Chapter 4
Active Learning and Mapping: A Survey and Conception of a New Stochastic Methodology for High Throughput Materials Discovery

Laurent A. Baumes
CSIC-Universidad Politecnica de Valencia, Spain

ABSTRACT

The data mining technology increasingly employed into new industrial processes, which require automatic analysis of data and related results in order to quickly proceed to conclusions. However, for some applications, an absolute automation may not be appropriate. Unlike traditional data mining, contexts deal with voluminous amounts of data, some domains are actually characterized by a scarcity of data, owing to the cost and time involved in conducting simulations or setting up experimental apparatus for data collection. In such domains, it is hence prudent to balance speed through automation and the utility of the generated data. The authors review the active learning methodology, and a new one that aims at generating successively new samples in order to reach an improved final estimation of the entire search space investigated according to the knowledge accumulated iteratively through samples selection and corresponding obtained results, is presented. The methodology is shown to be of great interest for applications such as high throughput material science and especially heterogeneous catalysis where the chemists do not have previous knowledge allowing to direct and to guide the exploration.

DOI: 10.4018/978-1-60960-860-6.ch004

1. INTRODUCTION

Data mining, also called knowledge discovery in databases (Piatetsky-Shapiro & Frawley, 1991;Fayyad, Piatetsky-Shapiro & Smyth, 1996) (KDD) is the efficient discovery of unknown patterns in databases (DBs). The data source can be a formal DB management system, a data warehouse or a traditional file. In recent years, data mining has invoked great attention both in academia and industry. Understanding of field and defining the discovery goals are the leading tasks in the KDD process. It can be distinguished two aims: *i*) *verifications*, where the user hypothesizes and mines the DB to corroborate or disprove the hypothesis; *ii*) *Discovery*, where the objective is to find out new unidentified patterns. Our contribution is concerned by the latter, which can further be either predictive or descriptive. The data mining technology is more and more applied in the production mode, which usually requires automatic analysis of data and related results in order to proceed to conclusions. However, an absolute automation may not be appropriate. Unlike traditional data mining contexts deal with voluminous amounts of data, some domains are actually characterized by a scarcity of data, owing to the cost and time involved in conducting simulations or setting up experimental apparatus for data collection. In such domains, it is hence prudent to balance speed through automation and the utility of the generated data. For these reasons, the human interaction and guidance may lead to better quality output: the need for *active learning* arises.

In many natural learning tasks, knowledge is gained iteratively, by making action, queries, or experiments. Active learning (AL) is concerned with the integration of data collection, design of experiment, and data mining, for making better data exploitation. The learner is not treated as a classical passive recipient of data to be processed. AL can occur due to two extreme cases. *i*) The amount of data available is very large, and therefore a miming algorithm uses a selected data subset rather than the whole available data. *ii*) The researcher has the control of data acquisition, and he has to pay attention on the iterative selection of samples for extracting the greatest benefit from future data treatments. We are concerned by the second situation, which becomes especially crucial when each data point is costly, domain knowledge is imperfect, and theory-driven approaches are inadequate such as for heterogeneous catalysis and material science fields. Active data selection has been investigated in a variety of contexts but as far as we know, this contribution represents the first investigation concerning this chemistry domain.

A catalytic reaction is a chemical reaction in which transformations are accelerated thanks to a substance called catalyst. Basically, starting molecules and intermediates, as soon as they are formed, interact with the catalyst in a specific/discriminating manner. This implies that some transformation steps can be accelerated while other can be kept constant or even slowed down. Catalytic processes constitute the fundamentals of modern chemical industries. Over 90% of the newly introduced chemical processes are catalytic. In the highly developed industrial countries, catalytic processes create about 20% of the Gross Domestic Product. Catalysis is responsible in the manufacture of over $3 trillion in goods and services. We will focus on heterogeneous catalysis which involves the use of catalysts acting in a different phase from the reactants, typically a solid catalyst with liquid and/or gaseous reactants. For further details, the reader is referred to (Ertl, Knozinger & Weitkamp, 1997). During the whole catalytic development, a very large number of features and parameters have to be screened and therefore any detailed and relevant catalyst description remains a challenge. All these parameters generate an extremely high degree of complexity. As a consequence, the entire catalyst development is long (~15 years) and costly. The conventional catalyst development relies essentially on fundamental knowledge and know-how. It implies a complete characterisation of the catalyst