The Influence of Chance on Data Mining Results

By: Ammar Y. Elnour *Supervisor : Prof. Joseph Davis*

ammar@it.sydney.edu.au

School of Information Technology

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

THE UNIVERSITY OF SYDNEY

> A single death is a tragedy. A million deaths is a statistic.
>
> (Joseph Stalin)

1. Objectives

- To develop a generalized conceptual model to asses the influence of chance on the results of data mining. The concept will be based on the theory of statistical inference and uncertainty when applied on scale-free distributions.
- To develop testing procedures to test the method using nonparametric and bootstrapping methods and techniques such as hypotheses testing, statistical estimation and Inference.
- To validate the methods using experiments on large datasets.

2. Introduction

Investigating uncertainty and randomness have been a major concern that puzzled scientist in various branches of sciences and human activities. The existence of uncertainty touches our every day life from all directions. Such degree of doubt may affect our decisions specially when the amount of uncertainty can not be quantified easily. Data mining is among these disciplines that suffer from the lack of standard tools to measure the influence of uncertainty (chance) among the generated results.

The rapid developments in information technologies have rendered it easy to gather and electronically capture a huge volume of data in different fields. These volume of data has the potential to produce useful information and new knowledge if appropriate mining techniques are used, however a good proportion of the results are redundant, obvious or statistically invalid.

This research addresses the validity of the results which is still a rich area for research. Statistical techniques alone are not capable enough to address mining issues [1], but they can play significant role in assessing the results.

3. Chance-setup in Data Mining

3.1 What is chance setup

The chance is a dispositional property of one or more trials, experiments, or observations. This property relates to the long run frequency of an outcome on the trials [2].

3.2 The influence of chance on data mining

Data mining datasets are observational data. They are not controlled by any method of statistical experimental techniques which means that data mining researchers have no control over the source of the data, which leads to the lack of the possibility for random assignment to different group of data [3]. Datasets used in data mining are relatively large. Large data sets are more likely to be less clean which is an initial requirement for most statistical methods of analysis. The identically and independently distribution of data items is also one of the major assumptions that most statistical methods based on. However, large data sets are unlikely to fulfil this assumption. Data sets used in data mining tasks have their own characteristics where their *chance set-up* exists and affects later manipulations or processes. Datasets used in data mining tasks may be the result of mixture systems [4], where some hidden factors, unrecorded causes or latent variables may be the cause of part of the discovered associations and patterns. Considering the many factors that may trigger *chance set-up* in observational datasets used by data mining tasks and the fact that data mining results offer candidate theories which need further validations using statistical approaches [6]. It is reasonable to conjecture that chance can influence the results of data mining methods.

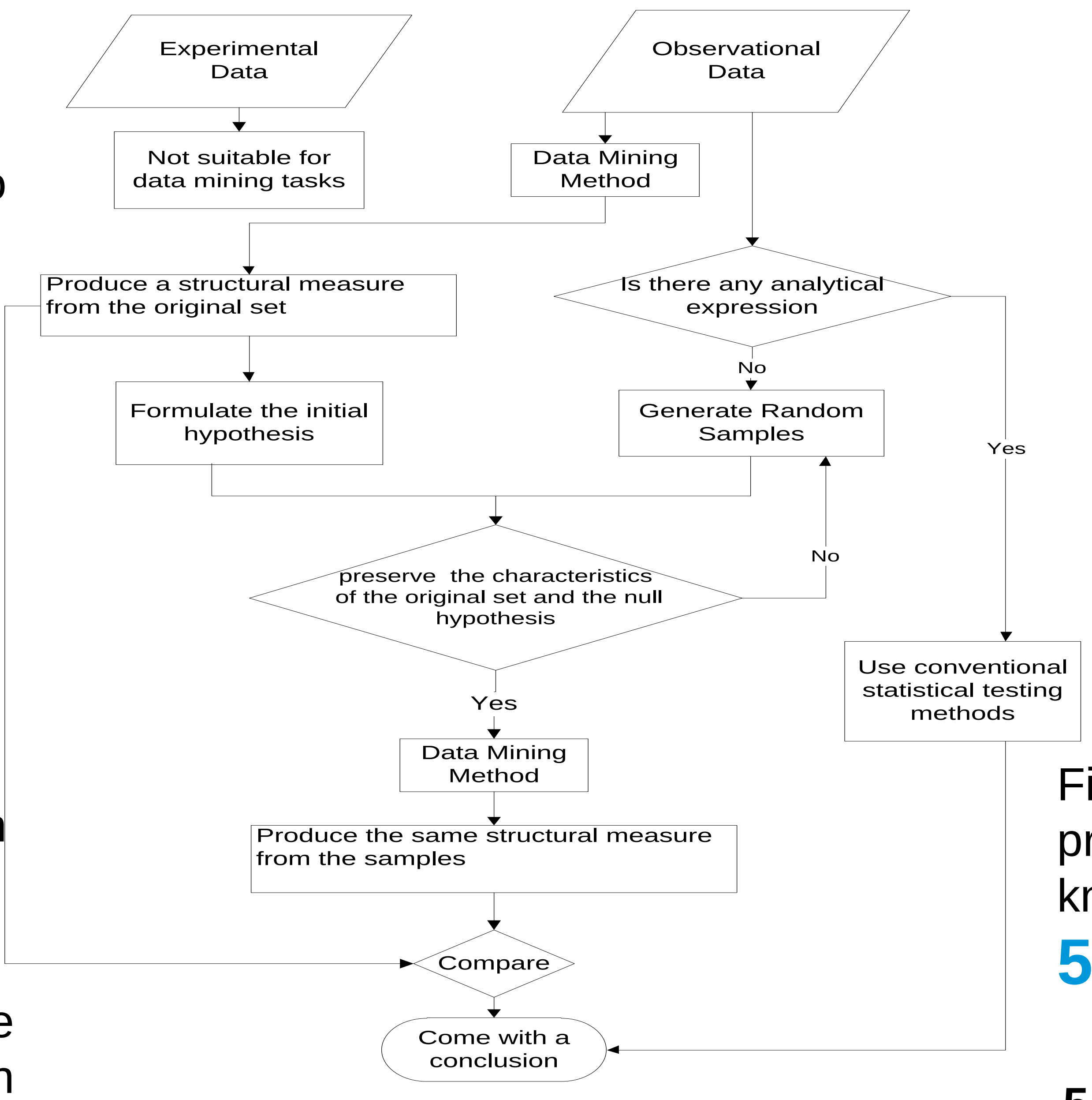

Experimental Data
Observational Data
Not suitable for data mining tasks
Data Mining Method
Produce a structural measure from the original set
Is there any analytical expression
No
Yes
Formulate the initial hypothesis
Generate Random Samples
preserve the characteristics of the original set and the null hypothesis
No
Yes
Use conventional statistical testing methods
Data Mining Method
Produce the same structural measure from the samples
Compare
Come with a conclusion

Fig1: Testing the validity of data mining results

4. Research Methods and Tools

Statistical inference methods are the main tools in our research. Bootstrap and resampling techniques will be used to simulate and approximate the distributions of the original datasets. Nonparametric hypotheses testing methods will also be used to test the validity of discovered knowledge and hence to quantify the influence of chance on the results.

Fig 3 :The Deviation of the Sampled Rule–confidences from the originals
(Mushroom Dataset)

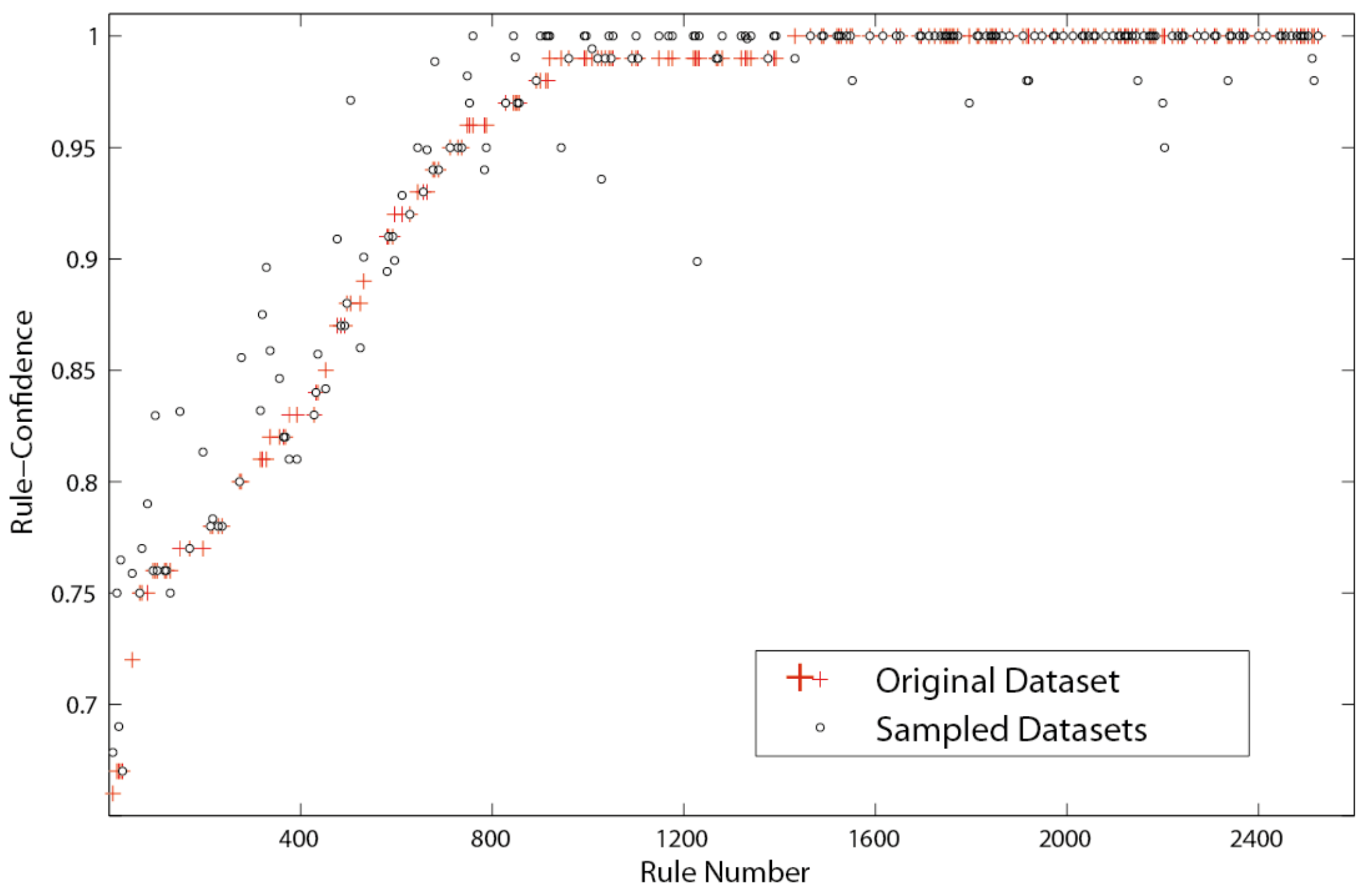


Figure 1 shows the inferential model that we are proposing to test the validity of discovered knowledge from a data mining task.

5. Chance Free Data Mining Model (CFDMM)

5.1 Problem Definition

Given a dataset *D(X)* where: $X = \{x_1, ..., x_n\}$ are the attributes that describe the dataset.

Denote a data mining task *T* that acts on the dataset X to produce a set of results (patterns) R_i by :

$R_i[m_i(\theta)] \leftarrow T[X, p(\theta)]$, where

$p(\theta)$ ≡ are the initial parameters that seed task *T* ,

$m_i(\theta)$ ≡ structure measure of result R_i

← ≡ denotes the term ”produces”

which is read as: Task *T* that is seeded by the set of parameters $p(\theta)$; works on the dataset *X* to produce result R_i with accuracy estimate $m_i(\theta)$.

5.2 Def: Chance-expressing datasets

Given an original dataset *D(X)* and a data mining task *T* There is “*part of the world* [5]” where a dataset *D(X')* of the type *D(X)* exists and it is by chance-expressing the original *D(X)* if :

- it is a result of random process on the original dataset
- the number of rows in *D(X')* is equal to the number of rows in *D(X)*.
- the number of attributes in *D(X')* is equal to the number of attributes in *D(X)*.
- preserving the initial characteristics of the original dataset.
- satisfying the requirements of the null hypothesis of task *T*, i.e. $\min \langle T[X, p(\theta)] - T[X , p(\theta)] \rangle < \epsilon$

5.2 Empirical Distribution for a data mining dataset

Given an original dataset *D(X)* and a data mining task *T*. The empirical distribution of *D(X)* denoted by $D(X_r')$ where each $D(X_r')$ is of the type *D(X)* ;$1 \le r \ge n$, is a set of *n* *chance-expressing* replicates of D(X) that satisfies $\epsilon_k < \epsilon_{k-1}$ $(k=2; 3; \ldots; r)$

6. Experimental Results

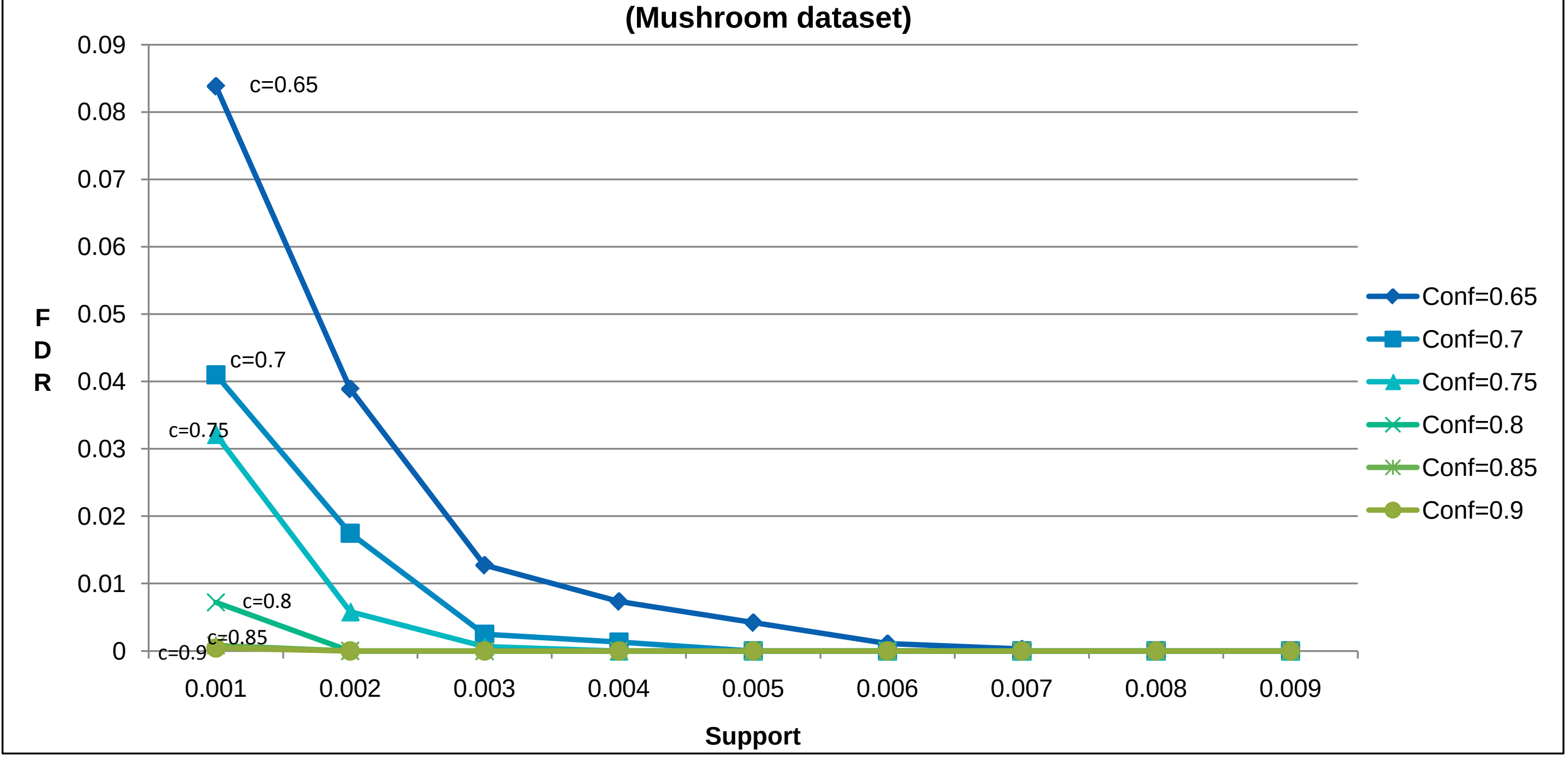
6.1 Testing the validity of specific Rule.

Figure 2 and 3 shows some of the results of our experiments on testing the significance of a single association rule (AR).

7. Further work:

We are working on a method to generate random samples that reflect the chance setup and suitable for assessing a general data mining task.

Fig 2 : FDR under Low Support ARM
(Mushroom dataset)



References:

[1] M. David J. Hand and P. Smyth, Principles of Data Mining. MIT Press, 2001.

[2]Hacking, Logic of Statistical Inference. Cambridge University Press, 1965.

[3] D. Hand, “Mining the past to determine the future: Problems and possibilities,” International Journal of Forecasting, October 2008.

[4] C. Glymour, D. Madigan, D. Pregibon, P. Smyth, and U. Fayyad, “Statistical themes and lessons for data mining,” Data Mining and Knowledge Discovery, vol. 1, no. 1, pp.11–28, 1997.

[5] Zhang and B. Padmanabhan, “Using randomization to determine false discovery rate for rule discovery,” in Proceedings of the Fourteenth Workshop On Information Technologies and Systems 11,12 December 2004. WITS, 2004, pp. 140–145.

[6] G. Pandey, S. Chawla, S. Poon, B. Arunasalam, and J. G. Davis, “Association rules network: Definition and applications,” Stat. Anal. Data Min., vol. 1, no. 4, pp. 260–279, 2009.