Preface

Humans have been "manually" extracting patterns from data for centuries, but the increasing volume of data in modern times has called for more automatic approaches. Early methods of identifying patterns in data include Bayes' theorem (1700s) and Regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. Data mining has been developed as the tool for extracting hidden patterns from data, by using computing power and applying new techniques and methodologies for knowledge discovery. This has been aided by other discoveries in computer science, such as Neural networks, Clustering, Genetic algorithms (1950s), Decision trees (1960s) and Support vector machines (1980s). Data mining commonly involves four classes of tasks:

- Classification: Arranges the data into predefined groups. For example, an e-mail program might attempt to classify an e-mail as legitimate or spam. Common algorithms include Nearest neighbor, Naive Bayes classifier and Neural network.
- Clustering: Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.
- Regression: Attempts to find a function which models the data with the least error. A common method is to use Genetic Programming.
- Association rule learning: Searches for relationships between variables. For example, a supermarket might gather data of what each customer buys. Using association rule learning, the supermarket can work out what products are frequently bought together, which is useful for marketing purposes. This is sometimes referred to as "market basket analysis."

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set; this is called overfitting. To overcome this, the evaluation

uses a test set of data which the data mining algorithm was not trained on. The learned patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish spam from legitimate e-mails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained, and the accuracy of these patterns can then be measured from the number of e-mails they classify correctly. A number of statistical methods such as ROC (receiver operating characteristic) curves may be used to evaluate the algorithm. If the learned patterns do not meet the desired standards, then it is necessary to reevaluate and change the preprocessing and data mining. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

In recent years, data mining has been widely used in areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering, as well as in businesses and governments, to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. The present volume, dedicated to the application of Data Mining in Crystallography, is organized as following.

The Chapter "An Introduction to Data Mining" written by J. Apostolakis provides an introduction and a short overview of the mathematical concepts and ideas behind the most relevant methods of data mining approach in crystallography. Some applications are described in the following chapters (and we hope these will make the mathematical concepts of introduction more accessible for a wide range of readers); other methods still do not find a large application in crystallography but we hope the chapter will open up possibilities for future discoveries in this field.

Crystallographers were among the first scientists to recognize the importance of collecting data on crystal structures. As a result nowadays a large amount of crystal structural data is collected in several big crystallographic data bases. The comprehensiveness of the data collection, the structure and quality of the data and the selection of relevant data sets are extremely important to get reasonable results from the fully automatic procedure of data mining. As one example, the Inorganic Crystal Structure Database (ICSD), a source of information for crystallographers, mineralogists, physicist and chemists, is presented in the Chapter "Data bases, the base for Data Mining" by Ch. Buchsbaum, Sabine Höhler-Schlimm, and S. Rehme, along with a short overview of all existing crystal structural data bases.

In the Chapter "Data Mining and Inorganic Crystallography" by K. Rajan, an overview of the types of information that can be gleaned by applying data mining and statistical learning techniques to inorganic crystallography is provided. The focus is on two broad areas, classification and prediction of data, the two primary roles of data mining as a field. A fundamental issue in inorganic crystallography is to understand the relationship between chemical stoichiometry and crystal structure. The relationship between specific compounds and specific crystal structures is usually developed heuristically by surveying the crystallographic data of known compounds. This process of structure–chemistry association has laid the historical foundations of identifying crystal structure prototypes and structural classifications.

This demonstrates how informatics can quantitatively accelerate the discovery of structure–chemistry relationships and also be used as the foundation for developing structure–chemistry–property relationships.

In the Chapter "Data Mining in Organic Crystallography" by D.W.M. Hofmann, two actual applications of Data Mining have been highlighted: the cluster analysis and the support vector machines (SVM). The SVMs are used to find errors in the Cambridge Structural Database of small molecule crystal structures and to derive force fields without any hypotheses on the functional form. Since the accuracy of the force fields derived by data mining depends on the number of known crystal structures, this approach should be favored in the long-term. The second method, clustering, has been introduced in this field only very recently. An obvious application is the screening of data bases to remove undesired repetitions of crystal structures. This is important for all, virtual as well as for experimental, data bases. Its application is interesting in crystal structure determination, where it can be used to find isostructural crystal structures. With this simple application, the knowledge about regularities between isostructural crystal structures gives very valuable information for crystal engineering. A third method, principal component analysis, might become more important in the future, as it is already in use nowadays in inorganic crystallography.

The last chapter of the volume, "Data mining in Protein Secondary Structure Prediction" written by A. Kloczkowski et al. is dedicated to the application of data mining techniques to extract the predictive information from the protein, DNA and RNA Data Bases. Data Mining in biology is a rapidly growing field of science, combining molecular biology, computational methods and statistics for analyzing and processing biological data. This has led to the development of a new field of science: bioinformatics. Mining for information in biological databases involves various forms of data analysis such as clustering, sequence homology searches, structure homology searches, examination of statistical significance, etc. Particularly, the data mining of structural fragments of proteins from known structures in the Protein Data Bank significantly improves the accuracy of secondary structure prediction. Since the crystallization of these objects is the most serious bottleneck in high-throughput protein-structure determination by diffraction methods, it is to be noted that the data mining approach is also used to characterize the biophysical properties and conditions that control and improve a protein crystallization.

The tendency to increase the accuracy of the crystal structure data promises a better quality of patterns obtained by Data Mining for the future because the quality of the result depends strongly on the amount, the quantity and reliability of the data used. The aim of the volume is to show the possibilities of the method used in knowledge discovery in crystallography. We hope that it will make Data Mining more accessible in crystallography and allow new applications in the field and the discovery of non trivial and scientifically relevant knowledge.

Pula, September 2009

Detlef W.M. Hofmann
Liudmila N. Kuleshova

http://www.springer.com/978-3-642-04758-9

Data Mining in Crystallography
(Eds.)D.W.M. Hofmann; L.N. Kuleshova
2010, XIII, 172 p., Hardcover
ISBN: 978-3-642-04758-9