

Nature Inspired Methods for Local Pattern Detection

Ira Assent, Ralph Krieger, Thomas Seidl
Department for Computer Science 9
RWTH Aachen University

Andreas Steffens
MIT GmbH
Aachen

Subspace clustering

Data produced in application domains like life sciences, meteorology, engineering, telecommunication, commerce and multimedia entertainment is rapidly growing, thus increasing the demand for data mining techniques which help users keep track of the information hidden in the data. Many applications ask for incoming data to be grouped according to some predefined criteria. Examples include customer shopping behavior and interests in a group of products or image segmentation for object recognition. For clustering, numerous techniques in data mining have been proposed and employed successfully in various application domains including k-mean [18], EM [10], DBSCAN [12] and OPTICS [2].

However, clustering is difficult in the presence of noise. Additionally, an effect dubbed "curse of dimensionality" hinders cluster detection [7]. The more attributes are recorded, the higher is the number of possible attribute value combinations. This leads to more and more similar distances between individual instances in sparsely populated high dimensional applications. Consequently, no meaningful clusters exist in all attributes of high-dimensional domains.

This has led to the development of subspace clustering research. The general idea is to locally project the instances into subspaces of the attributes to rediscover locally relevant patterns as performed by CLIQUE [1], ENCLUS [9], SUBCLU [15] and CLICKS [26]. By masking irrelevant or noisy attributes, meaningful clusters can be found. Our group studied the effects of subspace clustering in various applications, like medical image processing [17, 25] molecular classification [3] and hydrological applications [23].

Recent work on subspace clustering addresses new interestingness measures for subspace clusters. Thereby, in a statistical manner, a cluster is interesting if it is more dense than expected. One of the open problems still remaining is that the search for interesting subspace clusters is a tremendously complex task. It is typically not possible to evaluate all possible subspaces (i.e. the entire powerset) of high-dimensional datasets, nor is it feasible to consider all possible attribute value divisions. To remedy this, some heuristics have been proposed and evaluated (RIS [16], SURFING [5], SHISM [24], Classic [4]) but the quality of the heuristics still demands for improvement.

High-dimensional index structures have often been used to support clustering algorithms in finding the neighborhood of an object [6, 11]. Depending on the application specialized search algorithms [21, 22] and techniques for dimensionality reduction [8] are helpful. For subspace clustering methods appropriate index structures and search algorithms are still an open problem. Nature inspired search algorithms are also an interesting approach for evaluating the different kind of queries performed by subspace clustering algorithms.

Nature inspired solutions for subspace clustering

Nature inspired algorithms have already been used to tackle the task of detecting local patterns [20]. However many open questions still remain in this field of research and many nature inspired approaches may yield good solutions for this complex task.

The different subspace projections form a lattice. The rate of local patterns contained in a subspace can be interpreted as the gravity of a subspace. Thus one possibility to identify interesting subspaces is to apply gravity based learning models [19].

We propose to follow another nature inspired analogy which has been successfully employed for computationally complex optimizations such as efficient routing or wireless network optimization. Evolutionary algorithms mimic the natural evolution of specialized species which are well-adapted to their individual natural environments. In contrast to traditional subspace search methods, evolutionary approaches allow eventual prevalence of seemingly non-promising subspace combinations. Mutation can prevent locally optimal solutions. Moreover, less parameters have to be determined in advance, as long as gene diversity is ensured.

Populations evolve across several generations where individual fitness according to the respective local environment decisively influences the likelihood of reproduction. Diversity of genetic possibilities is ensured via sexual crossing and mutation of individual genes or crossing over of sequences of genes. While evolution is not directed, competition among individuals ensures survival of the fittest.

Taking an evolutionary perspective, subspace search is then the natural selection or final prevalence of those subspaces which are the fittest with respect to their cluster potential.

For evolutionary algorithm aiming at letting naturally evolve these subspaces best suited for clustering in high dimensional settings the following aspects have to be evaluated:

- What is the population's gene pool and how is the population initialized? We propose to admit the entire powerset as the population's gene pool, thus avoiding any bias in subspace selection. Random initialization will be compared with a priori knowledge on attributes to create individuals, and with populations which are constrained to cover all attributes. Moreover, we will study the size and dimensionality of the population.
- How is fitness determined? The fittest subspace combinations are those with the highest potential for containing meaningful patterns or clusters. We will thus measure the homogeneity of patterns contained in subspace combinations [4].
- Which individuals are admitted for reproduction? Our approach is to admit the fittest individuals. Additionally, to ensure that the gene pool's diversity is kept up, less promising individuals are randomly crossed in.
- How can genetic information from parents be passed on to children? We propose to represent genetic information of an individual as the binary encoding of the corresponding subspace and to mimic natural inheritance with recombination (cross-over) and mutation by using a probabilistic model.
- How can associations between genes be detected or measured? In evolving biological systems an interesting phenomenon is discovered, genes and genetic information are associated or linked to each other. The understanding of such a linkage leads to deeper understanding of subspaces in our context. The term coevolution indicates the parallel development of two species in evolution theory. This effect is produced by mutual interaction, which increases the fitness of both species. Therefore an interesting question in subspace search is: How can coevolution of genes be supported in our context?

Stable populations across several generations of subspaces indicate that the algorithm has indeed detected those subspaces most promising for subspace cluster identification.

Evaluation

We plan to evaluate our algorithm in a variety of settings, determining the influence of population size, initialization, and variations in reproduction and mutation strategies. Comparison with a brute force approach is generally not feasible, thus we will compare our evolutionary approach with well-established subspace detection algorithms such as [24, 16]. Subspace clustering can be seen as a pre-processing step in locally adaptive classification, i.e. the automatic filing of incoming, previously unseen, instances according to patterns learned from historic data [4]. By including this application, we are able to determine the classification accuracy (as a percentage of classified objects) of our approach as opposed to traditional methods. To measure the quality of subspace detection, we plan to generate synthetic data containing local patterns in subspaces and determine the improvement factor of evolutionary algorithms [14, 13].

Literatur

- [1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 94–105, 1998.
- [2] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 49–60, 1999.
- [3] M. Ankerst, G. Kastenmüller, Kriegel H.-P., and Seidl T. Nearest neighbor classification in 3d protein databases. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 34–43, 1999.
- [4] I. Assent, R. Krieger, P. Welter, T. Seidl, and J. Herbers. Subspace classification: Local pattern analysis for classification. *Submitted for publication*.
- [5] C. Baumgartner, C. Plant, K. Kailing, H.-P. Kriegel, and P. Kröger. Subspace selection for clustering high-dimensional data. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM)*, pages 11–18, 2004.
- [6] S. Berchtold, D.A. Keim, and H.-P. Kriegel. The X-tree: An index structure for high-dimensional data. In *Proceedings of the 22nd International Conference on Very Large Databases (VLDB)*, pages 28–39, 1996.
- [7] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, pages 217–235. Springer, 1999.
- [8] C. Brochhaus, M. Wichterich, and T. Seidl. Approximation techniques to enable dimensionality reduction for voronoi-based nearest neighbor search. In *Proceedings of the 10th International Conference on Extending Database Technology (EDBT)*, pages 204–221, 2006.
- [9] C.H. Cheng, W.-C.A. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM International Conference on Knowledge discovery and data mining (SIGKDD)*, pages 84–93, 1999.
- [10] A.P. Dempster, N. M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *J. of the Royal Statistical Society*, pages 1–38, 1977.
- [11] J. Enderle, M. Hampel, and T. Seidl. Joining interval data in relational databases. In *Proceedings of the ACM international conference on Management of data (SIGMOD)*, pages 683–694, 2004.
- [12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.
- [13] A.A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer, 2002.
- [14] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.
- [15] K. Kailing, H.-P. Kriegel, and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 246–257, 2004.
- [16] K. Kailing, H.P. Kriegel, P. Kröger, and S. Wanka. Ranking interesting subspaces for clustering high dimensional data. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 241–252, 2003.

- [17] T. M. Lehmann, D. Beier, and T. Seidl C. Thies and. Segmentation of medical images combining local, regional, global, and hierarchical distances into a bottom-up region merging scheme. In *Proceedings of the Int. Conference on Image Processing, part of the Symposium on Medical Imaging (SPIE)*, pages 546–555, 2005.
- [18] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symp. Math. stat. & prob.*, pages 281–297, 1967.
- [19] D. Ruta and B. Gabrys. Nature-inspired learning models. In *Proceedings of Nature-inspired Smart Information Systems Symposium (NiSIS)*, 2005.
- [20] I. A. Sarafis, P.W. Trinder, and A.M.S. Zalzal. Towards effective subspace clustering with an evolutionary algorithm. In *Proceedings of the Congress on Evolutionary Computation (CEC)*, pages 797–806, 2003.
- [21] T. Seidl. Nearest neighbor search on multimedia indexing structures. In *Invited Tutorial on 1st Int. Workshop on Computer Vision meets Databases (CVDB) in cooperation with ACM SIGMOD*, 2004.
- [22] T. Seidl and H.-P. Kriegel. Adaptable similarity search in large image databases. In *State-of-the-Art in Content-Based Image and Video Retrieval*, pages 297–317, 2001.
- [23] T. Seidl, R. Krieger, I. Assent, B. Glavic, and H. Nacken. Data mining zur entscheidungsunterstützung in der hydrologie. In *Tag der Hydrologie*, pages 137–145, 2005.
- [24] K. Sequeira and M.J. Zaki. SCHISM: A new approach for interesting subspace mining. In *Proceedings of the International Conference on Data Mining (ICDM)*, pages 186–193, 2004.
- [25] C. Thies, M. Schmitt Borreda, T. Lehmann, and T. Seidl. A classification framework for content-based extraction of biomedical objects from hierarchically decomposed images. In *Proceedings of the Int. Conference on Image Processing, part of the Symposium on Medical Imaging (SPIE)*, 2006.
- [26] M.J. Zaki, M. Peters, I. Assent, and T. Seidl. Clicks: an effective algorithm for mining subspace clusters in categorical datasets. In *Proceedings of the ACM International Conference on Knowledge discovery and data mining (SIGKDD)*, pages 736–742, 2005.