

A Review on Laws of Data Mining

Surbhi Arora¹, Joni Birla²

^{1,2}Department Of Computer Science & Engineering,
Ganga Institute Of Technology And Management,
Kablana, Jhajjar, Haryana, India

Abstract--Once, the term “data mining” was only used to describe a specific new culture of business people using innovative methods of uncovering useful patterns in data. The tools of the data miner were quite distinct from those of the traditionally trained analyst. When you heard “data mining,” you could be sure of knowing what it meant. Today, the term “data mining” is thrown around very casually, and it may be used to describe anything from a business person using modern pattern recognition methods to a database analyst making SQL queries. As interest in data mining grew, lots of people started claiming, “Us, too! We’ve got data mining, too!” Much of this change has been driven by vendors who didn’t want to be left out of the data mining party, but didn’t want to invest in new tools and processes, either.

I. INTRODUCTION

Data mining is the creation of new knowledge in natural or artificial form, by using business knowledge to discover and interpret patterns in data.

In its current form, data mining as a field of practise came into existence in the 1990s, aided by the emergence of data mining algorithms packaged within workbenches so as to be suitable for business analysts. Perhaps because of its origins in practice rather than in theory, relatively little attention has been paid to understanding the nature of the data mining process. The development of the CRISP-DM methodology in the late 1990s was a substantial step towards a standardised description of the process that had already been found successful and was (and is) followed by most practising data miners.

Although CRISP-DM describes how data mining is performed, it does not explain what data mining is or why the process has the properties that it does. In this paper I propose nine maxims or “laws” of data mining (most of which are well-known to practitioners), together with explanations where known. This provides the start of a theory to explain (and not merely describe) the data mining process.

It is not my purpose to criticise CRISP-DM; many of the concepts introduced by CRISP-DM are crucial to the understanding of data mining outlined here, and I also depend on CRISP-DM’s common terminology. This is merely the next step in the process that started with CRISP-DM.

II. LAWS OF DATA MINING

A. 1st Law of Data Mining – “Business Goals Law”:

Business objectives are the origin of every data mining solution

This defines the field of data mining: data mining is concerned with solving business problems and achieving business goals. Data mining is not primarily a technology; it is a process, which has one or more business objectives at its heart. Without a business objective (whether or not this is articulated), there is no data mining.

Hence the maxim: “Data Mining is a Business Process”.

B. 2nd Law of Data Mining – “Business Knowledge Law”:

Business knowledge is central to every step of the data mining process

For convenience I use the CRISP-DM phases to illustrate:

- 1) Business understanding must be based on business knowledge, and so must the mapping of business objectives to data mining goals. (This mapping is also based on data knowledge data mining knowledge).
- 2) Data understanding uses business knowledge to understand which data is related to the business problem, and how it is related.
- 3) Data preparation means using business knowledge to shape the data so that the required business questions can be asked and answered. (For further detail see the 3rd Law – the Data Preparation law).
- 4) Modelling means using data mining algorithms to create predictive models and interpreting both the models and their behaviour in business terms – that is, understanding their business relevance.
- 5) Evaluation means understanding the business impact of using the models.
- 6) Deployment means putting the data mining results to work in a business process.

3rd Law of Data Mining – “Data Preparation Law”:

Data preparation is more than half of every data mining process

It is a well-known maxim of data mining that most of the effort in a data mining project is spent in data acquisition and preparation. Informal estimates vary from 50 to 80 percent. Naive explanations might be summarised as “data is difficult”, and moves to automate various parts of data acquisition, data cleaning, data transformation and data preparation are often viewed as attempts to mitigate this “problem”. While automation can be beneficial, there is a risk that proponents of this technology will believe that it can remove the large proportion of effort which goes into data preparation. This would be to misunderstand the reasons why data preparation is required in data mining

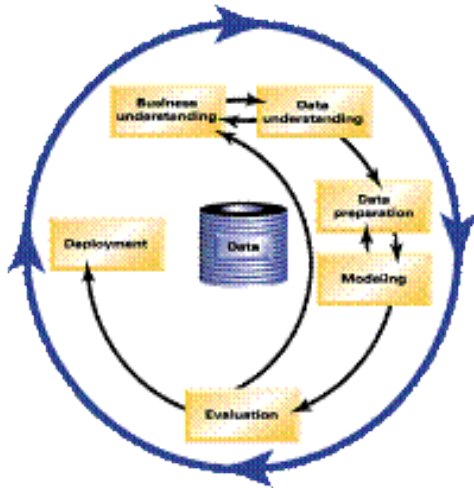


Fig 1. Process Cycle

C. 4th Law of Data Mining – “NFL-DM”:

The right model for a given application can only be discovered by experiment or “There is No Free Lunch for the Data Miner”

There are 5 factors which contribute to the necessity for experiment in finding data mining solutions:

- 1) If the problem space were well-understood, the data mining process would not be needed – data mining is the process of searching for as yet unknown connections.
- 2) For a given application, there is not only one problem space; different models may be used to solve different parts of the problem, and the way in which the problem is decomposed is itself often the result of data mining and not known before the process begins.
- 3) The data miner manipulates, or “shapes”, the problem space by data preparation, so that the grounds for evaluating a model are constantly shifting.
- 4) There is no technical measure of value for a predictive model (see 8th law).
- 5) The business objective itself undergoes revision and development during the data mining process, so that the appropriate data mining goals may change completely.

D. 5th Law of Data Mining: There are always patterns in the data.

As a data miner, you explore data in search of useful patterns. Understanding patterns in the data enables you to influence what happens in the future.

E. 6th Law of Data Mining, or “Insight Law”:

Data mining amplifies perception in the business domain.

Data mining methods enable you to understand your business better than you could have done without them.

F. 7th Law of Data Mining or “Prediction Law”:

Prediction increases information locally by generalization.

Data mining helps us use what we know to make better predictions (or estimates) of things we don't know.

G. 8th Law of Data Mining, or “Value Law”:

The value of data mining results is not determined by the accuracy or stability of predictive models.

Your model must produce good predictions, consistently. That's it

H. 9th Law of Data Mining, or “Law of Change”:

All patterns are subject to change.

Any model that gives you great predictions today may be useless tomorrow.

III. CONCLUSION

In this research paper we have studied data mining and its 9 laws. we have cover following topics: Goals Law, Knowledge Law, Data Preparation Law ,NFL _DM, Law Of Data Mining, Law of Chane , Insight Law, Prediction Law Etc.

REFERENCES

- [1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.
- [2] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.
- [3] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
- [4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.
- [5] Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 9781558604896. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long
- [6] See e.g. OKAIRP 2005 Fall Conference, Arizona State University, About.com: Datamining
- [7] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.
- [8] Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research **11**: 2533–2541. the original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons.
- [9] Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, and Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.
- [10] Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, and Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012