Mining for High Dimensional Clusters using Projections and Visualizations

Alexander Hinneburg
Institute of Computer Science, University of Halle
Kurt-Mothes-Str.1, 06120 Halle (Saale), Germany
Tel. (+49) 345 5524737, Fax (+49) 345 5527009
hinneburg@informatik.uni-halle.de

Abstract

Many applications require the clustering of large amounts of high dimensional data. Most automated clustering techniques, however, do not work effectively and/or efficiently on high dimensional (numerical) data, i.e. they are likely to miss clusters with certain unexpected characteristics. There are various reasons for this. First, it is difficult to find the necessary parameters for tuning the clustering algorithms to the specific applications characteristics. Second, it is hard to verify and interpret the resulting high dimensional clusters and third, often the concept of clusters inspired from low dimensional cases cannot be extended to high dimensional cases. A desired clustering method should be able to deal with the inherent sparsity of high dimensional spaces, which is due to the so-called "curse of dimensionality".

Existing approaches which partition the data corresponding to occurring clusters, can be categorized as follows: model and optimization based, density based and hybrid techniques [16, 17]. Hierarchical methods are not mentioned explicitly, because mostly all approaches can be extended to a hierarchical version.

Examples for model and optimization based approaches are k-means(or LBG) [8], CLARANS [19], Kohonen feature maps [18], (Growing) Neural Gas [6] and Growing Cell Structures [7]. All these algorithms can be run in batch or online mode [20] and try to adopt iteratively an assumed model to a given data set. The advantage of the algorithms is their low complexity, which is linear in the number of data points, dimensions, prototypes and iterations, making them suitable for large data sets. However, it is hard to verify, whether the used model is appropriate for the given data, especially if the data have high dimensionality and contain a significant amount of noise [15]. It is also difficult to tune necessary parameters such as number of prototypes, adaption and growth rate and to find a good abortion criteria

to stop the iterative optimization.

The second broad category are density based approaches. The theory for these methods discerns linkage based methods (single, average, centroid and complete–linkage) [4] and kernel density estimation [22, 21]. Recent examples for linkage based methods are BIRCH (Phase 1-2) [25], DBSCAN [5], DBCLASD [24], STING [23] and OPTICS [3]. Another recent clustering algorithm in this category is DENCLUE [9], but this is based on kernel density estimation. Kernel density estimation provides a general theory for the named algorithms which can be simulated by using specific kernels. However, it is known from the theory of kernel density estimation that a density estimate based on an arbitrary kernel becomes insignificant when the dimensionality of the data grows [22, 21]. So such algorithms are not able to deal with the inherent sparsity of high dimensional feature spaces.

Hybrid methods provide a framework to apply different clustering methods on a data set to combine the advantages of the used methods. Such algorithms are BIRCH (Phase 3-4), CLIQUE [2] and OptiGrid [10]. The concept of BIRCH is (1) to compress the data by using a fast, but sequence dependent centroid–linkage algorithm and (2) to apply another arbitrary cluster algorithm to the compressed representation of the data set to correct some artifacts from the compression phase. However, BIRCH fails to derive a good compressed representation in cases of high dimensional noisy data. In that cases it compresses the whole data set to one data item. An experimental evaluation of that behavior can be found in [10]. The CLIQUE algorithm constructs a subset of axes parallel linear projections (subspaces) which contains clusters. The projection construction works bottom up. First it starts to search one dimensional projections for good clusterings. In a second step it evaluates all combinations of the best found one dimensional projections to construct two dimensional subspaces. High dimensional projections are constructed by the same way. The hybrid structure comes from the fact, that for the evaluation process any density based algorithm can be applied. However, the bottom up concept falls short for larger dimensionality because the number of combinations of projections can grow very large.

In my PhD thesis I want to develop a system for exploring and clustering high dimensional data, which combines the advantages of an advanced clustering algorithm with novel visual mining techniques. The new clustering algorithm should be able to deal with the named problems. In recent papers we published first results [10, 11].

The new idea for high dimensional clustering is to use projections to specify separators which separate one or more clusters. A separator can be a simple hyperplane, a combination of multiples, can be based on center points using new metrics or the result of any cluster algorithm applied to the projected data. The advantage of the use of projections – instead of the full high dimensional space – is that first, projections of the data often contain a lot of information about the structure and are smaller than the original; second, lower dimensional projection of the data are not so sparse like the original space and traditional concepts of data exploration like kernel density estimation, histograms, data compression for instance with k-means are working more effectively. This leads to the concept of projected clustering, which extends the traditional cluster concept. It bridges a gap between projection pursuit [13] and clustering. In our group we also explore and extend the concept of proximity based on metrics in high dimensional spaces [14, 1, 12], which lead to a new understanding of high dimensional clusters.

There are several problems in the pursued approach, namely to find meaningful projections of the data, to determine the quality of a projection, to specify meaningful separators. These problems can be described as an optimization problem. Since the optimization criteria often contains application dependent factors we use the concept of an visual and interactive optimizer. This allows the user to influence the optimizer with application dependent knowledge. The system will be tested on various real data applications.

References

[1] Aggarwal C., Keim D. A., Hinneburg A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space, Submitted for publication.

[2] Agrawal R., Gehrke J. E., Gunopulos D. and Raghavan P.: Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications., In Proceedings of the 1998 SIGMOD Conference, Seattle, Washington, 1998.

[3] Ankerst M., Breunig M. M., Kriegel H.-P., Sander J.: OPTICS: Ordering Points To Identify the Clustering Structure, Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA, 1999, pp. 49-60.

[4] Bock H.H.: Automatic Classification, *Vandenhoeck and Ruprecht*, Göttingen, 1974.

[5] Ester M., Kriegel H.-P., Sander J., Xu X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc.

2nd Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press, 1996.

[6] Fritzke B.: A Growing Neural Gas Network Learns Topologies, in Tesauro G., Touretzky D.S. and Leen T.K.(eds.) Advances in Neural Information Processing Systems 7, *MIT Press* MA, 1995.

[7] Fritzke B.: Growing cell structures - a self-organizing network for unsupervised and supervised learning., *Neural Networks*, Vol. 7, No. 9, page 1441-1460, 1994.

[8] Gersho A. and Gray R.M.: Vector Quantization and Signal Compression, *Kluwer Academic Publishers*, 1992.

[9] Hinneburg A. and Keim D.A.: An Efficient Approach to Clustering in Multimedia Databases with Noise, Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York,*AAAI Press*, 1998.

[10] Hinneburg A. and Keim D. A.: Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering, Int. Conf. on Very Large Databases (VLDB'99), Edinburgh, UK, 1999.

[11] Hinneburg A., Wawryniuk M. and Keim D. A.: HD-Eye D. A.: Visual Mining of High-Dimensional Data, Computer Graphics & Applications Journal, Sept. 1999.

[12] Hinneburg A., Aggarwal C. and Keim D. A.: What is the nearest neighbor in high dimensional spaces, Submitted for publication.

[13] Huber P.J.: Projection Pursuit (with Discussion), *Ann. Statist.*, 13, pp. 435–525, 1985.

[14] Heczko M., Hinneburg A. and Keim D. A.: Multiresolution Similarity Search in Image Databases, Submitted for publication.

[15] Hinneburg A. and Keim D. A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise, submitted for publication.

[16] Keim D.A. and Hinneburg A.:(Tutorial) Clustering Methods for Large Databases: From the Past to the Future. SIGMOD Conference, Philadelphia, 1999: 509.

[17] Keim D.A. and Hinneburg A.:(Tutorial) Clustering Techniques for Large Data Sets: From the Past to the Future. SIGKDD Conference, San Diego, 1999: 435.

[18] Kohonen T., Mäkisara K., Simula O., Kangas J.: Artificial Networks, Amsterdam 1991.

[19] Ng R.T., Han J.: Efficient and Effective Clustering Methods for Spatial Data Mining, Int. Conf. on Very Large Databases (VLDB'94), pp. 144-155, 1994.

[20] Rojas R.,: Neural Networks – A systematic Introduction, *Springer*, Berlin, 1994.

[21] Scott D.W.: Multivariate Density Estimation, *Wiley and Sons*, 1992.

[22] Silverman B.W.: Density Estimation for Statistics and Data Analysis, *Chapman and Hall*, 1986.

[23] Wang W., Yang J., Muntz R.: STING: A Statistical Information Grid Approach to Spatial Data Mining, Proc. 23rd Int. Conf. on Very Large Data Bases,*Morgan Kaufmann*, pp.186-195, 1997.

[24] Xu X., Ester M., Kriegel H.-P., Sander J.: A Nonparametric Clustering Algorithm for Knowlege Discovery in Large Spatial Databases, Proc. IEEE Int. Conf. on Data Engineering,*IEEE Computer Society Press*, 1998.

[25] Zhang T., Ramakrishnan R., Linvy M.: BIRCH: An Efficient Data Clustering Method for very Large Databases, Proc. ACM SIGMOD Int. Conf. on Management of Data, ACM Press, 1996, pp. 103-114.