



Final Report on Task Force

Nature inspired Methods for Local Pattern Detection

TASK FORCE Activity responsible: Thomas Seidl

Focus group: Nature-inspired Data Technology (NiDT)

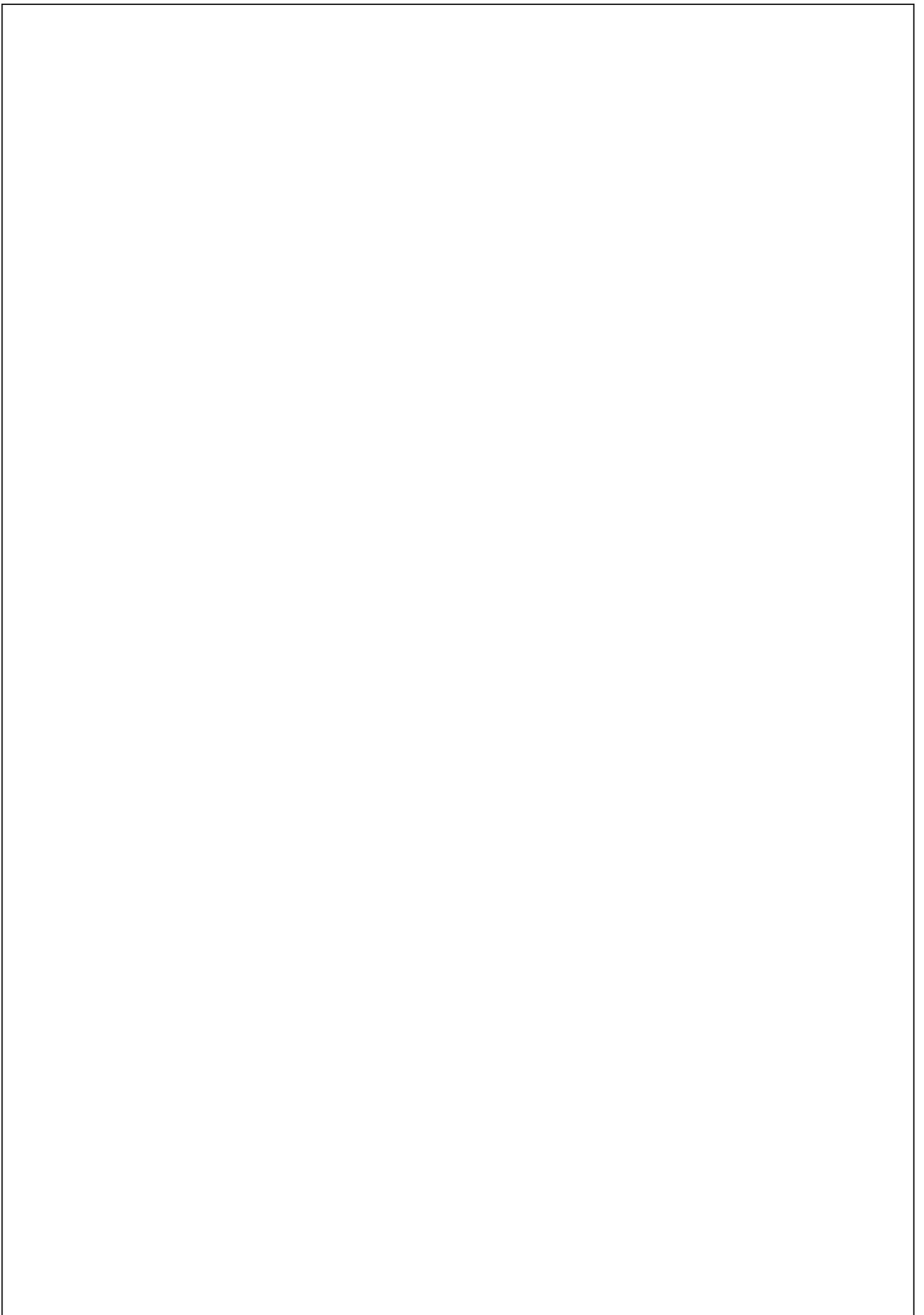
Activity ID: 81

Duration

Oct. 1st 2006 –Dec. 31st 2007

Main task force activities

- Position paper and introductory talk and report. (NiSIS Mallorca Brainstorming Meeting, June 2006 and NiSIS Symposium, November 2006)
- Participation in Workshops at NiSIS Symposium, November 2006
- Organization of Workshop on Nature inspired Methods for Local Pattern Detection, held in conjunction with NiSIS Symposium, November 2007.
- Participation in NiSIS Symposium, November 2007



Achievements & Conclusions

Context of the task force

Increasingly large data resources in life sciences, mobile information and communication, e-commerce, and other application domains require automatic techniques for gaining knowledge. One of the major knowledge discovery tasks is clustering. It aims at summarizing data base objects such that similar objects are grouped together while dissimilar ones are separated.

In scenarios with many attributes or with noise, clusters are often hidden in subspaces of the data and do not show up in the full dimensional space. For these applications, subspace clustering methods aim at detecting clusters in any subspace. As dimensionality of subspaces varies, approaches which do not take this into account fail to separate clusters from noise.

Focus 1: Eliminating dimensionality bias for local pattern detection

In high dimensional data, detecting patterns is challenging as distances get more and more similar. With increasing dimensionality the distance to the nearest neighbor approaches the distance to the farthest neighbor. This effect is one of the problems often termed as “curse of dimensionality”. For pattern detection algorithms like clustering this “curse of dimensionality” often obscures patterns and hence hinders the detection of meaningful clusters.

Local pattern detection algorithms like subspace clustering tackle the curse of dimensionality by identifying patterns in subspaces of the full attribute space. Pattern detection in different subspaces leads to incomparability of traditional approaches, termed “dimensionality bias”. We give a formal definition of “dimensionality bias” and analyze consequences for subspace clustering. A dimensionality unbiased subspace clustering definition based on statistical foundations is proposed. Experiments on synthetic and real world data sets demonstrate that our approach outperforms existing subspace clustering models in terms of accuracy and redundancy removal.

Focus 2: Efficient local pattern detection

To identify meaningful patterns in high-dimensional spaces, subspace clustering algorithms search for clusters in arbitrary subspaces. As the number of subspaces is exponential with the dimensionality efficiency is an important issue for subspace clustering algorithms. Working in a top down manner (starting from the full space) on the lattice of the dimensions or bottom up (starting from the empty space), monotonicity properties may be exploited to prune subtrees of the lattice from further consideration.

Alternatively, the performance of subspace clustering algorithms may be improved by proceeding in two steps. In the first step subspaces with a high cluster tendency are identified. Clusters are then detected in a second step. With increasing dimensionality pruning criteria have to cut off large portions of the search space to obtain reasonable runtime. As a consequence top down algorithms typically identify high dimensional clusters while bottom up algorithms only consider low dimensional spaces.

Focus 3: Nature inspired local pattern detection

To identify medium dimensional clusters with respect to the full dimensionality, we develop nature inspired methods. One promising approach for subspace search is to mimic evolution. Our subspace search method focuses clustering to relevant subspaces. In a first step, we search subspaces which are clustered in a second step. Mimicking evolutionary theory, a population of subspaces is subjected to nature-inspired optimization. Evaluation of subspaces is according to a fitness function which reflects the clustering tendency of the subspace.

Not just single optimal subspaces are interesting in subspaces clustering. Different local solutions represent interesting subspace combinations. These combinations are not necessarily related, yet constitute regions of interest to clustering algorithms. Such diverse local optima can be modeled by biological niches in our evolutionary approach. Evolutionary multi-objective optimization (EMOO) aims at detecting a set of near-optimal solutions. Our approach ensures that the population may split according to locally optimal conditions. In these sub-populations, evolutionary optimization is restricted to one such niche.

Moreover, subspace search is not always uninformed. In many cases, user guidelines on the usefulness of subspaces exist. Additionally, certain fuzzy properties such as an

unclear preference for higher dimensional subspaces cannot be modeled by restricting the population as such. Instead, these preferences should be incorporated into the evolutionary model.

Environmental change is used to model these guided optimization processes. As the environment changes, populations adapt by different selective processes. These are reflected in changing fitness functions and consequently in changing reproductive combinations.

An additional extension to these environmental changes is the evaluation of a number of different such guidelines during the lifetime of a population. This ensures a broader genetical diversity than typically observed in a few generations, extending the search space in optimization.

Summary of achievements

Nature inspired evolutionary algorithms are successfully used for dimensionality unbiased subspace clustering. Challenges arising from the large number of different subspaces in terms of comparability (“dimensionality bias”) as well as those arising from performance issues (“efficient pattern detection”) are handled by nature inspired methods (“nature inspired local pattern detection”) that model core properties of relevant subspaces for clustering in high dimensional spaces. For very high dimensional spaces, however, scalability is still an issue, requiring novel techniques that allow for directing of subspace search toward unexplored subspaces.

Future research directions

As part of the workshop “Nature Inspired Local Pattern Detection (NiLoP2007) held in conjunction with the annual NiSIS Symposium in Malta, promising nature inspired approaches for local pattern detection in these very high dimensional spaces were discussed. Using population dynamics to model an incentive toward unexplored subspaces, informed search strategies may be modeled. Using models of overpopulation to steer emigration of individuals to less populated niches in multiobjective evolutionary algorithms, unexplored subspaces are modeled as available niches that individuals are drawn to. Taking environmental change into account, reassignment of individuals to varying subspaces allows for diverse subspace exploration, a promising approach for search in very high dimensional settings.

These strategies could be evaluated in a variety of settings, determining the influence of various biological parameters such as population size, reproduction and mutation strategies on the quality and runtime performance, especially in different practical applications. Comparison with brute force approaches is generally not feasible as this outruns typically available computation resources, thus comparison of evolutionary approaches with well-established, traditional subspace detection algorithms should provide insights into the performance and accuracy gains of population dynamics for subspace search.

Publications

1. Assent I., Krieger R., Steffens A., Seidl T.: A Novel Biology inspired Model for Evolutionary Subspace Clustering, Proc. Annual Symposium on Nature inspired Smart Information Systems (NiSIS 2006), Puerto de la Cruz, Tenerife

2. Assent I., Krieger R., Müller E., Steffens A., Seidl T.: Evolutionary Subspace Search in biologically-inspired Optimal Niches, Proc. Annual Symposium on Nature inspired Smart Information Systems (NiSIS 2007), St Julians, Malta

3. Assent I., Krieger R., Müller E., Seidl T.: DUSC: Dimensionality Unbiased Subspace Clustering, Proc. IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA

4. Krieger R.: Nature inspired local pattern detection: new approaches to open data mining questions, Proc. Workshop on Nature-inspired Methods for Local Pattern Detection (NiLOP 2007) in conjunction with NiSIS 2007, St. Julians, Malta

5. Müller E.: Density-based clustering in arbitrary subspaces, Proc. Workshop on Nature-inspired Methods for Local Pattern Detection (NiLOP 2007) in conjunction with NiSIS 2007, St. Julians, Malta

6. Assent I.: Population dynamic coverage for subspace search, Proc. Workshop on Nature-inspired Methods for Local Pattern Detection (NiLOP 2007) in conjunction

with NiSIS 2007, St. Julians, Malta

7. Assent I., Krieger R., Müller E., Seidl T.: Removing Dimensionality Bias in Density-based Subspace Clustering, Dutch-Belgian Data Base Day (DBDBD 2007), Eindhoven, NL

8. Assent I., Krieger R., Seidl T.: Evolutionäre Clustering-Algorithmen: Bionik-Methoden für Knowledge Discovery in Datenbanken, RWTH Themen 1/2008: Bionik.

9. Assent I., Krieger R., Welter P., Herbers J., Seidl T.: SubClass: Classification of Multidimensional Noisy Data Using Subspace Clusters, Proc. 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008), Springer LNCS/LNAI, Osaka, Japan, 2008

Further References

1. Agrawal, R.; Gehrke, J.; Gunopulos, D.; Ragha van, P., 1998: Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings ACM SIGMOD, pages 94-105.

2. Bennett, Kristin; Ferris Michael C.; Ioannidis Yannis E., 1991: "A genetic algorithm for database query optimization", Proceedings International Conference on Genetic Algorithms, San Mateo, CA, pp. 400-407

3. Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang, 1999: Entropy-based subspace clustering for mining numerical data. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, pp 84-93. ACM Press.

4. 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 (KDD'96), Portland, OR, 1996, pp. 226-231.

5. Garai G.; Chaudhuri B.B; 2004: "A novel genetic algorithm for automatic clustering" , Pattern Recognition Letters, Volume 25, Number 2, 19 January 2004, pp. 173-187

6. Hall LO, Ozyurt IB, Bezdek JC, 1999: "Clustering with a genetically optimized approach". *IEEE Trans. Evolutionary Computation* 3(2), 103-112.

7. Hinneburg A., Keim D.A., 1998: "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", *Proc. 4rd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 58-65

8. Jin, Y, 2002: "A Comprehensive Survey of Fitness Approximation in Evolutionary Computation", *Soft Computing* 9, pp 3-12

9. K. Kailing, H. Kriegel, P. Kroeger and S. Wanka, 2003: Ranking interesting subspaces for clustering high dimensional data. In PKDD, pp 241-252.

10. Kröger P., Kriegel H.-P., Kailing K.: Density-Connected Subspace Clustering for High-Dimensional Data, *Proc. SIAM Int. Conf. on Data Mining (SDM'04)*, Lake Buena Vista, FL, 2004, pp. 246-257.

11. Lawrence O. Hall, Ibrahim Burak, Özyurt and James C. Bezdek, 1999: "Clustering with a Genetically Optimized Approach", *IEEE Trans. on Evolutionary Computation*, pp. 103-112

12. Park Y and Song M, 1998: "A genetic algorithm for clustering problems." *Genetic Programming 1998: Proc. 3rd Annual Conf.*, 568-575.

13. Punch WF, Goodman ED, Pei M, Chia-Sun L, Hovland P, Enbody R, 1993: "Further research on feature selection and classification using genetic algorithms." *Proc. 5th Int. Conf. Genetic Algorithms (ICGA-93)*, 557-564.

14. Sarafis I. A. , Trinder P. W., Zalzala A. M. S., 2003: "Towards effective subspace clustering with an evolutionary algorithm", *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003*, pp. 797--806

15. Sequeira, K., Zaki, M., 2004: SCHISM: a new approach for interesting subspace mining. In *Proceedings of the IEEE International Conference on Data Mining ICDM*, page(s): 186- 193.

16. Shannon C., Weaver W. 1949: *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois.

17. Ursem, Rasmus K., 2002: "Diversity-Guided Evolutionary Algorithms", *Proceedings of Parallel Problem Solving from Nature VII*, pp 462-471

18. Uyar Sima Etaner, Harmanci A. Emre, 1999, "Investigation of New Operators for a Diploid Genetic Algorithm" Proceedings of SPIE, Volume 3812,1999, pp 32-43