Data Analytics using *KNIME* open source tool

Dörheit, Eric
TU Berlin
Berlin, Germany
eric.doerheit@campus.tu-berlin.de

Bode, Olga
TU Berlin
Berlin, Germany
olga.bode@mailbox.tu-berlin.de

Poljak, Dorothea
TU Berlin
Berlin, Germany
@mailbox.tu-berlin.de

ABSTRACT

We present our results of the evaluation of the open source tool *KNIME* which is used for data analytics and data mining. We choose anomaly detection as the subject to evaluate *KNIME* as many methods of data analytics such as clustering, classification, time series analysis and statistical techniques are applicable to anomaly detection [1]. As a data set for the analysis we use the data provided for the *DEBS Grand Challenge 2012* [2].

1. INTRODUCTION

For the evaluation of *KNIME* we first investigated the tool by following several white-papers provided by *KNIME*. We start with describing the functionalities and the usage of *KNIME*. Then we explain the data used for the evaluation. Subsequently we provide an overview on anomaly detection based on [1]. In Section 2 we evaluate *KNIME* through applying anomaly detection on the data set described in Section 1.2.

1.1 The Open Source Tool KNIME

In this section we give an overview on *KNIME* based on three white-papers provided by *KNIME*:

- Big Data, Smart Energy, and Predictive Analytics
- Anomaly Detection in Predictive Maintenance
- KNIME opens the Doors to Big Data

TODOS:

- General features of *KNIME* → what can you do with *KNIME* (ETL, Mining, Analysis, Visualization etc.)
- How to use *KNIME*? → Workflows, Nodes, ... (describe the usage of *KNIME* in general)
- Example workflow(s) based on the *KNIME* white-papers
- Overview on the nodes that are available and name, that there is an API to develop your own nodes
- Tiny summary (2 sentences) if possible

1.2 DEBS 2012 Grand Challenge

TODOS:

- Describe the challenge / the origin of the data
- Explain the data set

1.3 Anomaly Detection

Anomalies in data are patterns which do not conform to the expected behavior and anomaly detections deals with finding this patterns [1]. There are many techniques that can be applied to detect anomalies. Subsequently we describe classification based, nearest neighbor based and clustering based techniques as well as statistical anomaly detection techniques.

1.3.1 Based on Classification

1.3.2 Based on Nearest Neighbor

1.3.3 Based on Clustering

1.3.4 Statistical Anomaly Detection Techniques

2. ANOMALY DETECTION WITH KNIME

TODOS:

- How to input the big files into *KNIME*? → Split files and iterate over them, input into MySQL etc.
- Describe the ETL process with *KNIME* of the data
- Describe how to do anomaly detection with *KNIME*

3. RESULTS

TODOS:

- *KNIME* is not directly suitable for Big Data Processing
- Easy tool to do advanced data analytics without deep knowledge of underlying algorithms and math
- Enables users which are no data scientists or have a strong background in this field to do data analysis
- Good integration with various other tools (R, Weka) and adoptable to own needs with Java, Python, ... snippet nodes and the API to create own nodes
- Relatively slow

4. CONCLUSIONS

5. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[2] Z. Jerzak, T. Heinze, M. Fehr, D. Gröber, R. Hartung, and N. Stojanovic. The debs 2012 grand challenge. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, DEBS '12, pages 393–398, New York, NY, USA, 2012. ACM.