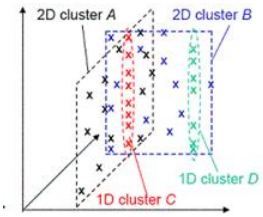
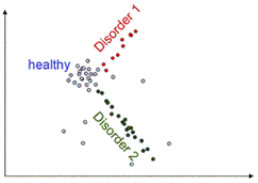
Data stream mining

There are many examples of data streams, including computer network traffic, telecommunication network calls, supermarket transactions, ATM transactions, web searches and sensor data. A key characteristic of this data is *variability*, i.e., data change over time as new instances arrive and old instances become obsolete. This has a strong impact on the extracted data mining models, as they are not able to model the underlying data in the long run and therefore, they should be maintained in an online fashion. Such an online maintenance is not straightforward as traditionally data mining and machine learning are applied over fixed datasets (batch processing). There is a need therefore to develop new methods and algorithms that can deal with data variability, in an efficient way.

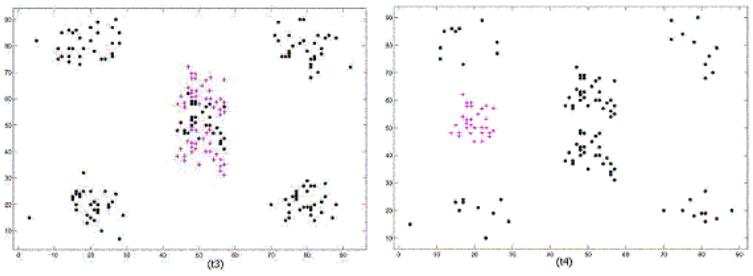
1. Clustering over high dimensional data streams [Master thesis]

In high dimensional spaces, it is typically difficult to find clusters in the complete full dimensional feature space. But, one can detect clusters in subspaces of the original feature space. Some examples are presented below: on the left, we see axis-parallel subclusters, on the right we see arbitrary-oriented subclusters:





In a stream setting, as the data change over time, both cluster members and subspaces where these clusters exist might change. An example of clustering change (here, in the full dimensional feature space) is presented below:



The goal of this project is to develop clustering methods for high dimensional data streams.

2. Feature ranking in data streams [Master thesis]

Feature evaluation (and in general, feature space selection) is an important part of data mining. Typically feature evaluation is performed upon the whole dataset of instances (batch evaluation). In a stream setting though, there is no fixed dataset as the stream evolves over time. Moreover, the importance of features might change over time as the underlying data population changes.

The goal of this project is to evaluate features in a stream environment and maintain the top-*k* most relevant/important features over time.

3. Outlier detection in data streams [Master thesis]

Outlier detection is an important problem, especially in a stream environment. As an example, you can consider monitoring of abnormal credit card transactions or spam behaviour or suspiciously high traffic nodes in a network. The main challenge for data streams is that the role of outliers is subject to time, as what is an outlier now might turn into a cluster later on or different sort of outliers might arise with time.

The goal of this project is to monitor top-*k* outliers in a stream environment over time.

4. Multi-label stream classification [Master thesis]

Classification or supervised learning is one of the most important tasks in data mining and machine learning. In multi-label classification, instead of a single class-label as is typical in traditional classification, each instance can be associated with multiple labels, e.g., a document-instance is associated with multiple topics-classes.

The goal of this project is to develop classification methods for multi-label data streams.

Requirements

- Very good knowledge of data mining (at least one of the DM 1, DM2 lectures)
- Very good programming knowledge (Java, Python, R)
- Motivation, hard and independent work

If you are interested, please contact: Prof. Dr. Eirini Ntoutsi (ntoutsi@kbs.uni-hannover.de). Please include your CV with information on your data mining knowledge and programming skills.