Introduction

One thing a child has to learn is to divide and to group objects based on their color, form or size, i.e. based on their attributes. Such an ability is very important for the improvement of abstract and logical human thinking. But it is also a very helpful ability in economics, industry, science or politics, where the identification and description of homogeneous groups — so called *clusters* — of customers, products, events or situations helps to structure information about these objects and therefore generates knowledge, which allows to make special, group depending offers or decisions. Unfortunately, the ad hoc identification and description of clusters by human beings usually gets impossible with increasing numbers of objects and attributes.

Clustering methods have been studied first in statistics[1], but nowadays, where the improvement of technology allows to store the data of millions of objects with hundreds of attributes in single databases, new techniques for *cluster analysis* are also suggested by researchers from the *machine learning/neural networks* area[2] and the *database* community[3]. Furthermore, a new direction of research called *Data Mining* or — according to the more general definition of Fayyad and Piatetsky-Shapiro [21] — *Knowledge Discovery in Databases (KDD)*, has been established, where algorithms are developed that are able to scan huge databases and to extract *knowledge patterns* within the data. Since clusters are important examples of such knowledge patterns, the development of fast and efficient clustering techniques is part of this fast growing research area.

The most popular clustering method is k*-means*, and most of the suggested algorithms in the literature are variants of this method. The basic idea of k-means is to determine k cluster representatives and to assign each object to the cluster with its representative closest to the object so that the sum of the squared distances between the objects and their corresponding representatives is minimized.

---

[1] See, e.g., the introductory textbooks by Duran and Odell [18] or Fukunaga [27].

[2] For an overview see the complementary textbooks by Bishop [7] and Ripley [53].

[3] Important research is not only done by database groups at university [64, 19, 32], but also from industrial groups like IBM'S Quest group [1].

An investigation of algorithms based on the k-means method or other frequently used clustering methods leads to the following observations:

- The computed clusters are *geometrically* based, i.e. the objects within the same cluster have the property that their distance is small if they are interpreted as points in a suitable metric space. For non-geometric cluster problems, the computed clusters are usually not satisfactory. An important example are *dynamic* cluster problems, where one is interested in the identification of *metastable* clusters. Here, the objects within the same cluster should exhibit a high probability for transitions between each other with regard to an underlying *dynamic system*.

- If the numbers of objects and attributes is high, heuristics are used to speed up the cluster identification process. Many of these heuristics are designed for special applications and therefore not generally usable. Further, a mathematical justification is very often missing.

- A correct number of clusters k has to be known a priori.

In the case of reversible dynamic cluster problems, the theory of *Perron Cluster* analysis that has been recently developed by DEUFLHARD ET AL. offers a new access. The key concept of Perron Cluster analysis is the identification of metastable clusters by computing *almost invariant aggregates* of a suitable stochastic matrix $\mathcal{S}$. Via an investigation of the eigenvalues and the eigenvectors of the matrix $\mathcal{S}$, not only a correct number of clusters k can be determined, but also the metastable clusters themselves. Without a problem reduction, the size of the matrix $\mathcal{S}$ depends on the number of objects that have to be clustered. Therefore Perron Cluster analysis is directly usable only for very small reversible dynamic cluster problems[4].

Self-organized neural networks, especially KOHONEN'S *Self-Organizing Maps*, can be used to replace groups of similar objects by single representatives. The representatives are related to each other in a way that tries to preserve the original cluster structure, i.e. a fitting clustering of the representatives should correspond to a fitting clustering of the original objects. In contrast to the k-means method and its variants, the number of representatives is usually much larger than any correct number of clusters. Therefore, self-organized neural networks can be used as a kind of pre-clustering process to reduce the complexity of a cluster problem.

The aim of this thesis is a fruitful combination of Perron Cluster analysis and self-organized neural networks within an *adaptive multilevel clustering approach*

[4] As a first remedy, the use of essential degrees of freedom in the spirit of [4] made it possible to identify metastable clusters of a small molecule via Perron Cluster analysis [59].

that allows a fast and robust identification and an efficient description of clusters in *high-dimensional* data. In a general variant that needs a correct number of clusters k as an input, this new approach is relevant for a great number of cluster problems since it uses a cluster model that covers geometrically, but also dynamically based clusters. Its essential part is a method called *representative clustering* that guarantees the applicability to large cluster problems: Based on an *adaptive decomposition* of the object space via self-organized neural networks, the original problem is reduced to a smaller cluster problem. The general clustering approach can be extended by Perron Cluster analysis so that it can be used for large reversible dynamic cluster problems, even if a correct number of clusters k is unknown a priori. The basic application of the extended clustering approach is the *conformational analysis* of biomolecules, with great impact in the field of *Drug Design*. Here, for the first time the analysis of practically relevant and large molecules like an *HIV protease inhibitor* becomes possible.

This thesis is divided into five chapters. It starts with a general mathematical definition of cluster analysis in high-dimensional data. The scalability problem of the identification step will be addressed and the idea of representative clustering will be presented. In the section following, a rigorous definition of efficient cluster description will be given. The first chapter closes with a survey of the difficulties that arise, if a correct number of clusters is not known a priori.

The second chapter establishes a concept of decomposition within cluster analysis. Based on a general definition we will present a special variant called approximate box decomposition. It will be shown that the concept of decomposition gives way to a significant cluster problem reduction via representative clustering without destroying the original cluster structure. In addition, the usefulness of approximate box decompositions for the computation of efficient cluster descriptions will be demonstrated.

In the following chapter, KOHONEN'S Self-Organizing Maps are used for the computation of adaptive decompositions. Further, a powerful extension called *Self-Organizing Box Maps* will be suggested that computes approximate box decompositions.

In the fourth chapter, we are going to present a multilevel clustering approach using representative clustering based on successively refined adaptive decompositions. After an introduction to the basic theory, we combine Perron Cluster analysis with our clustering approach so that it includes an automatic computation of a correct number of cluster for cluster problems with a stochastic homogeneity function.

The final chapter gives a comprehensive presentation of applications. Especially the conformational analysis of biomolecules will be described in detail and illustrated with numerical results.

Acknowledgment

Foremost, I would like to thank P. Deuflhard for guiding me into a fascinating research area and giving me the chance to work in his group. His constant encouragement and confidence were extremely helpful for the progress of this thesis.

Furthermore, I am indebted to J. Weyer, who has taught me to become a real applied mathematician. During the last years he spent much of his rare time, on giving me advice and support.

Finally, I have to thank my parents and my girl-friend Simone for showing me that live is a wonderful present.