Data Preparation: Art or Science?

Gunjan Mansingh
Department of Computing
The University of the West Indies
Kingston, Jamaica
gunjan.mansingh@uwimona.edu.jm

Kweku-Muata Osei-Bryson
School of Business
Virginia Commonwealth University
Richmond, U.S.A.
kmuata@isy.vcu.edu

Lila Rao
Mona School of Business and Management
The University of the West Indies
Kingston, Jamaica
lila.rao@uwimona.edu.jm

Maurice McNaughton
Mona School of Business and Management
The University of the West Indies
Kingston, Jamaica
maurice.mcnaughton@uwimona.edu.jm

***Abstract*—Data preparation is often cited as the most time consuming phase of a Knowledge Discovery and Data Mining (KDDM) process. This is attributed to the fact that this phase is highly dependent on the expertise of the analyst. Although process models exist for KDDM the description of their phases of the process focus on outlining what must be done but often do not detail how this should be done. While there is some research in addressing the how of the phases, the data preparation phase is thought to be the most challenging and is often described as an art rather than a science. The tasks defined in this phase are thought to be highly dependent on the expertise of the analyst and the context. While we are of the view that there will always be an art to data preparation we will demonstrate that the science can actually enhance the art. We further contend that as more research of this kind is published, that demonstrates a variety of data preparation techniques that enhance the data mining process, the more effective will be the science of data preparation.**

Keywords— IKDDM, Data Mining, Data Preparation

I. INTRODUCTION

Increasingly more organizations are recognizing the need for data-driven decision making to harness knowledge that is hidden in their large data assets. The strategic benefits that data mining or analytics can bring to organizations are numerous and its value increasingly apparent to business and technology leaders. However, these organizations have recognized that these benefits will only be realized if there is a clear understanding of the objectives of the data mining process and of the data itself. As data mining becomes an integral decision making tool it is important that Knowledge Discovery and Data Mining (KDDM) standards are established for this process [1, 2]. Some standards for various aspects of the data mining process are emerging (e.g. models, settings and processes) and using these standards will help to ensure that the process of data mining is reliable and repeatable.

Data mining process models, such as CRISP-DM provide a comprehensive detailed description on what needs to be done in the various stages of the data mining process [3, 4] and provides a structure for organizing the data mining effort by describing the tasks involved in data mining. The phases include understanding the business problem, capturing and understanding data, applying data mining techniques, interpreting results, and deploying the knowledge gained in operations. The tasks and activities that have to be carried out are often presented in a checklist manner. The focus of these process models is on "what" needs to be done and not the "how". These process models have a fragmented approach and lack an integrated view [5]. To address these issues IKDDM (see figure 1) presented an integrated knowledge discovery process model which set out to identify the intra and inter task dependencies between the different phases in the process model. The phases identified by IKDDM are; Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment. Sharma et al. (2012) focused on explicating the task dependencies in the business understanding and the evaluation phases. According to them understanding dependencies in the process is a step towards semi-automation. While a considerable amount of work has been done is understanding the dependencies in the other phases, the ones in data preparation are still not explicit. Even though in most projects 50%-80% of the time and effort is spent on data preparation phase [6] and it remains an art as the tasks in this phase are dependent on the expertise of the analyst. Data preparation process has a lot of variability from one data mining objective to another, often the steps in preparation are not explicitly presented. An objective of a scientific process management is to reduce variability in the process [7]. Hence to reduce the variability in the data preparation phase and to make it scientific it is essential to identify the intra- and inter task dependencies amongst the essential tasks and provide guidance on how they can be carried out. Therefore, in this study we focus on the data preparation phase of the IKDDM process model and identify the task dependencies in this phase with a specific emphasis on the "how" i.e. the steps to prepare the data in a way that the

978-1-5090-1281-7/16/$31.00 ©2016 IEEE

benefits of the specific data mining technique can be maximized.

Given that the data preparation is influenced by the choice of the data mining technique, for the purposes of this paper the focus is on association rule induction and sequential pattern mining. For these techniques we will explicate the steps that should be used in integrating and preparing the data. The steps will be demonstrated by showing how a financial institutions data can be prepared and used for marketing applications.

II. BACKGROUND

A. Overview of the IKDDM model

CRISP_DM and SEMMA are the among the most popular process models that are used by data mining experts to solve real world data mining problems. In these process models "what" needs to be done is clearly articulated, however they lack guidance on "how", therefore an Integrated Knowledge Discovery and Data Mining process model (IKDDM) was developed. IKDDM has the same phases as CRISP-DM but was found to be more effective and efficient in performing the phases of KDDM. IKDDM improves the fragmented approach to the different tasks in the various phases and provides a framework (see table 1) which includes the output of the tasks, the tools that can be used to implement them and an indication as to whether a given task can be a candidate for semi-automation.

The *Business Understanding* (BU) phase is considered to be the most critical phase as the quality of the data mining activity is largely dependent on the choices made in this phase and the *Data Preparation* (DP) phase is the most time consuming phase. In the DP phase identifying the data that is to be included, and how it is transformed and formatted is dependent on the skills of the data mining analyst thus making this phase more of an art than science. In the next section to reduce the variability in the tasks of this phase we identify the task dependencies in DP. Since this phase is dependent on the BU phase, and the choice of data analysis method is made in this phase (BU – Task e), data preparation is guided by the choice of the data analysis method.

B. Data Preparation for Data Mining

The focus of the data preparation phase is on identifying quality data and formatting it appropriately, which can lead to generation of quality patterns by the chosen data mining algorithms [8]. Data preparation generates a dataset smaller than the original dataset but with better quality and relevant data which can significantly improve the efficiency of the modeling phase.

As shown in fig. 1 the data preparation phase focuses on i) Cleaning the data, ii) Construct the data (i.e. create derived variables, discretize where relevant, integrate if necessary), iii) Convert data to the format that the selected tool requires to satisfy the requirements of the given DM tool. It is quite clear that these focus on "what" needs to be done but we recognise that there is not a great deal of literature on "how" it should be done. For example, there are many possibilities in constructing the data whether in terms of the derived variables that can be constructed, the ways in which the data can be discretised and the how the data can be integrated. Converting the data to the format required for the selected tools also requires a great deal of expertise in understanding and identifying the various ways this can be done.

The objective of IKDDM is on providing the "how" thus in this paper we aim to extend this process model by refining the DP phase of IKDDM. This phase plays a pivotal role in maximizing the value that can be harnessed from existing data sources.

III. STEPS IN DATA PROCESSING – DEVELOPING AN INTEGRATED VIEW

The three mains tasks in data preparation phase are clean, construct and format data (see figure 1). These tasks can vary based on the "Data analysis method" from BU phase. In this paper the scope is restricted, primarily to Association rule induction (ARI) and Sequential pattern mining (SPM). For ARI and SPM the features of a well-formulated data mining objective is "itemsets" [9].

An association rule (AR) shows relationships among items in a transaction of a database [10]. In organizations these patterns or rules have been used to improve marketing campaigns by grouping products that target particular market segments. Therefore, while constructing and formatting data the focus is on determining what can be included as items. In many organizations data that is relevant for decision making typically includes both date and demographic variables, as organizations want to know the sequencing among the items in an "itemset" and also the demographic variables that occur more frequently with certain items. Sequencing between items generates sequential patterns instead of association rules and with inclusion of demographic variables and numeric variables their values have to be discretized to be included as items.

With these requirements in mind the identified data sources and their quality reports are examined by the data mining analyst in the data preparation phase to determine what variables can be added as a basket item to a dataset which will be used in the modelling phase.

A typical dataset will contain:

- An id field.
- Sequence number (for sequential patterns).
- Target variable (i.e. items).

For each id there can be multiple items which represent all the items that are associated with a single basket. A sequence number may or may not be present based on whether ARI or SPM is being performed.

The steps in constructing and formatting data for ARI and SPM, are as follows:

TABLE 1. Phases of IKDDM Process Model

Phase	Tasks
Business Understanding (BU)	a) Define research goals and success criteria. b) Use relevant existing theory to identify variables that are likely to be relevant to the phenomena of interest. c) Do preliminary identification of relevant data including sources of the data. d) If relevant: • Use existing theory & extant research to: provide guidance for the development of data collection instrument. • Develop, test & refine data collection instrument. e) Identify specific data analysis methods (e.g. regression, DT induction, Regression Splines, Clustering, Association Rules, structural equation modeling, Data Envelopment Analysis) plus their parameter settings for use in the Modeling (i.e. Data Mining) phase. f) Determine whether available DM software offer adequate facilities for applying the selected data analysis methods. g) Identify the performance measures and elicit from researcher Value Functions that may be relevant for the measures and DM methods (e.g. trapezoidal value functions for Simplicity) h) Elicit Preference Functions from researcher (e.g. weights obtained using the Analytic Hierarchy Process- AHP) that will be used in the Evaluation step for comparing causal models.
Data Understanding (DU)	a) Collect initial data b) Describe data (e.g. the format of the data, number of records and variables in each table, names of the variables) c) Explore data (e.g. determine data distributions using histograms, simple statistical analysis; Find outliers; do Factor Analysis & validity tests; Determine if there are natural Groups) d) Explore relationships between pairs of variables using correlation analysis, etc. e) Assess data quality
Data Preparation (DP)	a) Clean the data b) Construct the data (i.e. create derived variables, discretize where relevant, integrate if necessary) c) Convert data to the format that the selected tool requires to satisfy the requirements of the given DM tool
Modeling or Data Mining (DM)	a) Apply to the prepared data, each DM method that was selected in the Business Understanding (BU) phase. b) Record the resulting data that corresponds to the DM performance measures elicited in the Business Understanding phase.
Evaluation (EV)	Evaluation of the generated knowledge from the business perspective a) Exclude models that do not satisfy the relevant threshold for any of the Performance Measures. b) For each model, use the Preference Function to generate a Composite Performance Score for that model. c) Rank models in descending sequence based on Composite Score.

Step 1. The variable to be considered as an item (i.e. the original Target variable) in the dataset is categorical.

a. Examine the frequency counts.
b. Generate a concept hierarchy or if one exists examine to see its appropriateness.
c. In collaboration with the business analyst determine if they categorical variables are being analyzed at the right level of granularity. Use the concept hierarchy to identify the level of analysis.
d. By using the concept hierarchy a new categorical value will be generated instead of the original item e.g. XO-50" Smart TV and WHOSUNG – 52" Smart TV can be mapped to a concept "Big Size Smart TV".

For example, an appliance company contains data on several items that are bought at the shop. The objective is to see what items are more frequently bought together. However, the number of items is quite large (250) and that number needs to be reduced. For the subset of items shown in table 2, a concept hierarchy could be used to categorize the TV based on their sizes, brands, type or a combination. The business analyst will be needed to determine which concept should be used to reduce the number of items.

Step 2. The variables to be considered as an item (i.e. the original Target variable) in the dataset are numeric.

a. Discretize the numeric variables.
b. A Composite Variable is created that contains data from each original Target variable.
c. There is thus multiple rows in the transformed dataset for each row in the original dataset (i.e. one row for each original Target variable).

TABLE 2. Example Frequency counts of items in a dataset

Items	Count
XO – 50" Smart TV	140
WHOSUNG – 52" Smart TV	45
XO – 52" TV	300
WHOSUNG – 32" TV	450
XO – 32" TV	200

For example: Given the Zoo dataset, objective is to generate ARs that involve the Predator, Toothed, and Legs attributes, where the values of the Legs attribute is broken down into three (3) categories: 0, 1-2, and greater than 2 (see table 3).

Step 3. Fig. 1 shows iterations between the data preparation phase and the modelling phase, therefore while preparing data some preliminary modelling has to be performed. Generate AR and determine whether co-occuring items are highly correlated, i.e. they occur together most of times only with each other (e.g. lift > 90% and confidence > 90%).

a. If yes then and go back to step 1 and re-examine concept hierarchy with business analyst.
For example, in an insurance company all persons who buy gold scheme get a free premium windshield coverage. It is likely that most occurrences of item "Gold scheme" will have item "premium windshield coverage" and vice versa hence it might be better to reexamine the concept hierarchy to determine how to remove the highly correlated items.

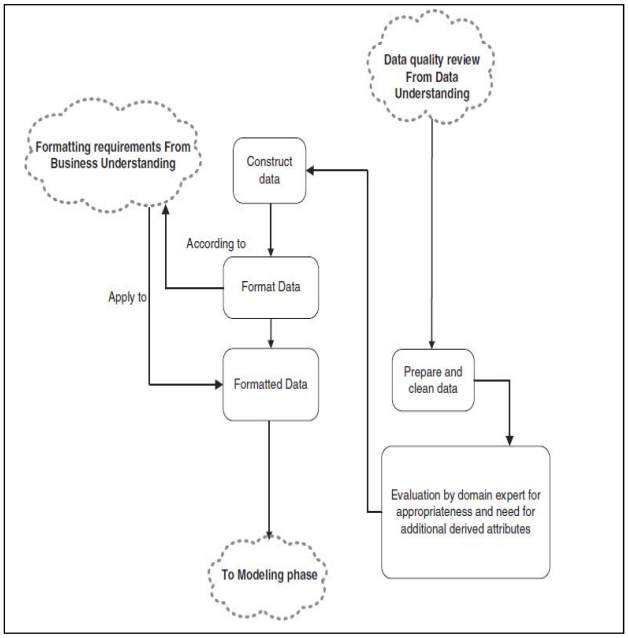

Fig 1. Data Preparation Phase (Source: Sharma and Osei-Bryson, 2010)

Step 4. Determine if demographic variables need to be added. If yes go back to either step 1 or 2 based on the type of variable. If categorical variable (e.g. Occupation) go to Step 1 and if they are numeric (e.g. age or income) go to step 2. Add these as items to the basket.
For sequential patterns a date field becomes a required field. Perform step 5 and 6 if data preparation is being done for SPM. If only doing ARI, skip to step 7.

Step 5. Determine if demographic variables have to be added. If yes, add demographic variables with an arbitrary low sequence number where each demographic variable will have a unique sequence number. For example, education can have a sequence number 3 and income can have sequence number 4. All education values and all income values will have sequence number 3 and 4 respectively in the dataset.

Step 6. For transactional items in a dataset a date field exists, hence transform the date field associated with an item to a sequence number. Convert the date field in a way that the items with a later calendar date have a higher sequence number. Take the year, month and day to create this number. For example, 19 December 2004 will become 20041219.

Step 7. Format data based on data mining tool selected in BU phase. Some tools require data to be organized in column format while some want it in the row format.

- Tool: SAS Enterprise Miner – Column Format

Cust ID	Items
A10002	Income - VALUE
A10002	Car Insurance
A10005	Education- High School
A10005	House Insurance

- Tool: RapidMiner – Row Format

Cust ID	Income-Emerging	Education - High School	Car Insurance	House Insurance
A10002	Y	N	Y	N
A10005	N	Y	N	Y

IV. CASE STUDY

A financial institution currently offers several banking products, some of which have moderate or low penetration. There is currently a business emphasis on building a sales culture across the organization with the strategic organizational priorities focusing on extending wallet share and improving customer satisfaction. Although there is anecdotal awareness of repeat customer take-up of related products, there is currently no systematic means of identifying related products or directing sales personnel to offer product bundles. In order to improve the success rate of current targeted marketing campaigns, market basket analysis can be employed as a customer analytic technique.

This technique uses association rule mining to identify the products which customers currently purchase together and therefore can help to identify those products that go well together (in terms of bundles) and should be marketed accordingly. In this case, since a customer does not buy a set of financial products all at once, the basket contains products bought over time, which makes it conducive to sequential mining which not only shows the products that were bought together but also the sequence in which they were bought. Both association rule induction (ARI) and sequential pattern mining (SPM) will help to increase the effectiveness of sales campaign and target marketing process.

A. Applying Data Preparation Steps

Based on the illustrative example in the business understanding phase ARI and SPM are the selected problem type. Also from the data understanding phase a quality report on all the data sets is received. All data sources that have fields which can represent items in an itemset were identified. All data sources were examined to determine if they had a common unique id which could be used to identify a customer. The data sets that contain customer data on types of loans, credit cards, insurance and investments were identified to be considered as items. To these data sets we apply the steps outlined in section 3 to create a data set for modelling phase.

TABLE 3. Activities in Step 2

Attribute	Discretization Function					
	Value	Discrete Value	Value	Discrete Value	Value	Discrete Value
Predator	**0**	Non_Predator	**1**	Predator		
Toothed	**0**	Non_Toothed	**1**	Toothed		
Legs	**0**	No_Legs	**1 – 2**	1_2_Legs	**> 2**	More_than_2_Legs

Initial Data	Animal_ Name	hair	feathers	eggs	milk	aquatic	Predator	Toothed	Backbone	breathes	venomous	fins	Legs
	Aardvark	1	0	0	1	0	1	1	1	1	0	0	4
	Antelope	1	0	0	1	0	0	1	1	1	0	0	4

Data Preparation Steps

Step

2a *ID variable (e.g. Animal_Name) & relevant Target Variables (e.g. Predator, Toothed, Legs)*

Animal_Name	Predator	Toothed	Legs
aardvark	1	1	4
antelope	0	1	4

2b *Create & Add relevant Discretized Target Variables (e.g. D_, D_Toothed, D_Legs)*

Animal_Name	Predator	Toothed	Legs	D_Predator	D_Toothed	D_Legs
aardvark	1	1	4	Predator	Toothed	More_than_2_Legs
antelope	0	1	4	Non_Predator	Toothed	More_than_2_Legs

2c *Do Transformation resulting in new Transaction type dataset with ID variable & new Composite Target Variable (e.g. Animal_Category) and 1 row per original Target variable.*

Aardvark	Predator
Aardvark	Toothed
Aardvark	More_than_2_Legs
Antelope	Non_Predator
Antelope	Toothed
Antelope	More_than_2_Legs

TABLE 4. Sample Baskets

Cust Id	Sequence	Target
A10002	20/03/2012	Mileage Card
A10002	14/03/2015	Loan Type X
A10002	24/03/2015	Car Insurance
A10005	20/04/2010	Low interest Card
A10005	22/01/2015	WE Mutual Funds

Step 1. In the data sources *cust id* was used as a unique id, for customers the products they acquired (i.e. card type, insurance type, type of investment and type of loan) with the institution were considered to be the items to be included in the data set. If a customer had taken up a particular credit card type or a loan that was added to that customer's basket with a corresponding date. The product types were all categorical (see Table 4). Frequency count for all items were examined. Concept hierarchies were generated by the business analyst for credit cards and insurances schemes as their counts were greater than 50. No concept hierarchy was used for loan types and investment types. As their count was less than 10.

Step 2. No numeric fields were considered to be transformed to items.

Step 3. Determine whether any relationships exist between items.

a. Association rules were generated using the baskets that were created in step 1. Several rules especially

TABLE 5. Co-related Association Rules

CONF	SUPPORT	LIFT	COUNT	RULE
100	1.11	90.16	255	M FUND ==> B FUND
100	1.11	90.16	255	E FUND ==> B FUND
100	1.11	90.16	255	M FUND & E FUND ==> B FUND

those that contained investment items had several items highly co-related (see table 5). These preliminary results were discussed with the business analyst to determine the rationale behind this occurrence. As a result a concept hierarchy was created for different investment types and highly logical co-related items were removed.

Step 4. Add demographic data to the basket data. Each numeric demographic variable such as age, income was discretized using the steps outlined in step 2. The categorical variables such a profession and education

were processed based on step 1. The basket for each customer now also contains their demographics.

Step 5. Since all items in the basket are acquired not at one time but over a period it was important to keep the date field. However, we added the demographics which would not have a date. Therefore, for each demographic field such as education, profession, income, sex, marital status and age we assigned sequence number 1,2,3,4,5 and 6 respectively (see table 6).

Step 6. The date field was transformed to a sequence number by using the process outlined in step 6. On a calendar a later date will have a higher sequence number than earlier date (see table 4 and 6).

Step 7. The tool selected was SAS Enterprise miner which requires the data to be in column format. The created dataset was is in the correct format (see table 6). This data is now prepared and ready to be sent to the modeling phase for sequential patterns to be generated.

TABLE 6. Sample of prepared data set

Cust Id	Sequence	Target
A10002	1	College / University
A10002	2	Clerical/Administrative
A10002	3	VALUE
A10002	4	Male
A10002	5	Married
A10002	6	YOUNG
A10002	20120320	XYZ Card
A10002	20150314	Loan Type X
A10002	20150324	Car Insurance
A10005	1	High School
A10005	2	Teacher/Lecturer
A10005	3	EMERGING
A10005	4	Female
A10005	5	Married
A10005	6	MIDDLE
A10005	20150122	WE Mutual Funds
A10005	20100429	ABC Card

B. Discussion

The creation of the baskets required joining several data sources. The quality of this data can be affected by either poor data quality of the data sources or a poor concept hierarchy. The data quality issues with the joining fields and duplicate rows of data, reduce the number of rows of data. This reduces the number of baskets that can be produced from the data sources. A good concept hierarchy is essential for ARI and SPM which essentially captures the context of the problem domain. It should be developed in collaboration with the business analysts, which ensures that the right context for the business problem is captured in the categorization. Hence, data preparation requires knowledge of both data mining and business analyst. In this study we provided a practical example of how data from a financial institution can be prepared for ARI and SPM. The data preparation method performed in this study is not unique to this study. The steps that we have outlined are generalizable and can be used to prepare data for ARI and SPM in any domain.

V. CONCLUSION

In this study we have outlined the steps for preparing data for ARI and SPM. Both these techniques are similar as they require data in itemsets format. By outlining the steps we are reducing the variability in this process thus ensuring that the tasks of data preparation phase, such as *construct* and *format* can easily be duplicated in other studies and are not dependent on the context and expertise of the data mining analyst. There are still some steps that require business analyst's intervention (for example the concept hierarchy) hence there will always be some art in this phase however by publishing more research of this type and capturing the common patterns of data preparation techniques we can explicate the science in this art of preparing

REFERENCES

[1] S. Okazaki, "What do we know about mobile internet adopters? A cluster analysis," Information and Management, vol. 43(2), 2006, pp. 127-141.

[2] L. A. Kurgan and P. Musilek, "The survey of Knowledge Discovery and Data Mining process models," The Knowledge Engineering Review, vol. 21(1), 2006, pp. 1-24.

[3] R. L. Grossman, M.F. Hornick, and G. Meyer, "Data Mining Standards Initiatives," Communications of the ACM, vol. 45(8), 2002, pp. 59-61.

[4] C. Clifton and B. Thuraisingham, "Emerging standards for data mining," Computer Standards & Interfaces, vol. 23(3), 2001, pp. 187-193.

[5] S. Sharma, K.-M. Osei-Bryson, and G.M. Kasper, "Evaluation of an integrated Knowledge Discovery and Data Mining process model," Expert Systems with Applications, vol. 39(13), 2012, pp. 11335-11348.

[6] "|n Big Data Preparing The Data Is Most Of The Work", http://www.datasciencecentral.com/profiles/blogs/in-big-data-preparing-the-data-is-most-of-the-work, 2015. Retreived on 24th November 2015.

[7] J. M. Hall and M.E. Johnson, "When should a process be art, not science?", Harvard Business Review, March 2009.

[8] S. Zhang, C. Zhang, and Q. Yang, "Data Preparation for Data Mining", Applied Artificial Intelligence, vol. 17, 2003, pp. 375-381.

[9] S. Sharma and K.-M. Osei-Bryson, "Toward an integrated knowledge discovery and data mining process model," The Knowledge Engineering Review, vol. 25(1), 2010, pp. 49-67.

[10] M. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from a Database Perspective," IEEE Transaction on Knowledge and Data Engineering, vol. 8(6), 1996, pp. 866-883.