Towards a reverse engineering approach for guiding user in applying data mining*

Roberto Espinosa[1], Jose-Norberto Mazón[2], and José Zubcoff[2]

[1] Lucentia, University of Matanzas, Cuba
respinosa@umcc.cu
[2] Lucentia, DLSI/DCMBA, University of Alicante, Spain
{jnmazon,jose.zubcoff}@ua.es

Abstract. Data mining is at the core of the knowledge discovery process. However, an initial preprocessing step is crucial for assuring reliable results within this process. Preprocessing of data is a time-consuming and non-trivial task since data quality issues should be considered. This is even worst when dealing with complex data, not only because of the different kind of complex data types (XML, multimedia, and so on), but also because of the high dimensionality of complex data. Therefore, to overcome this situation, in this position paper we propose using mechanisms based on data reverse engineering for automatically measuring some data quality criteria on the data sources. These measures will guide user in selecting the most adequate data mining algorithm in the early stages of the knowledge discovery process. Finally, it is worth noting that this work is a first step towards considering, in a systematic and structured manner, data quality criteria for supporting data miners in applying those algorithms that obtain the most reliable knowledge from the available data sources.

1 Introduction

According to the seminal work in [3], data mining is the process of applying data analysis and discovery algorithms to find knowledge patterns over a collection of data. Importantly, the same authors explain that data mining is only a step of an overall process named knowledge discovery in databases (KDD). KDD consists of using databases in order to apply data mining to a set of already preprocessed data and also to evaluate the resulting patterns for extracting the knowledge. Indeed, the importance of the preprocessing task should be highlighted due to the fact that (i) it has a significant impact on the quality of the results of the applied data mining algorithms [10], and (ii) it requires significantly more effort

* This work has been partially supported by the following projects: MANTRA (GV/2011/035) from Valencia Ministry, MANTRA (GRE09-17) from the University of Alicante, SERENIDAD (PEII-11-0327-7035) from Junta de Comunidades de Castilla La Mancha (Spain) and by the MESOLAP (TIN2010-14860) project from the Spanish Ministry of Education and Science.

than the data mining task itself [7]. When mining complex data the preprocessing task is even more time-consuming, not only because of the different kind of complex data types (XML, multimedia, and so on), but also because of the high dimensionality of complex data [10]. High dimensionality means a great amount of attributes difficult to be manually handled and making the KDD awkward for non-experts data miners. Specifically, high dimensionality implies several data quality criteria to deal with in the data sources, such as incomplete, correlated and unbalanced data. For example, incomplete data could easily appear in complex data such as microarray gene expression [16] or in data coming from sensors [6]. Several statistical techniques have been proposed to deal with dimensionality reduction issue [4], such as PCA (Principal Component Analysis) or Regression Trees, among others. However, by using those techniques there are an important information lost: the functional dependencies and data structure. To overcome this situation, in [10] the definition of user-friendly data mining applications is suggested. To this aim, data preprocessing should be automated, and all steps undertaken should be reported to the user or even interactively controlled by the user, at the same time that useful information is not lost.

Bearing these considerations in mind, in this position paper we propose mechanisms based on data reverse engineering for automatically measuring some data quality criteria (e.g., completeness, correlation or balance) on the data sources in order to guide user in selecting the most adequate data mining algorithm in the early stages of the knowledge discovery process. It is worth noting that our focus is on data quality criteria different from those related to the cleaning process. Therefore, our approach assumes that data is already cleaned.

Finally, it is worth noting that this work is a first step towards considering, in a systematic and structured manner, data quality criteria for supporting data miners in applying algorithms for obtaining the most reliable knowledge from the available data sources.

2 Related work

To the best of our knowledge, it is the first attempt to consider data reverse engineering for guiding user in selecting the best data mining algorithm based on data quality criteria. However, data reverse engineering has been used in other related fields such as data warehousing. There are several approaches [5, 8, 2, 11] that suggest mechanisms, as algorithms or guidelines, to specify the schema of the data warehouse starting from an operational database described by an Entity-Relationship (ER) model. However, these mechanisms have to be manually applied, thus resulting costly to apply when the designer is not an expert in the domain. Only in the algorithm proposed in [13], the level of automation to discover elements for the data warehouse schema in an ER model has been increased. Apart from the level of automatization, every of these current approaches presents a major drawback: it is assumed that well-documented diagrams of data sources are available. Unfortunately, the operational data sources are usually real legacy systems and the documentation is not generally avail-

able or it can not be obtained [1]. Moreover, if the data sources are complex, although the documentation exists it may be not easily understandable. Therefore, the application of these approaches is unfeasible if data sources are too large and complex, even though if expert designers in the application domain take part in the development process.

One approach that tries to ameliorate the above-presented problems has been proposed in [9] where a set of algorithms is presented to automatically discover data warehouse structures in the data sources. Furthermore, this approach suggests a "metadata obtainer" as a reverse engineering stage in which relational metadata is obtained from data sources.

Therefore, in this position paper, our hypothesis is that data reverse engineering techniques can be used in order to automatically (i) obtain a common representation of data sources, and (ii) measure data quality criteria on the data sources. Our approach aims to guide user in selecting the most adequate data mining algorithm in the early stages of the KDD process.

3 Data reverse engineering for guiding user in applying data mining

This section describes our approach for data reverse engineering in preprocessing data for guiding user in applying mining algorithms. It consists of two steps: (i) creating a common representation of the data sources, and (ii) measuring data quality criteria of data sources and adding them to the common representation. As shown in Fig. 1, our approach aims to give advice to data miners for selecting the most appropriate data mining algorithm for the available data sources.

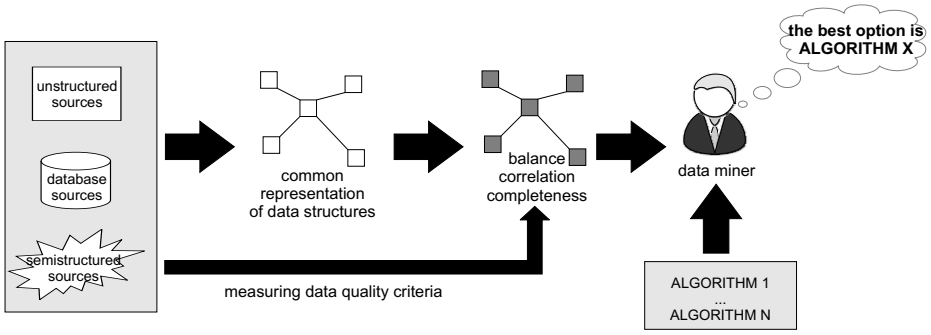

Fig. 1: Overview of our data reverse engineering approach for data mining

3.1 Defining a common representation

Several standards could be used for representing heterogeneous data sources metadata in a common format. For example, the *Common Warehouse Meta-*

model (CWM) [12] consists of a set of metamodels that allow up to represent data structures and related information. In common scenarios, data sources can be implemented according to a plethora of techniques, such as relational databases, XML files or even text files. Depending on the kind of data sources a specific metamodel of CWM can be used in order to allow elements to be represented in a model. This model will guide user in applying preprocessing algorithms for dealing with data quality criteria or even for choosing the most adequate data mining algorithm.

3.2 Measuring data quality

Data quality means "fitness for use" [14] which implies that the data should accomplish several requirements to be suitable for a specific task in a certain context. There are several data quality criteria which should be measured to determine the suitability of data for being used [15]. In KDD, this means that data sources should be useful for discovering reliable knowledge when data mining techniques are being applied. Our hypothesis is therefore, that data miner must measure data quality criteria to avoid discovering superfluous, contradictory or spurious knowledge. This is specially true for high dimensional data, since a non-expert data miner without knowing in detail the domain of data can apply a data mining technique that provides misleading results. For example, if some attributes are selected as input for a classification algorithm (being some of them strongly correlated), the resulting knowledge pattern, though correct, will not provide the useful expected value. Therefore, those data quality criteria that may affect the result of data mining techniques should be determined in order to be reported to the data miner in early stages of the design.

Data quality criteria for data mining are related to two important preprocessing issues that should be addressed in data mining [3]. On one hand, missing or noisy data should be considered (which is related to the completeness data quality criteria), and, on the other hand, complex relationships among data should be detected (which is related to correlation and balance). These three quality criteria can be measured from the data sources and stored in the *CWM* model previously obtained. This model can be then used by data miners for supporting their decisions on what data miner algorithms should be applied, thus increasing the success of the knowledge discovery.

Completeness For our purpose, completeness is defined as the percentage of non-null values taken by certain attribute in the data sources. For example, in relational data sources, for each column in every table in the database, the amount of non-null and null values are obtained. Therefore, completeness could be computed for each column "columnName" of each table "tableName" by using the SQL code in query 1.1.

```
1 select count(*) as TotalValues,
2 (select count(*) from 'tableName'
3 where 'columnName' Is Null) as NullValues,
```

```
4 100 - (NullValues * 100 /TotalValues) as NonNullPercentage
5 from 'tableName';
```

Query 1.1: Query for computed Non-null Percentage

Each corresponding element in the *CWM* model can store this value.

Correlation Two attributes are correlated if changes in the value of an attribute are associated with changes in another attribute, thus being a measure of association between two attributes. For example, the Pearson correlation coefficient can be implemented in our approach by means of the following formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

where

- r_{xy}: it is the correlation coefficient. It gives a value between $+1$ and -1 inclusive. A value of 1 implies that x and y are perfectly correlated (y increases as x increases), while a value of -1 implies that y decreases as x increases. A value of 0 implies that there is no correlation between x and y.
- x and y are two variables from which the correlation is measured.
- n is the number of values to be considered.

The value of the correlation coefficient of each attribute with regard to the other attribute can be stored in the *CWM* model.

Balance Data stored in a certain attribute are balanced if the number of different values representing each different instance are significantly equal. This means that similar numbers of instances are expected for each value. To know how balanced data are, a method that returns the Chi-square for each attribute can be implemented. Then, a statistic Chi-square test can be performed to know if the instances are uniformly distributed. In this case, the null hypothesis is that all positions have equal (similar) number of instances. Then, the data would be uniformly distributed. The alternative hypothesis therefore states they are different. The level of significance (the point at which you can say with 95% of confidence that the difference is not due to chance alone) is set at 0.05. The Chi-square formula is as follows:

$$\chi^2_{obs} = \sum_{i=1}^{n} \frac{(f_i - np_i)^2}{np_i}$$

where

- χ^2_{obs}: calculated value from the sample for the Chi-square statistic.
- f_i: number of observed frequencies, i.e. number of instances per category.
- p_i: number of expected frequencies, i.e. the values estimated by a uniform distribution of the instances per category.
- n is the number of categories to be considered.

Each attribute in the *CWM* model can store the value of Chi-square.

4 Conclusions

In this position paper, an approach based on data reverse engineering is proposed for automatically measuring data quality criteria (e.g., completeness, correlation or balance) on the data sources in order to support user in selecting the most adequate data mining algorithm in the early stages of the knowledge discovery process. This work intends to be a first step towards considering, in a systematic and structured manner, data quality criteria for supporting data miners for obtaining reliable knowledge. Finally, our short-term future work consists of conducting several experiments to show the feasibility of our approach.

References

1. Alhajj, R.: Extracting the extended entity-relationship model from a legacy relational database. Inf. Syst. 28(6), 597–618 (2003)
2. Böhnlein, M., vom Ende, A.U.: Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems. In: DOLAP. pp. 15–21 (1999)
3. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: Towards a unifying framework. In: KDD. pp. 82–88 (1996)
4. Fodor, I.K.: A survey of dimension reduction techniques. LLNL technical report (June), 1–24 (2002)
5. Golfarelli, M., Maio, D., Rizzi, S.: The Dimensional Fact Model: A conceptual model for data warehouses. Int. J. Cooperative Inf. Syst. 7(2-3), 215–247 (1998)
6. Halatchev, M., Gruenwald, L.: Estimating missing values in related sensor data streams. In: COMAD. pp. 83–94 (2005)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)
8. Hüsemann, B., Lechtenbörger, J., Vossen, G.: Conceptual data warehouse modeling. In: DMDW. p. 6 (2000)
9. Jensen, M.R., Holmgren, T., Pedersen, T.B.: Discovering multidimensional structure in relational data. In: DaWaK. pp. 138–148 (2004)
10. Kriegel, H.P., Borgwardt, K.M., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A.: Future trends in data mining. Data Min. Knowl. Discov. 15(1), 87–97 (2007)
11. Moody, D.L., Kortink, M.A.R.: From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In: DMDW. p. 5 (2000)
12. Object Management Group: Common Warehouse Metamodel Specification 1.1. `http://www.omg.org/cgi-bin/doc?formal/03-03-02`
13. Phipps, C., Davis, K.C.: Automating data warehouse conceptual schema design and evaluation. In: DMDW. pp. 23–32 (2002)
14. Strong, D.M., Lee, Y.W., Wang, R.Y.: 10 potholes in the road to information quality. IEEE Computer 30(8), 38–46 (1997)
15. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM 40(5), 103–110 (1997)
16. Troyanskaya, O.G., Cantor, M., Sherlock, G., Brown, P.O., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for dna microarrays. Bioinformatics 17(6), 520–525 (2001)