Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

Master Thesis

Evaluation of Automated Feature Generation Methods

Feature engineering (a.k.a. feature construction/generation) is usually a manual process. It is the most time-consuming part in machine learning projects or data mining competitions. It is also arguably the most interesting and creative part and the key to success.

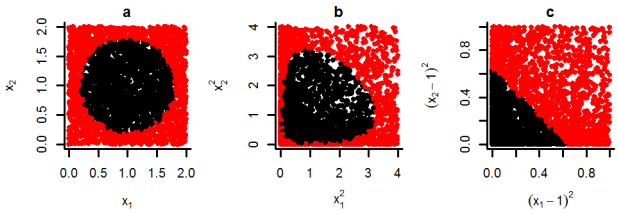
A number of automated feature generation methods have appeared recently with the aim to shift the burden from the researcher to the machine, which is less expensive to operate. These methods usually include three main steps:

- create new features from existing attributes;
- select a subset of features in supervised or unsupervised manner;
- train the model.

Though respective publications provide some evidence that these algorithms improve performance of some methods on some exemplary datasets, the following questions remain open:

1. How to set parameters of a feature generation method to get the best result?
2. How to compare different automated feature generation methods to each other?
3. Do machine learning techniques on average benefit from these methods?
4. Can automation replace manual feature generation, at least to some extent?

One may be very skeptical regarding Questions 3 and 4. The reason is that an increasing number of features leads to the well-known problem of overfitting, not fully solved by feature selection. Moreover, the infinite number of possibilities to create new features makes it unlikely to obtain relevant ones in an automated manner.



Think of the classification task depicted on the figure (a). Direct creation of new features by taking squares of initial ones produces a space (b), which is not much better than the original one. This would be the case for many other operations performed by automated methods. On the other hand, when looking at the data more thoroughly, one might have designed better features, by squaring the attributes after a shift (c).

You are asked to come up with a meaningful answers to the questions formulated above. To do so, the following steps are important:

- review the existing automatic feature generation methods and classify them;
- design and perform a set of experiments, explain results theoretically and develop your own method combining the best solutions;
- compare the performance of winning codes for different data mining competitions (kaggle etc.) to that achieved with automated methods on the same tasks.

Throughout this work, you will get familiar with the best practices in data mining applications as well as theory behind various machine learning models.

Contact

Vadim Arzamasov vadim.arzamasov@kit.edu Room: 340

Am Fasanengarten 5 76131 Karlsruhe Building: 50.34