A Comprehensive Methodology to Implement Business Intelligence and Analytics Through Knowledge Discovery in Databases

Fernando Paulo Belfo[1(✉)] and Alina Banca Andreica[2]

[1] Polytechnic Institute of Coimbra,
ISCAC Coimbra Business School, Coimbra, Portugal
fpbelfo@gmail.com
[2] Faculty of European Studies, Babes-Bolyai University of Cluj-Napoca,
Cluj-Napoca, Romania
alina.andreica@ubbcluj.ro

Abstract. Business intelligence is used by companies for analysing business information, providing not only historical or current views on business operations, but also providing predictions about the business. Consequently, knowledge discovery in databases can support the implementation of business intelligence solutions, especially in order to deal with the reality of big data, using diverse data mining techniques that can help to better prepare the data and to create improved models. The current paper proposes a methodology to implement business intelligence and analytics solutions, based on the CRISP-DM methodology, where the application of simplification and equivalence algorithms in modelling data representations can be used for improving the process of business management. This promising approach can boost business intelligence and analytics by using alternative techniques for discovering and presenting new knowledge about the business. The application of simplification and equivalence algorithms within the business context enables finding the most comprehensive or relevant knowledge, represented for instance as association rules, and bringing a real competitive advantage for the stakeholders.

Keywords: Business intelligence · Knowledge discovery in databases
Data mining · Equivalence algorithm · Canonical representation

1 Introduction

Data mining is one of the steps that compose the process of knowledge discovery in databases (KDD). Data mining can be defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in stored data within structured databases. The most important problem addressed by the KDD process is the one of mapping low-level data (usually too big to easily understand) into other forms that might be more compact [1, 2].

Various economic activities have increasingly developed projects of knowledge discovery in databases to solve different problems or to provide them with previously unknown opportunities of finding solutions. Among such economic areas, we can

A. Groza and R. Prasath (Eds.): MIKE 2018, LNAI 11308, pp. 102–111, 2018.
https://doi.org/10.1007/978-3-030-05918-7_10

highlight examples from the financial area [3], the insurance area [4], the accommodation and catering [5], the academic area [6] or the medical field [7]. These knowledge discovery techniques helped the organizations which developed such projects to define strategies that contributed to increasing their performance.

Data mining applications are vast and implementing such techniques in business gains more and more impact. Furthermore, data mining techniques can be combined with business intelligence and analytics (BI&A).

Business intelligence (BI) encompasses the technologies and strategies that are used by companies for the data analysis of business information, providing not only historical or current views of business operations, but also providing prediction about the business. Business intelligence commonly uses technologies like data mining, text mining, process mining or complex event processing. Since the 1990's, business intelligence and analytics and the related field of big data analytics have also become increasingly important and have consequently evolved [8].

The conjunction of these two areas (data mining and business intelligence and analytics) continues to be promising. Data mining allows discovering patterns within data, like association, prediction, segmentation or sequential relationships.

Each type of these datamining tasks may use several types of techniques and popular algorithms. For example, discovering association rules (link analysis) is usually performed using an algorithm like the popular Apriory. For instance, imagine that there is a datamining project with the goal of determining certain association rules among customers, rules which contribute to better defining customer credit. In this respect, a sample of customers of a commercial bank is used. One possible association rule derived from the sample of customers used is represented below and has a confidence of 100% and a support of 36%.

$$IF\ competence = high\ AND\ personality = good\ THEN\ credit\_history = good \quad (1)$$

Yet, traditional association rules are somewhat rigid because they are anchored on the specific instances of attributes and their relations within the available business databases. Simplification and equivalence algorithms bring a more flexible perspective within the businesses realities.

The motivation behind this paper is that equivalence algorithms and canonical representation can present an interesting alternative way to help discovering and presenting more comprehensive knowledge, like enhanced association rules among businesses data. They can contribute to discovering common customer similar (and not necessarily equal) behaviours or characteristics. These operations may be performed using structured content, organized in the database management systems of companies (BI&A 1.0) or by searching web-based content, typically an unstructured content (BI&A 2.0) or by using mobile and sensor based content (BI&A 3.0) [8].

The current paper proposes the application of simplification and equivalence algorithms in modelling data representations for improving business management either at an operational, or at a tactical or strategic level. This promising approach could boost business intelligence and analytics by using alternative techniques for discovering and presenting business knowledge. We propose a specific application of

equivalence algorithms within the business context in order to find the most comprehensive knowledge, like association rules, and to represent them for the stakeholders.

2 A Methodology to Implement Business Intelligence and Analytics

Based on the CRISP-DM methodology, an acronym for "Cross-Industry Standard Process for Data Mining" [9], we propose a methodology for implementing business intelligence and analytics that is represented on Fig. 1. Our proposal of business intelligence and analytics (BI&A) implementation has a dynamic perspective. It is based on the idea that advanced BI&A may have a model (or several models) to support the creation of novel and powerful views of business operations. This model (or models) that support the BI&A system may be implemented with the support of a KDD process (or several processes if several models are necessary). This section will explain it better.

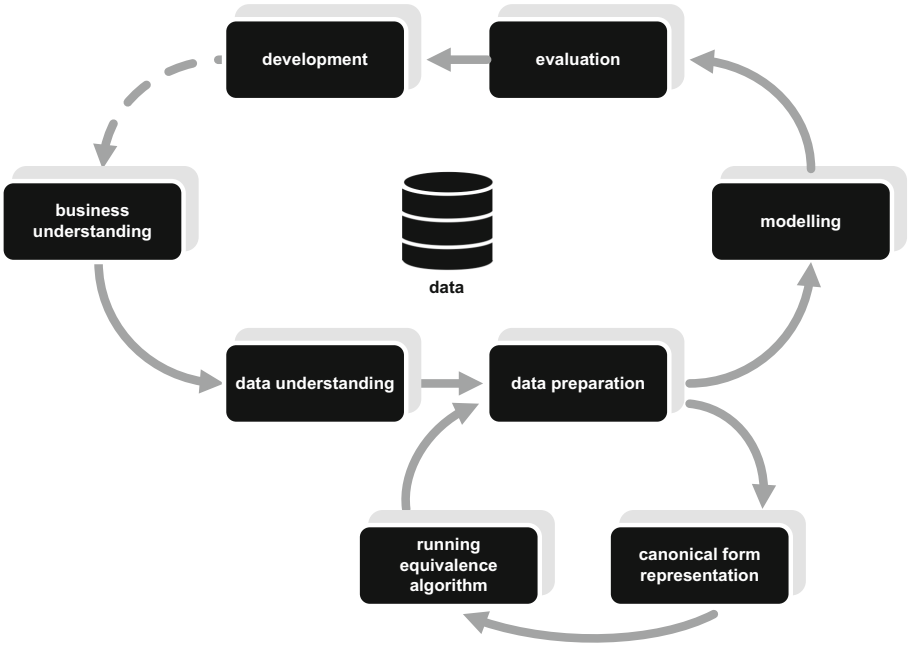

Fig. 1. Proposed methodology to implement business intelligence and analytics.

2.1 From Business Understanding to Development

Based on the CRISP-DM methodology, an acronym for "Cross-Industry Standard Process for Data Mining" [9], we propose the following methodology for implementing business intelligence and analytics - see Fig. 1. Business intelligence and analytics may be viewed as a sequence of processes including: business understanding, data

understanding, capture and preparation, determining the canonical form representation and running the equivalence algorithms, data mining modelling and development, testing and evaluating the results – see Fig. 1.

As proposed within CRISP-DM methodology [9], the initial phase of this methodology focuses on understanding the project objectives and requirements from a business perspective. As this new methodology intends to support the implementation of advanced business intelligence solutions, it proposes the conversion of the knowledge into a data mining problem definition.

The second phase is the data understanding phase. It starts with an initial data collection and proceeds with activities that will help to better know the data, to identify potential quality problems within it, to discover first perceptions on data and formulate initial hypotheses.

The data preparation represents the third phase. It covers all the activities to construct the final dataset that will feed the modelling tool. Data preparation tasks are usually performed several times, whenever they are needed. The data preparation phase includes tasks like the selection of tables, records and attributes, as well as transformation and data 'cleaning'. Our methodology proposes the application of canonical representation and equivalence algorithms as a mean of obtaining improved data, which will allow obtaining enriched models. These two steps will be further explained.

The modelling phase is the next phase and uses several techniques, as further described. These techniques are selected and applied, while their parameters are calibrated to optimal values. After the data mining problem type has been defined, one or more tasks and techniques can be selected among the candidates for that type of problem. The main tasks are classification, segmentation, prediction or association. Then, one algorithm from the set of available ones should be chosen in order to perform the modelling. Usually, data preparation has to be improved, therefore, often the process should resume to that previous phase.

After building the model (or the models), it (they) should be evaluated. Within this phase it is important to carefully assess the model. The steps executed prior to the construction of the model can be reviewed in order to ensure that the business objectives are properly achieved. The main output of this phase concerns a decision regarding the use of the data mining results. If these results are satisfying, they can be further used in the development of the business intelligence solution. If not, the algorithm goes back to the modelling phase, or sometimes, even to the beginning of the project, reformulating the preliminary phases as well.

The last phase of the process is the development of the business intelligence and analytics solution. The process only ends after the knowledge that was discovered is organized and presented in a way that the customer can use it. This is the final goal of business intelligence. There are many types of organizing the way in which the data is presented to the user. It may involve applying "live" models within an organization's decision making processes, for example aggregating real-time analysis of customers' behaviour or presenting real-time and predictive analytics for the operational activity of a factory's shop floor. The choice of an adequate visualization technique is very important. For example, if the business intelligence is supported on the exploration of association rules, there are some innovative ways of doing it [10]. Figure 2 presents one of these examples.

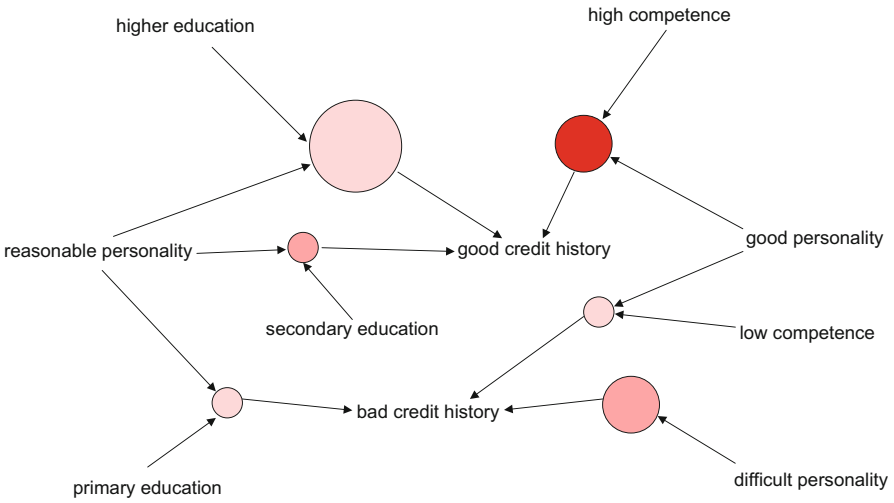

Fig. 2. Graph-based visualization with items and rules as vertices.

Although the proposed methodology is inspired from the CRISP-DM, main differences concern the data preparation and the development phases of the BI&A solution. The data preparation is the step that provides this methodology with a more comprehensive approach. Data preparation will be supported by canonical form representation and running the equivalence algorithm. The development phase involves specific aspects related to the construction of an artefact like a business intelligence and analytics solution.

2.2 Deepening the Data Preparation and Modelling

The data mining phase consists of choosing data mining tasks and corresponding algorithms. Its tasks may belong to predictive or descriptive types. Predictive (or supervised) activities learn decision criteria in order to be able to classify unknown cases (future trends) using the knowledge acquired from a set of samples which belong to already known classes. Descriptive (or unsupervised) activities work with a set of data that does not have a given output class seeking to identify common unknown patterns in these data. The most important predictive tasks are classification, prevision and tendency analysis, while the most important descriptive tasks are clustering, association, summarization and visualization [11–14].

Let's look again at the association rule example that was previously presented in Sect. 1. Within this rule, we can observe that if the bank customer has a high competence and if the customer's personality is good, then, most probably (with a confidence of 100% and a support of 36%), this would mean that he customer has a good credit history.

Within a traditional KDD project, conducted under the CRISP-DM model, it is common that the data preparation phase includes tasks like new data construction. Yet, this construction is usually composed by very simple and direct procedures. For

example, the creation of a new attribute for the age, computed using the customer's date of birth.

Within the previously presented example, the customers of the commercial bank could be previously classified with a specific personal competence that could be low, medium or high. This classification is traditionally already performed before the KDD project starts. After the data preparation phase, there should be a unique table where the chosen algorithm, for example, the Apriori algorithm, will run and using the necessary attributes.

Instead of previously defining a customer as having a high competence, the canonical form representation and the equivalence algorithm may help evaluating it. This computation will be performed within the data preparation phase, specifically within the step of constructing new data. The canonical form representation defines entities' equivalences that can help to classify old and new objects. After the canonical form representation is computed, the equivalence algorithm will run and help to define the personal competence of bank customers as low or high.

2.3 Canonical Form Representation

Within this section we overview the methods we propose [15, 16] for implementing equivalence algorithms at the database level for various entities and entity classes, including hierarchical structures, the entities being retained in database tables. This solution is very useful, since the volume of data that has to be processed is often very large, being consequently retained in databases.

The equivalence algorithms we have implemented are based on the theoretical framework given in [17] and the principles for representing) and processing hierarchical are presented in [16].

We have implemented post order type n-ary tree algorithms using the pointer tree representation [16] in order to parse the hierarchy of entities. We proved that the function implementing the equivalence test for two entities based on the property that they have the same canonical representative is much more efficient than the one using the definition [16].

The categories of entities and the principles for implementing equivalence algorithms on these categories are also described in [15, 16]. Intuitively, the canonical set of a category of entities may be obtained by "flattening" its category sub-tree and computing the union set of all canonical sets if its descendant leaf entities. In the case of categories of entities, the canonical representative is recursively computed [18].

In many practical cases, entity equivalences or even canonical elements have to be mapped, sometime with user assistance. In [18, 19], we address mappings and pattern matching issues. We overview here the principles proposed in this respect.

We designed pattern matching rules for equivalent entities by reducing the mapping between two elements, belonging to the two equivalence classes that are to be mapped, to mapping their canonical representatives [18, 19]. The formalization is given in [18, 19].

Intuitively, instead of dealing with the equivalence of any elements between two classes, we reduce the problem to the simpler one of dealing with the equivalence of the two classes' representatives – see Fig. 3.

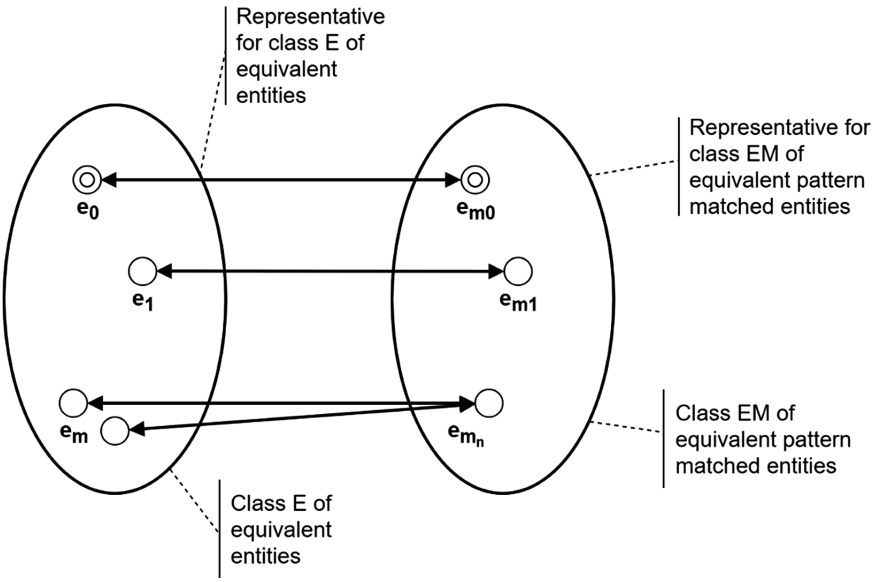

Fig. 3. Pattern Matching scheme for Equivalence classes [18].

The canonical set of a category of entities or events may be achieved by "flattening" its category sub-tree and computing the union set of all canonical sets of its descendant leaf entities. The process of obtaining the canonical representative of categories of entities or events is recursively computed [18, 20].

2.4 Principles for Building Associations and Running Equivalence Algorithm

We have previously addresses the topic of building the cloud database representation in [18, 19]. Within this section, we describe the way in which we can use a canonical for in order to perform associations within the database that is processed.

Let us suppose that in the database we have S_1 and S_2 equivalent sets of data with the canonical representation S and S_2 and S_3 equivalent sets of data that have also been mapped into the same canonical representation S, then sets S_1 and S_3 do not have to be further checked if they are equivalent, since they have the same canonical representative S. The mapping and equivalence association scheme is graphically represented in Fig. 4 [18].

Instead of using specific instances that exist on each attributes, equivalence algorithms allow to combine detailed cases on to several types of entities or events. The principles behind implementing equivalence algorithms on categories of entities [15, 16] may also be used to implement equivalence on categories of any types of information objects as entities, like customers or products, or events, like travels or purchases.

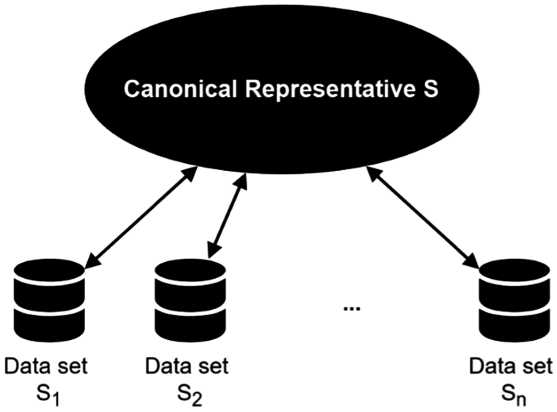

Fig. 4. Building the associations with the canonical representation

3 Conclusions

A project of knowledge discovery in databases is usually triggered by challenges or opportunities from the business environment and, to be truly effective, the business field must properly support and be supported by the information technology tools, based on the fundamental principle of business alignment with information [21, 22]. In fact, KDD projects are good examples of initiatives based on good business alignment with information technology, one of the major concerns of information technology managers in companies in recent years [23]. The equivalence algorithms and the canonical representation give the necessary flexibility to define new business concepts which can be used in order to capture more comprehensive knowledge, like enhanced association rules comparatively to the traditional approach.

The simplification and equivalence algorithms represent an opportunity of boosting business intelligence and analytics through an adequate process of knowledge discovery in databases. This paper presents a methodology to implement business intelligence and analytics solutions using the approach of knowledge discovery in databases processes and incorporating the advantages of canonical form representation and of the equivalence algorithms. Using the proposed methodology, the implementers of BI&A solutions may better prepare the data that will be used in the modelling phase and that supports the BI&A system.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
2. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. Springer, Heidelberg (1996). https://doi.org/10.1007/978-3-319-93040-4

3. Ngai, E., Hu, Y., Wong, Y., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decis. Support Syst. **50**(3), 559–569 (2011)
4. Azadmanesh, S., Tarokh, M.J.: Labeling customers using discovered knowledge case study: automobile industry. Int. J. Manag. Value Supply Chain. (IJMVSC) **3**(3), 13–24 (2012)
5. Loureiro, A., Lourenço, J., Costa, E., Belfo, F.: Indução de Árvores de Decisão na Descoberta de Conhecimento: Caso de Empresa de Organização de Eventos. In: VI Congresso Internacional de Casos Docentes em Marketing Público e Não Lucrativo. ISCAC Business School, Coimbra, Portugal (2014)
6. Pimenta, C., Ribeiro, R., Sá, V., Belfo, F.P.: Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa. In: 18ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2018) Associação Portuguesa de Sistemas de Informação: Santarém, Portugal (2018)
7. Cios, K.J., Moore, G.W.: Uniqueness of medical data mining. Artif. Intell. Med. **26**(1–2), 1–24 (2002)
8. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**, 1165–1188 (2012)
9. Chapman, P., et al.: CRISP-DM 1.0: Step-By-Step Data Mining Guide. SPSS, CRISP-DM Consortium: U.S.A (2000)
10. Hahsler, M., Chelluboina, S.: Visualizing association rules: introduction to the R-extension package arulesViz. R project module, pp. 223–238 (2011)
11. Galvão, N.D., Marin, H.D.F.: Data mining: a literature review. Acta Paulista de Enfermagem **22**(5), 686–690 (2009)
12. Berry, M.J., Linoff, G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. Wiley, New York (1997)
13. Fu, Y.: Data mining: tasks, techniques, and applications. IEEE Potentials **16**(4), 18–20 (1997)
14. Kivikunnas, S.: Overview of process trend analysis methods and applications. In: ERUDIT Workshop on Applications in Pulp and Paper Industry. Citeseer (1998)
15. Andreica, A., Stuparu, D., Miu, C.: Design techniques in processing hierarchical structures at database level. In: Proceedings of Iadis Information Systems, pp. 483–488 (2010)
16. Andreica, A., Stuparu, D., Miu, C.: Applying mathematical models in software design. In: 2012 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE (2012)
17. Buchberger, B., Loos, R.: Algebraic simplification. In: Buchberger, B., Collins, G.E., Loos, R., Albrecht, R. (eds.) Computer Algebra, pp. 11–43. Springer, Vienna (1982). https://doi.org/10.1007/978-3-7091-7551-4_2
18. Andreica, A.: Designing uniform database representations for cloud data interchange services. In: Proceedings of CLOSER 2017 - 7th International Conference on Cloud Computing and Services Science, pp. 554–559 (2017)
19. Andreica, A.: Applying Equivalence Algorithms in Solving Pattern Matching Problems. Case Study for Expert System Design, p. 255. ICT, Society, and Human Beings (2016)
20. Andreica, A., Belfo, F.: Building cloud data interchange services for E-learning systems: applications on the moodle system. In: Proceedings of the 8th International Conference on Cloud Computing and Services Science (CLOSER 2018), pp. 565–572 (2018)

21. Reich, B.H., Benbasat, I.: Measuring the linkage between business and information technology objectives. MIS Q. **20**(1), 55–81 (1996)
22. Belfo, F., Sousa, R.D.: Reviewing business-IT alignment instruments under SAM dimensions. Int. J. Inf. Commun. Technol. Hum. Dev. **5**(3), 18–40 (2013)
23. Kappelman, L., et al.: The 2016 SIM IT Issues and Trends Study. MIS Q. Exec. **16**(1), 47–80 (2017)