Dynamic Data Mining

Richard Weber
University of Chile, Chile

INTRODUCTION

Since the First KDD Workshop back in 1989 when "Knowledge Mining" was recognized as one of the top 5 topics in future database research (Piatetsky-Shapiro 1991), many scientists as well as users in industry and public organizations have considered data mining as highly relevant for their respective professional activities.

We have witnessed the development of advanced data mining techniques as well as the successful implementation of knowledge discovery systems in many companies and organizations worldwide. Most of these implementations are static in the sense that they do not contemplate explicitly a changing environment. However, since most analyzed phenomena change over time, the respective systems should be adapted to the new environment in order to provide useful and reliable analyses.

If we consider for example a system for credit card fraud detection, we may want to segment our customers, process stream data generated by their transactions, and finally classify them according to their fraud probability where fraud pattern change over time. If our segmentation should group together homogeneous customers using not only their current feature values but also their trajectories, things get even more difficult since we have to cluster vectors of functions instead of vectors of real values. An example for such a trajectory could be the development of our customers' number of transactions over the past six months or so if such a development tells us more about their behavior than just a single value; e.g., the most recent number of transactions.

It is in this kind of applications is where dynamic data mining comes into play!

Since data mining is just one step of the iterative KDD (Knowledge Discovery in Databases) process (Han & Kamber, 2001), dynamic elements should be considered also during the other steps. The entire process consists basically of activities that are performed before doing data mining (such as: selection, pre-processing, transformation of data (Famili et al., 1997)), the actual data mining part, and subsequent steps (such as: interpretation, evaluation of results).

In subsequent sections we will present the background regarding dynamic data mining by studying existing methodological approaches as well as already performed applications and even patents and tools. Then we will provide the main focus of this chapter by presenting dynamic approaches for each step of the KDD process. Some methodological aspects regarding dynamic data mining will be presented in more detail. After envisioning future trends regarding dynamic data mining we will conclude this chapter.

BACKGROUND

In the past a diverse terminology has been used for emerging approaches dealing with "dynamic" elements in data mining applications. Learning from data has been defined as *incremental* if the training examples used become available over time, usually one at a time; see e.g., (Giraud-Carrier, 2000). Mining *temporal* data deals with the analysis of streams of categorical data (e.g., events; see e.g., Domingos, Hulten, 2003) or the analysis of time series of numerical data (Antunes, Oliveira 2001; Huang, 2007). Once a model has been built, *model updating* becomes relevant. According to the CRISP-DM methodology such updating is part of the monitoring and maintenance plan to be performed after model construction.

The following listing provides an overview on applications of dynamic data mining.

- Intrusion detection (Caulkins et al., 2005).
- Traffic state identification (Crespo, Weber, 2005).
- Predictive maintenance (Joentgen et al., 1999).
- Scenario analysis (Weber 2007).
- Time series prediction (Kasabov, Song, 2002)

Dynamic data mining has also been patented already, e.g., the dynamic improvement of search engines in internet which use so-called rating functions in order to measure the relevance of search terms. "Based upon a historical profile of search successes and failures as well as demographic/personal data, technologies from artificial intelligence and other fields will optimize the relevance rating function. The more the tool is used (especially by a particular user) the better it will function at obtaining the desired information earlier in a search. ... The user will just be aware that with the same input the user might give a static search engine, the present invention finds more relevant, more recent and more thorough results than any other search engines." (Vanderveldt, Black 2001).

MAIN FOCUS

As has been shown above dynamic data mining can be seen as an area within data mining where dynamic elements are considered. This can take place in any of the steps of the KDD process as will be introduced next.

Feature Selection in Dynamic Data Mining

Feature selection is an important issue of the KDD process before the actual data mining methods are applied. It consists in determining the most relevant features given a set of training examples (Famili et al., 1997). If, however, this set changes dynamically over time the selected feature set could do so as well. In such cases we would need a methodology that helps us to dynamically update the set of selected features. A dynamic wrapper approach for feature selection has been proposed in (Guajardo et al., 2006) where feature selection and model construction is performed simultaneously.

Preprocessing in Dynamic Data Mining

If in certain applications feature trajectories instead of feature values are relevant for our analysis we are faced with specific requirements for data preprocessing. In such cases it could be necessary to determine distances between trajectories within the respective data mining algorithms (as e.g., in Weber, 2007) or alternatively to apply certain preprocessing steps in order to reduce the trajectories to real-valued feature vectors.

Dynamic Clustering

Clustering techniques are used for data mining if the task is to group similar objects in the same classes (segments) whereas objects from different classes should show different characteristics (Beringer, Hüllermeier 2007). Such clustering approaches could be generalized in order to treat different dynamic elements, such as e.g., dynamic objects and/or dynamic classes. Clustering of **dynamic objects** could be the case where trajectories of feature vectors are used as input for the respective data mining algorithms. If the class structure changes over time we speak about **dynamic classes**. We present first approaches for clustering of dynamic objects and then a methodology to determine dynamic classes.

In order to be able to cluster dynamic objects, we need a distance measure between two vectors where each component is a trajectory (function) instead of a real number. Functional fuzzy c-means (FFCM) is a fuzzy clustering algorithm where the respective distance is based on the similarity between two trajectories which is determined using membership functions (Joentgen et al., 1999).

Applying this distance measure between functions FFCM determines classes of dynamic objects. The respective class centers are composed of the most representative trajectories in each class; see the following figure.

Determine dynamic classes could be the case when classes have to be created, eliminated or simply moved in the feature space. The respective methodology applies the following five steps in order to detect these changes.

Here we present just the methodology's main ideas; a detailed description can be found e.g., in (Crespo, Weber, 2005). It starts with a given classifier; in our case we chose Fuzzy c-means since the respective membership values provide a strong tool for classifier updating.

Step I: Identify Objects that Represent Changes

For each new object we want to know if it can be explained well by the given classifier. With other words