Interactive Data Quality Assistance – An Approach for Min(d)ing the Quality of Data

Nadia El Bekri, Elisabeth Peinsipp-Byma
Fraunhofer Institute of Optronics System Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany

Abstract - *In this paper we introduce the concept of an interactive data quality system. Today one of the most challenging goals of data processing is to achieve and to maintain high data quality. Especially when the data is added manually by multiple users. The idea was originated out of the need to analyze the data set from a recognition assistance system. The system supports aerial image analysts in the task of the object recognition. Depending on which object features the aerial image analyst selects the solution set of the object types gets more precise. Especially in this field mechanisms that improve high data quality are important. Therefore we developed an interactive data quality system that helps experts generating potential rules through correlation that describe the whole data set. More precisely, we search for rules within the data set that are in general valid for a certain group of object types.*

Keywords: data quality, interactive data analysis

1 Introduction

The main idea was to develop a system that analyzes a given set of data with the help of an interactive data quality system and thereby derive rules that are valid for the entire data set. The underlying data set was taken from the WDI (World Development Indicators). This data set is a collection of development indicators from international resources. It presents the most current and accurate global development data available, and includes national, regional and global estimates. The database contains more than 900 indicators for over 210 countries [1]. For example the country “Switzerland” contains the indicators “GDP per capita” and many other indicators. In this case, the countries represent our objects and the indicators are the features of the objects that describe every country in a particular way. What we did was the first introductory step in a whole quality assurance process. The idea was originated out of the need to analyze the data set from a recognition assistance system. The system supports aerial image analysts in the task of object recognition by allowing them to describe single object features. Thereby the aerial image analyst can interactively classify the objects by selecting the visually extracted object features. The solution set contains only the amount of objects that match the selected features. In a previous step the objects are added manually by multiple users to the database or multiple databases are fused. Obviously, this is a critical issue. Why? First, the user can assign the features to the wrong objects or can forget to assign a potential feature to an object. Second, the feature values can be out of range for a certain group of objects. This means that probably for certain object types only a specific range is right. All this causes can lead to wrong or incomplete solution sets. The potential benefit for the user, in this case to recognize a specific object, is getting lost. We took the data from the WDI as underlying data set because the military data set is confidential and cannot be released for publication purposes. In addition it is irrelevant which data set is used as long there is an object-feature relation between them.

2 Quality assurance process

In order to be able to achieve and maintain a high quality data set we build up a quality assurance process. From analyzing the data set with algorithms, we receive certain rules that are probably generally valid for a certain group of objects containing the same object features and correlate very strongly. First, the data needs to be analyzed and rules have to be derived. The second step is then to apply the potential rules on the data set in order to prevent misentries in advance.

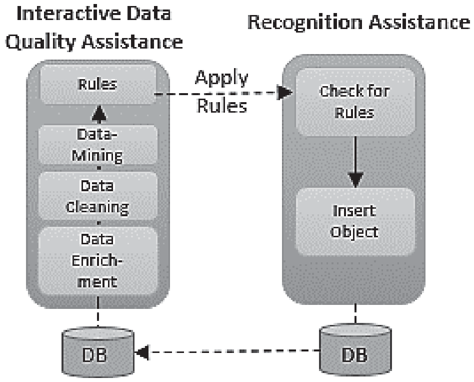


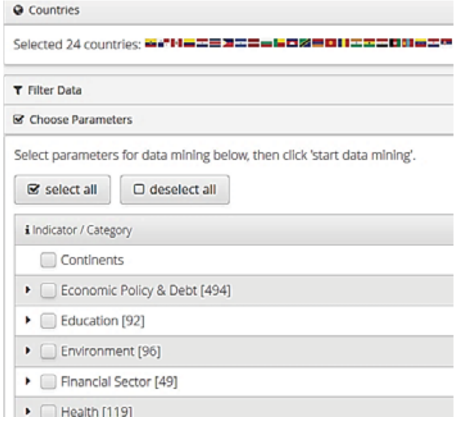
Fig. 1. Data quality assurance process

Figure 1 illustrates the general process of the data quality assurance. The data set we regard to analyze is the data set that already contains objects from the assistance system. The objects in this case are the countries, the features of the objects are the indicators that describe them. First, the data needs to be enriched. The data set of WDI e.g. did not contain the continents within the data point of a single country.

Therefore we enriched every country with the corresponding continent to be able to group them afterwards also by their continents. The Data Cleaning includes that empty data points are deleted from the data set that are considered into the data mining analysis. The next stage is to apply the data mining algorithm to find specific rules. The last step is to visualize the found rules for the analyst in a structured way. A strong correlation does not always imply causality and therefore the found rules within the data set need to be reviewed by the analyst before they can be applied to the whole data set as guideline.

2.1 Structure of the underlying data set

Figure 2 illustrates the structure of the underlying WDI data set in the interactive data quality system. In this case the countries are our objects symbolized with flags and the indicators listed below are the features of the objects.



Countries

Selected 24 countries:

Filter Data

☑ Choose Parameters

Select parameters for data mining below, then click 'start data mining'.

☑ select all ☐ deselect all

Indicator / Category
☐ Continents
☐ Economic Policy & Debt [494]
☐ Education [92]
☐ Environment [96]
☐ Financial Sector [49]
☐ Health [119]

Fig. 2. Structure of the WDI data (Screenshot of the interactive data quality assistance system)

For example the country Colombia contains data points for the indicator “Education”.

2.2 User interface

The user interface is split up into four different sections. The first section “Filter countries” offers the analyst the possibility to filter the data regarding a specific indicator. In case of the WDI data set it is possible to filter between specific ranges of numbers, e.g., within the “GDP per capita”. This function is quite useful to find rules only within a specific filtered data set. The second section “Choose parameters for data mining” offers the analyst the possibility to only consider the relevant indicators for the data mining analysis. Although it is possible to consider all indicators for data mining. The section “Navigation and Results” provides the opportunity to navigate through the decision tree by selecting the provided intervals. In decision trees, the data is represented by a hierarchal tree, where each leaf refers to a concept [2]. Furthermore the section shows the diagrams of the selected rule and the corresponding countries. The intervals are generated automatically on the basis of information gain. Furthermore it displays all rules that are found within the data set. The section “Countries” illustrates which countries match with the selected rule symbolized by the national flags.

2.3 Applied algorithm

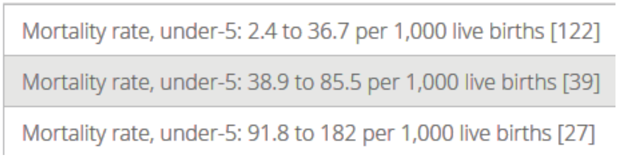
The underlying algorithm we used is k-means clustering with the application of the Lloyd algorithm. Clustering in general means partitioning a group of data points into different arrays. K-means clustering is a method to automatically partition a data set into k groups. The application of the Lloyd algorithm is efficient and resulted in the optimal solution for our specific problem to find rules that serve as guidelines within the data set. The algorithm has four major steps that need to be done in order to cluster each data point [3] [4]:

1. Initialize the centroids of the clusters
2. Search for every data point the closest cluster
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat the steps 2 and 3 until convergence

After performing the algorithm on the specified data set each object, in this case the countries, are assigned to a specific cluster. In order to navigate through the decision tree containing the different indicators, we split them automatically on the basis of the information gain. The information gain is predicated up on the reduced entropy after a data set is split up on an indicator. The major task during building up a decision tree is to find the attribute that returns the highest information gain. So in a first instance the entropy needs to be calculated. The second step is to split the data set on the different indicators. After this, the entropy for each branch is computed and added proportionally to get the entire entropy for the split. This entropy is then subtracted from the entropy before the split. The outcome is the information gain.

2.4 Example

In this section we want to illustrate an example rule generated by the interactive data quality assistance system. To control the variety of rules in advance we chose the parameters “GDP per capita” and „ Mortality rate, under -5 (per 1,000 live births) “for the data mining algorithm. Under five years mortality rate is the probability per 1000 that a newborn baby will die before reaching age five [1]. After starting the data mining algorithm, the highest information gain is found at „Mortality rate, under -5 (per 1,000 live births)”. Figure 3 illustrates the separation for „Mortality rate, under -5 (per 1,000 live births)”.



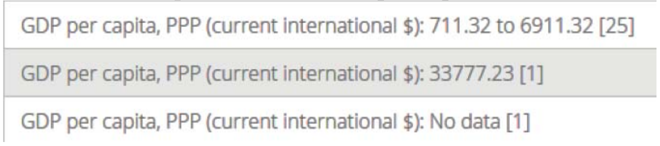
Mortality rate, under-5: 2.4 to 36.7 per 1,000 live births [122]

Mortality rate, under-5: 38.9 to 85.5 per 1,000 live births [39]

Mortality rate, under-5: 91.8 to 182 per 1,000 live births [27]

Fig. 3. Separation for „Mortality rate, under -5 (per 1,000 live births)”.

In the next step we chose the interval with the highest mortality rate. After choosing this interval the next recommended separation is at “GDP per capita”. Figure 4 illustrates the separation for “GDP per capita”.



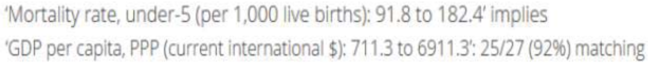
GDP per capita, PPP (current international $): 711.32 to 6911.32 [25]

GDP per capita, PPP (current international $): 33777.23 [1]

GDP per capita, PPP (current international $): No data [1]

Fig. 4. Separation for „GDP per capita”.

Countries with a high “Mortality rate, under -5 (per 1,000 live births)” seem to have a lower “GDP per capita”. The derived rule is then:



‘Mortality rate, under-5 (per 1,000 live births): 91.8 to 182.4’ implies
‘GDP per capita, PPP (current international $): 711.3 to 6911.3’: 25/27 (92%) matching

Fig. 5. Derived rule from data mining algorithm.

Figure 5 illustrates for this derived rule a 92 per cent match for the countries within.

3 Conclusions

At the first stage we build up the interactive data quality assistance system with the underlying data set of the WDI. The system delivers rules that are supposedly general valid for the data. The next step will be to substitute the data set with the military data set and to apply the data mining algorithms on it. We need this step to be able to perform pilot studies with experts to improve the correctness of the derived rules and to compare different algorithms and the results. Furthermore this derived rules after being checked by the experts then will be applied as a guideline to the recognition assistance system while inserting a new object.

4 Acknowledgment

The underlying project to this article is funded by the WTD 81 of the German Federal Ministry of Defense. The authors are responsible for the content of this article.

5 References

[1] http://data.worldbank.org/data-catalog/world-development-indicators.

[2] Maimon, O., Rokach L. Data Mining and Knowledge Discovery Handbook. Springer Sciene + Business Media, 2010, pp. 284.

[3] Faber, V. Clustering and the continuous k-means algorithm. Los Alamos Science, 1994, pp. 140–142.

[4] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability, Berkeley, CA: University of California Press, 1967, pp. 281-297.