A Study of the Factors Considered when Choosing an Appropriate Data Mining Algorithm

Teressa T. Chikohora

Abstract: ***A lot of data is generated and collected in today's organizations .Data mining has helped a lot of businesses to extract knowledge from data and use it to make decisions and gain competitive advantage. Businesses now analyse the data to make business decisions. Various algorithms may be used to analyse the data, however some of them do not yield useful knowledge. Choosing the appropriate algorithm remains a problem given the diversity in available algorithms. There are many algorithms, making it difficult for analysts and researchers who may not know which algorithm will be suitable for their needs. As a way of optimizing the chances of extracting useful knowledge, this study focuses on how the data analysts and researchers may choose appropriate algorithms that will yield desired knowledge. A number of factors to be considered when selecting an algorithm are discussed to help analysts in choosing appropriate algorithms.***

Key words: algorithm, factors, tool, data mining

I. INTRODUCTION

Data mining is a computer-assisted process of discovering patterns in data through analysing huge amounts of data and extracting meaning from these patterns, [1]. It has become common practice in organisations because of the enormous amounts of data collected and the strong competitive pressure. Businesses are now being forced to use the data they have to their advantage over their competitors. As such, data mining allows businesses to extract patterns in the data, classify their customers and study their behavioural trends. These patterns help businesses to make proactive, knowledge driven decisions which helps them to act quickly to their customer needs, thus giving them competitive advantage, [10].[5] mentioned that data mining helps to uncover hidden information that businesses can use to recognise important facts, relationships, trends, patterns, exceptions and anomalies in order to make better decisions and judgements. The uncovered information may be used by the organisations for market segmentation, fraud detection, market basket analysis and marketing. In order to extract this useful information, there are many algorithms and tools that may be used. They may be categorised into classification, regression, segmentation, association and sequence analysis algorithms, [2]. The data mining process follows a set of steps that must be executed irregardless of the algorithm that will be implemented. Several algorithms may be used to perform the same task and still produce different results. In this study, data mining algorithms are described, the mining techniques are analysed in order to identify the key input for their implementation and a set of factors to be considered when choosing an algorithm are identified and described.

II. BACKGROUND INFORMATION

Data mining involves the following steps shown in Figure1, which must be executed in sequence until the results are documented and reported, [9].

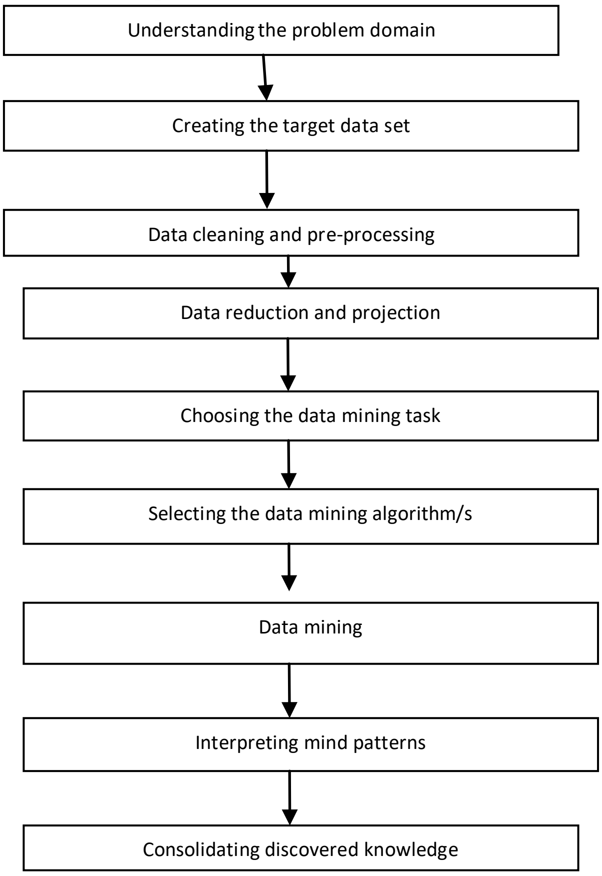
Understanding the problem domain
Creating the target data set
Data cleaning and pre-processing
Data reduction and projection
Choosing the data mining task
Selecting the data mining algorithm/s
Data mining
Interpreting mind patterns
Consolidating discovered knowledge

Fig 1: Steps Involved in the Data Mining Process

Most organisations are exposed to a number of data mining algorithms such that selecting an appropriate algorithm is a difficult task. When an inappropriate algorithm is implemented, the discovered knowledge is often meaningless to the organisation as it does not uncover the correct information leading to wrong business decisions, [9]. There are no specific guidelines as to how a researcher or analyst may choose an algorithm .As a result, inappropriate algorithms have been used and no useful information uncovered.

III. LITERATURE REVIEW

It is very important that an appropriate algorithm be used as it has an impact on the results and knowledge derived. The

Manuscript Received on July 2014.

Teressa T. Chikohora, Department of Software Engineering, Faculty Of Computing, Botho University, P. Bag F451Francistown, Botswana.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

algorithm defines the concrete method/s that will be used to search for patterns in data. [3] defined an algorithm as a mathematical procedure for solving a specific kind of a problem. In other words, the algorithm used should be able to yield the desired patterns. However, [9] mentioned that choosing an appropriate algorithm is a challenge for most researchers and there are not many tools that may be used to help them in selecting the algorithm. Researchers and data miners choose the data mining algorithm by considering the main goal of the problem to be solved and the structure of the data set, [9]. Algorithms have been defined to work with specific data sets, for instance, data that is grouped into clusters will require algorithms defined in the clustering techniques like decision trees, [5]. [10] mentioned the fact that algorithms are data set specific has forced some companies to acquire several tools for specific purposes. An organisation may use a feature extraction model to create a set of predictors and then a classification model to make a prediction, [3]. Selecting an algorithm has become so complex that other researchers adopt more than one technique in an effort to improve the data mining results, [6]. In making their choice , researchers may consider using an algorithm that will produce the correct result , is fast and the code well documented and clean, [8].The algorithm must produce the desired results based on the data mining task that needs to be accomplished and must be fast enough to search through the enormous amounts of data .[8] mentioned that it is important to understand the algorithm and if possible perform test with other input files before implementing it on the actual data set. This helps to check whether the selected algorithm will be appropriate or not.

Evaluation measures and the configuration parameters are also critical in choosing a data mining algorithm. [10] mentioned that data miners must select an algorithm that may be fully integrated with the database or data warehouse so as to avoid cost of extracting, importing and analysing the data. An appropriate algorithm based on this criterion will be one that may be integrated fully with the organisation’s current database. However, there are many computing tools available that may not work with every database, hence the need to consider the configuration parameters of the database in selecting an algorithm. The task to be performed also influences the choice of algorithm, [7].For instance if the task is a classification task, then the possible algorithms would be logistic regression, Naïve Bayes, Support vector machine or decision trees. When a task is an association task, the Apriori algorithm may be implemented [4]. The data miner’s familiarity with an algorithm may influence him or her into selecting this algorithm for all the data mining tasks assigned to them. [3] stated that a researcher may be able to use an algorithm without understanding the inner details of how it works but just an understanding of the algorithm’s general characteristics. In another study, [7] mentioned that data miners need to understand the algorithm they choose for a successful data mining task. Another factor that may be considered in choosing a data mining algorithm is the expected result set as well as what you want to use the data for, [12]. An appropriate algorithm is one that will provide the data miners with the correct results and useful knowledge that will inform the business decisions. Research shows that there are no predefined guidelines that may be followed in selecting a data mining algorithm. A careful consideration of the different factors identified by different authors must help data miners in selecting an algorithm. Ideally, a combination of the factors will increase the chances of selecting the most appropriate data mining algorithm.

IV. METHODOLOGY

A literature survey was conducted to inform this study. Literature on data mining techniques and algorithms was reviewed to give an insight into how data miners may select an algorithm to use in their tasks. The different authors mentioned what was required for the algorithms to be implemented successfully. An analysis of these requirements helped to draw a list of factors that could be considered when selecting an appropriate data mining algorithm. To understand how the data mining process is carried out, documentation for two tools , DM assistant, [9] and GESCONDA tool , [11] was reviewed .The documentation explained how a user would use the tools for data mining. There were no specific details of how the algorithm could be chosen but the result sets from the various algorithms were provided. It is from these result sets that the factor proved to be important in selecting an algorithm for use.

V. RESULTS DISCUSSION

This study revealed that there are factors that researchers may consider when selecting a data mining algorithm.

5.1 Main goal of the problem to be solved

This factor considers the reason why we are mining the data as well as the nature of the problem we are trying to address. A loan company may use a statistical decision procedure to determine whether to accept or reject cases for loan applications. The company may need to use classification rules to predict the number of customers who are likely to default their loan repayments based on information such as age, years with current employer, years with the bank and other credit cards possessed, [13] .So depending on the problem we are trying to solve data miners must select an appropriate algorithm. [2] stated that multiple algorithms may be used in a single solution to perform multiple tasks, for example using regression to obtain financial forecasts and then use neural network algorithm to perform an analysis of factors that influence sales. Table1 below shows how the nature of problems to be solved may be matched to possible algorithms, [13].

Table 1: Example Problems and the Possible Algorithms that may be Adopted to Solve them

Problem to be solved	Data mining Technique	Possible algorithm / s
identify anomalies in data, to find outliers	Classification	Decision trees One Class Support Vector Machine

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Find items that tend to co-occur in the data and the rules that govern their occurrence	Association rules	Apriori
Find groupings in data	Clustering	K-Means
Create new features using linear combinations of the original attribute	Feature extraction	Non- Negative Matrix factorization

To find anomalies in data, one may choose to use the decision trees algorithm.

5.2 Structure of the available data set

The data you provide is first analysed to identify specific types of patterns or trends before defining the mining model. The results of the analysis are used to define the parameters for the mining model hence the need to consider the data set, [2]. The relationships between the objects / data, relationships between variables and the way that the data is stored influences the choice of an algorithm , [9].Table 2 (below) shows examples of the data that will be required if a data miner is to implement a specified algorithm, [2] .

Table 2: Examples of the Required Structure of Data Sets for Some Algorithms

Algorithm	**Structure of data set**
Association algorithm	Single key column Single predictive column Input columns contained in two tables
Linear Regression	A key column Input column At least one predictable column
Clustering algorithms	Single key column At least one input column with values that are to be used to build the clusters Optional predictable column
Naïve Bayes	Single key column At least one predictable attribute At least one input attribute None of the attributes can be continuous numeric data as it will be ignored.
Neural networks	One input column One output column Data can be continuous , cyclical, discrete , key table or ordered
Time Series	Key time column that contains unique values Input columns At least one predictable column

The table above shows the recommended structure of the data if a researcher is to adopt a specific technique. The single key column is the column that uniquely identifies each record, in other words it is the primary key in a table. The predictive column is the key column in the nested table, the foreign key .Therefore based on these examples, to implement a time series algorithm, there must be one column with time data, input columns and at least one predictable column.

5.3 Expected results.

Every data mining task must yield a desired solution hence the importance of selecting an appropriate algorithm. The major aim of data mining is to identify patterns and trends in data so as to use the knowledge in decision making, [5]. Depending on the type of results expected , data miners may select an algorithm that is closest to produce the desired results. It is the result of the data mining process that determines the success or failure of a data mining task. In other words, if data mining does not produce the required results then, it would have failed. The most common techniques produce specific type of results, so a researcher may consider the type of results in assessing whether the algorithm is appropriate or not. Table 3 below shows the result types for some of the common techniques, [5].

Table 3: Some Data Mining Techniques with the Type of Results they Produce

Algorithm	**Type of results**
Classification	Maps data into predefined groups or classes by analysing multiple attributes
Clustering	Groups similar data together into clusters by examining one or more attributes
Association	Make a correlation between 2 or more items of the same type
Regression	Maps data item to a real valued prediction variable where a variable is valued based on the values of other variables.

5.4 What the Information Would be used for

Results of data mining tasks inform the business's decisions such that if the uncovered information is wrong, incorrect decisions may be made. The information may be used for market segmentation, market basket analysis and trend analysis to mention a few examples. In order for the information to be useful, it must be presented in the correct structure, for example to decide on the market segments, the information must be grouped together into clusters thus requiring the adoption of clustering algorithms for the data mining task.

5.5 Familiarity with an Algorithm

Having experience in implementing an algorithm may make the selection much easier. Data miners may adopt those algorithms that they are familiar with although there is a risk that the chosen algorithms may not be suitable for the task to be performed. This factor however becomes useful when the same type of tasks is to be performed. A case base where all the experiences with different algorithms are documented may be useful when adopting this factor in selecting an appropriate algorithm. Researchers may get an idea of which algorithms work best in specific domains as well as the challenges they are likely to face, well before the data mining resumes.

5.6 Configuration Parameters

When considering this factor, select an algorithm that may be integrated to the data source at minimal costs. An algorithm that may be fully integrated to the organisation's database would be more ideal, where extracting, importing and exporting the data will be easy to automate without incurring extra costs. However, it does not mean that an algorithm must be selected if it can only be integrated without satisfying other criteria. To make an informed choice, researchers may consider more than one of the factors in selecting a data mining algorithm to use. Considering only one aspect may increase chances of yielding undesired results, rendering the data mining process meaningless.

VI. CONCLUSION AND FUTURE WORK

Taking the proposed factors into consideration may help data miners to select an appropriate algorithm and increase the chances of extracting correct, useful knowledge that organisations may use to gain competitive advantage. This stage in the data mining process is very important as it determines the kind of results that will be generated which will influence the business decisions. Therefore if the results are incorrect, there are chances that the business will make wrong decisions posing a negative impact on its operations. However, there may be other factors that were not discussed in this study that are worth considering. Further work may be conducted to identify more factors that data miners may consider when choosing a data mining algorithm. The study is based on literature review and did not experiment to verify whether the suggested factors really assist in selecting an appropriate algorithm. I therefore recommend that a tool that will accept the parameters as defined by the factors be developed. After considering the values that the data miner would have specified, the tool will match these values to already set criteria, specified for each algorithm and then suggest possible algorithms that may be adopted. A case base may be maintained by organisations, miners and researchers would review this base in order to get guidance in choosing an algorithm for their task.

ACKNOWLEDGMENT

This paper was originally submitted as an assignment for an MSC degree in Information Systems at the National University of Science and Technology (NUST), Zimbabwe.

REFERENCES

1. Alexander, D. (n.d.): Data Mining [online] http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ (accessed 13/12/2013 234 p.m.)
2. Anon (2012): Data Mining Algorithms (Analysis Services – Data Mining) [online] http://technet.microsoft.com/en-us/library/ms175595.aspx (accessed on 08/01/2014) 2012.
3. Anon, (2008): Oracle Data Mining Concepts 11g Release 1 (11.1).
4. Anon, (n. d.): An overview of Data Mining Techniques.
5. Brown, M. (2012): Data Mining Techniques [online] http://www.ibm.com/developerworks/library/ba-data-mining-techniques/ (accessed 23/12/2013).
6. Cheeseman, P. and R. W. Oldford (1994): Selecting models from data. LNStats 89, Springer.
7. Chung, M.H. and P. Gray (1999): Special Section Data mining. Journal of Management Information Systems Volume 16 No.1.
8. Fournier-Viger, P. (2013): What are the steps to implement a data mining algorithm? [Online] http://data-mining.philippe-fournier-viger.com/what are the steps to implement a data mining algorithm (accessed 30/01/2014)
9. Gibert, K., Sànchez-Marrè, M. and V. Codina (2010): Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation.
10. Silltow, J. (2006): Data Mining 101: Tools and Techniques online] http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/ (accessed on 13/12/2013 at 409pm).
11. Miquel Sànchez-Marrè, Karina Gibert and Ignasi Rodríguez-Roda (n.d.): GESCONDA: A Tool for Knowledge Discovery and Data Mining in Environmental Databases.
12. Parthasarathy, S. (n.d.): CIS 674 Introduction to Data Mining.
13. Witten, I. H. and E. Frank (2005): Data Mining Practical Machine Learning.

AUTHORS PROFILE

Teressa T. Chikohora is a Computing Lecturer at Botho University in Botswana with more than 7 years experience in teaching and training. She is a holder of a Bachelor of Science Honors Degree in Information Systems from the Midlands State University in Zimbabwe. She is currently studying for a Master of Science Degree in Information Systems at the National University of Science and Technology (NUST) in Zimbabwe. Her research interests include assessment in education, security in cloud computing, data mining, networking and ecommerce.