membership degrees to different clusters. Usually it is assumed that the membership degrees of a data object to all clusters sum up

to one, so that a membership degree can also be interpreted as the probability the data object belongs to the corresponding cluster.

Noise clustering: An additional noise cluster is induced in objective function-based clustering to collect the noise data or outliers. All data objects are assumed to have a fixed (large) distance to the noise cluster, so that only data far away from all other clusters will be assigned to the cluster.

Robust clustering: Refers to clustering techniques that behave robust w.r.t. noise, i.e. adding some noise to the data in the form changing the values of the data objects slightly as well as adding some outliers will not drastically influence the clustering result.

Cluster centres (prototypes): Clusters in objective function-based clustering are represented by prototypes that define how the distance of a data object to the corresponding cluster is computed. In the simplest case a single vector represents the cluster, and the distance to the cluster is the Euclidean distance between cluster centre and data object.