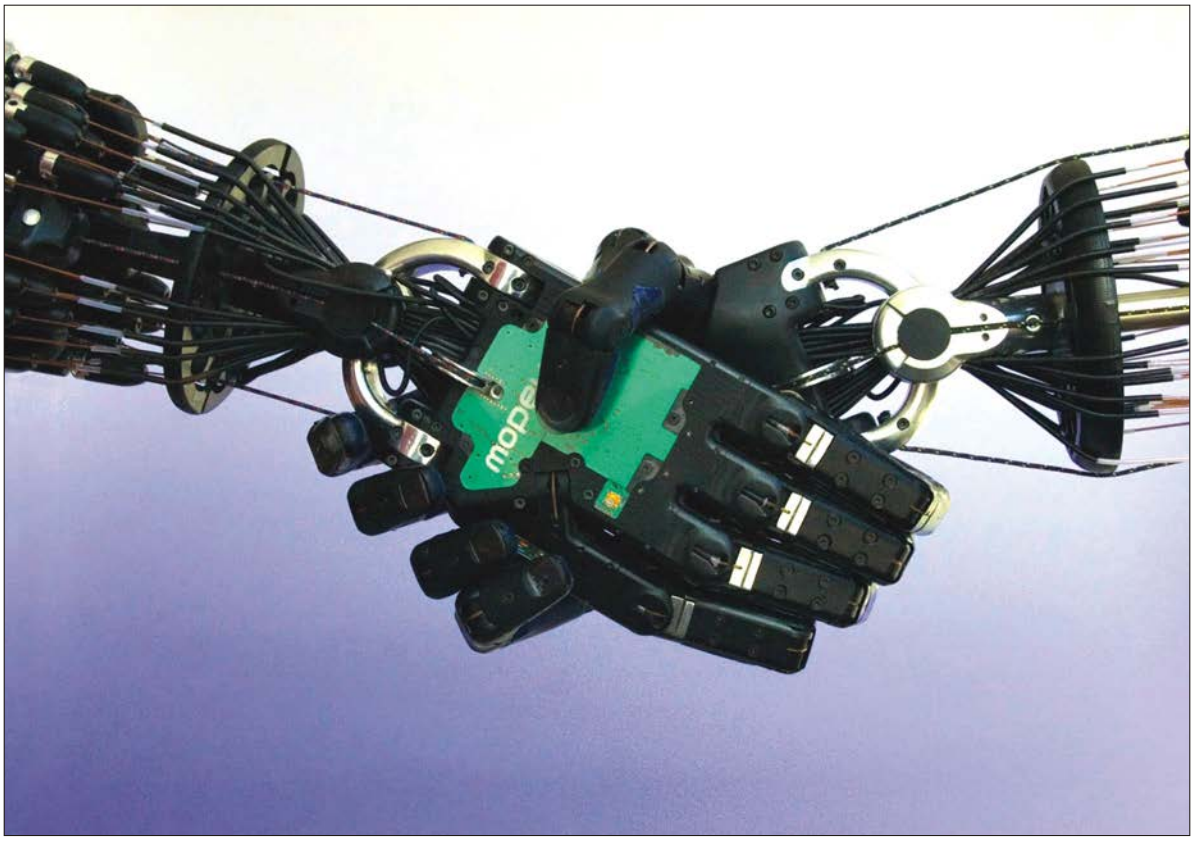

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 4 (1): 21-28, 2011
ISSN 1994-5450 / DOI: 10.3923/jai.2011.21.28

Knowledge Discovery Process: Guide Lines for New Researchers

L. Al-Shalabi
Faculty of Computer Studies, Arab Open University, Kuwait Branch, Kuwait

ABSTRACT

Guide lines for new researchers who are interested in the field of knowledge discovery and especially the data mining. Three engines were described; the preprocessing engine which describes the preparation of data in a dataset that will be called later as training dataset, the processing engine which is the data mining engine that describes the process of training the dataset (the training dataset) and the after-processing engine which describes and represents the new knowledge as a knowledge discovery or a data mining model. The challenge is how to prepare the data for data mining. The data should be of high quality so it will help in getting a high accuracy for a data mining system.

Key words: Knowledge discovery, data mining, missing values, normalization

INTRODUCTION

The increasing computerisation in all parts of life creates an ice berg of data. Transactions such using a stock market transactions or banking transactions are being routinely stored in a log file or in a database with all its details. Many companies are now looking at ways to use this big amount of data.

Data mining is to discover unseen knowledge from big amount of data. It is represented in a high level language. It is used to discover knowledge from given data which help decision makers in taking the important decisions and then to increase the profit of the company. One way of presenting the discovered knowledge is the use of rules. Discovery systems have been applied to real databases in medicine (Pieter and Dolf, 1996; Merz and Murphy, 1996), astronomy (Han and Kamber, 2000), the stock market (White, 1987) and many other areas.

This study give a brief idea about preprocessing data for data mining (preprocessing engine), processing data by a specific data mining algorithm (processing engine) and after-processing engine which is represented by the data mining model.

THE DATA

In any computerized system, the priority is always given to the quality of data which is used or manipulated. When we start to build any system, we first study the data that will be populated in the database and then design the database to suite such kind of data. From this point, it is assured that the quality of data is the starting point to get a reasonable knowledge which should come up with benefits for any company.

Data always changed based on different reasons such as changing of the measured variables of an instrument, changing of prices, changing of personal data and many other reasons. Companies usually keep all data (old and new which will be old after awhile) in an electronic archive similar to old physical archives but with more arrangement. Without any doubt, this big amount of data has some kind of useful knowledge to the company. The problem is that we may

face many problems inside this data which could make it difficult to get the desired knowledge from it. We should not use noisy data to come up with quality knowledge. Instead, we may use either one of the followings:

- Invest only high quality data which does not have any kind of noise
- Solve the noise inside the data including the missing data and the redundancy and then use the cleaned data to discover knowledge

PREPROCESSING, PROCESSING and AFTER PROCESSING ENGINES

These engines represent the knowledge discovery process. Before data mining can be used, a target dataset must be assembled. The training dataset must be large enough to contain the desire knowledge since this knowledge only exists in this dataset. Log files or data warehouses are common sources for data because they have a huge amount of data. Preprocessing is important to analyze the datasets before data mining process starts in order to make it ready to be used and in order to make sure that it is of high quality so we can get what we wish to get.

The dataset which will be used for mining is divided into two different datasets, the training dataset which we also called target dataset and the testing dataset. The training dataset is used to be trained by the data mining algorithm, while the testing dataset is used to verify the accuracy of the knowledge discovered.

The training dataset is then cleaned and prepared for data mining process. Cleaning removes the observations with noise and missing data.

The cleaned data should be reduced. This could be made by reducing the size of the data by normalizing the data itself and/or by reducing the data vertically or horizontally. For example, the address of labors in Toyota Company will make the dataset very big and then the processing time for such dataset will be long. So, we may change the address of each labor by a number representing his/her location. It could be 1 for East location, 2 for West location, 3 for North location and 4 for South location. We reduce the data vertically by removing not related or weak related features. Also, we reduce the data horizontally by removing duplicate records. It is essentially reducing the size of the dataset to be mind and hence reducing the processing time and effort.

Figure 1 summarizes the knowledge discovery process which consists of three important engines. Preprocessing engine is represented by blue color part which is the biggest part because of its high importance. Its importance leads to better quality of discovered knowledge. Processing engine is represented by red color part and the after processing engine is represented by green color part. The size of each part represents the importance of each of them based on the difficulty of processing and implementation.

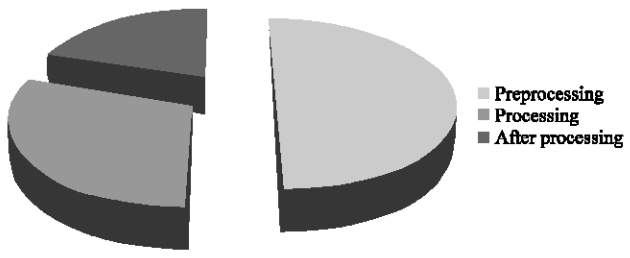

Fig. 1: The knowledge discovery engines. Preprocessing, processing and after processing steps

PREPROCESSING ENGIEN

Here, the three main important steps show that are usually used to prepare the dataset for the mining process. The discussed below gives some details of each.

Step 1: Choose the most suitable data for data mining: It is the first step which should be dealt with. The following characteristics of data are highly recommended:

Representative data: Training dataset should be representative. It should include all possible varieties of data. For example, if rules for diagnosing patients are being created and only male people are registered in the training set, the result for diagnosing a female based on these data probably will not be good. If the data is small then this problem could appear and this usually happen in machine learning. When using large datasets, the knowledge created which is presented as rules probably is representative, as long as the data being classified belongs to the same domain as those in the training set.

Data has boundary cases: One or more boundary cases should be present if the real differences between two different classes need to be found. For example, if an animal dataset is used and the target is to classify animals, the conditional attribute to determine if the class value is bird might be that it has wings and not that it can fly. This kind of detailed distinction will only be possible if e.g., penguins are stored in the dataset.

Data does not have limited information: In this case a data mining system does not have any way to distinguish between two types of records. This is usually happen when two records with the same values for condition attributes have a different classification value. It is clear to say that the two records have some properties, which are not among the attributes in the training set, but still make a difference. This could be arising if the dimension of the dataset is reduced and the deleted properties are the ones that make difference.

Step 2: Solving the noise in the dataset especially the missing data: One or more of the attribute values may be missing both for records in the training set and for records which are to be classified (Quinlan, 1986). Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns (Agrawal and Srikant, 2000).

The problem of missing values has been investigated since many times ago (Quinlan, 1989; Little and Rubin, 1987). The simple solution is to discard the data instances with some missing values (Liu *et al*., 1997). A more difficult solution is to try to determine these values (Kerdprasop *et al*., 2003). Several techniques to handle missing values have been discussed in the literature (Ragel and Cremilleux, 1999; Al-Shalabi, 2000; Kerdprasop *et al*., 2003; Little and Rubin, 1987; Liu *et al*., 1997).

Once the information about the records of missing values is known, the missing values can be replaced with appropriate values. There are several methods that can be used to find an appropriate value for a specific attribute that has missing value. Some methods are more appropriate if they get more information than others. The complexity of those methods is usually high and leads to time consuming especially if the training set is large. Using fast computer is not a practical solution. Other methods are powerful for specific datasets and not for others.

Some well known methods that can be found in practice are as follows:

- Exclude records that have missing data: this is usually done when the class label is missing. Also, it is recommended if the tuple contains several attributes with missing values.
- Replace all missing attribute values by a specific global constant, such as 'Missing', 'Unknown', '?', or others
- For quantitative values, use a generic figure, such as the mean or the mode value

    - Use the attribute mean value of all values in a specific column to fill in the missing value
    - Use the attribute mean value of all values of a specific column that belong to the same class
    - Use the most probable value of all values in a specific column to fill in the missing value
    - Use the most probable value of all values of a specific column that belong to the same class

- Use a specific algorithm or model to determine a value, One of these models is to split the original dataset into two new datasets; one contains records that have no missing values and the other one contains records that have missing values. Find the reduct of each of the two datasets and then merge the new reducts into one new dataset. The new dataset will have the minimum number of missing values that can be solved by any of the methods mentioned before. The dataset is then spitted into training dataset and testing dataset. Training dataset is then becomes ready to be trained (Al-Shalabi, 2009)

Step 3: Normalizing the data: Normalization is one of the important data transformation processes. It may improve the accuracy and efficiency of mining algorithms. Data mining algorithms provide better results if the data to be analyzed have been normalized (Al-Shalabi *et al.*, 2006a). Normalization process was studied by different researchers including Al-Shalabi *et al.* (2006a). Al-Shalabi *et al.* (2006a, b) were researched the importance of normalization process. They normalized a dataset using three different normalization methods and then they calculated the accuracy of the data mining system for each normalized dataset. They compared the results with the original dataset. Results showed that the accuracy of the trained normalized dataset is better than that if it is not normalized. They also compared between the three different methods of normalization.

In order to understand normalization, this example is given. If we have the age and sex attributes then we may find that the age attribute can take many more values than sex attribute. If we did not normalize the age attribute, then distance measurements taken on age will generally out weight distance measurements taken on sex attribute.

In normalization, the data are transformed into forms appropriate for mining. An attribute is normalized by scaling its values so that they fall within a small-specified range such as 0.0 to 1.0.

If neural network back propagation algorithm is used for classification mining, normalizing the input values for each attribute in the training set will help speed up the learning phase. For distance-based methods, normalization prevents attributes with large range from out weighting attributes with smaller ranges (Han and Kamber, 2000).

There are many methods for data normalization and present study mention here the most well-known methods such as mini-max, z-score and normalization by decimal scaling. Al-Shalabi *et al.* (2006a) summarize them as in the following.

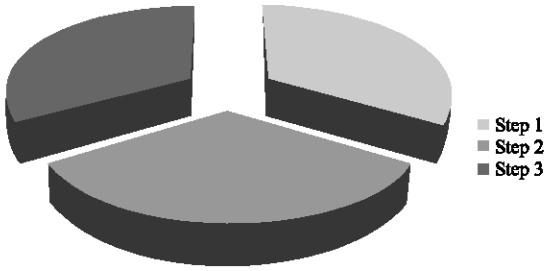

Fig. 2: The preprocessing steps

Min-max normalization: It performs a linear transformation on the original data. Suppose that min_a and max_a are the minimum and the maximum values for attribute A. Min-max normalization maps a value v of A to v' in the range [new-min_a, new-max_a] by computing:

$$v' = ((v\text{-}min_a)/(max_a\text{-}min_a)) * (new\text{-}max_a\text{-}new\text{-}min_a) + new\text{-}min_a \quad (1)$$

Z-score normalization: The values for an attribute, let say A, are normalized using the z-score normalization based on the mean and standard deviation of A. A value v of A is normalized to v' by computing:

$$v' = (v - \breve{A})/\sigma_A \quad (2)$$

where, $\breve{A}$ and σ_A are the mean and the standard deviation of attribute A. z-score normalization is useful when we know or when we can find the minimum and the maximum values of attribute A.

Normalization by decimal scaling: The value of attribute, let say A, is normalized by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing:

$$v' = v/10^j \quad (3)$$

where, j is the smallest integer such that $Max(|v'|)<1$. Normalization hides the original data that we don't want others to see. But we have to keep a record of the parameters used such as the mean and the standard deviation if using the z-score normalization and the minimum and the maximum values if using the min-max normalization so that future data can be normalized in the same manner.

Figure 2 summarizes the preprocessing steps as three different parts. Each part described by a color represents one of the three main steps of the preprocessing engine of knowledge discovery.

PROCESSING ENGIEN

This section shows the two main important steps that are to be taken into account when a mining process starts. The discussed below gives some details of each.

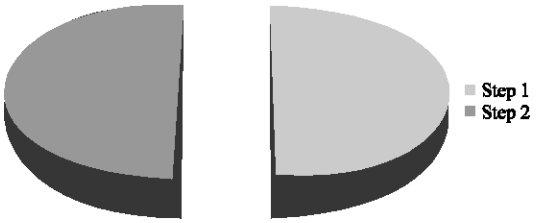

Fig. 3: Processing steps

Step 1: Choose the suitable data mining algorithm: The choice of a data mining algorithm is not an easy task. The best algorithm suite for a dataset may not be the most advanced algorithm and it may not be the one with that gives the highest accuracy. The important is to choose the algorithm that is simple, gives acceptable accuracy and able to perform all the required tasks.

Simple algorithm means that it is easy to be understood, does not need many calculations so that the complexity of the algorithm is low and it does not take long time for training process.

If we are about to use the data mining system to give an ordinary knowledge which is not a core to the company then the company may not interested much on the high accuracy especially if the cost is high. The company may agree on accepted accuracy but not the highest if it is possible to save money, time and effort. But if the data mining system required by the company is to find hidden valuable knowledge that may increase the benefits (money, time and image of the company) then the best algorithm might be the most accurate one.

There is no best data mining algorithm for all datasets (Al-Shalabi *et al.*, 2006b). Each algorithm has its strengths and weaknesses. So, each algorithm may be the best for a specific dataset or for particular needs in particular companies. Accuracy, functionality and cost of the algorithm are three main important points to be considered when choosing the data mining algorithm.

Step 2: Training process: In this step, a data mining algorithm will be applied to a chosen training dataset in order to train it and find the hidden knowledge that may exists inside it. The learnt model which could be represented by rules would be applied to the testing dataset which it had not been trained on. The accuracy of the data mining algorithm can then be measured by how many instances it correctly classifies.

Figure 3 summarizes the processing steps as two different parts. Each part described by a color represents one of the two main steps of the processing engine of knowledge discovery.

AFTER PROCESSING ENGIEN

This section shows the three main important steps to be considered after the mining process finished. The discussed below gives some details of each.

Step 1: Formalize a model based on the resulted rules from the previous step: If the structure of the rule resulted from the previous step is complex, then it is difficult to interpret it. The difficulty of interpreting the rule makes it uninterested and then it is better to ignore it. Rule length is a measure of simplicity. Rule length is defined as the number of concatenations in the rule. Rules whose lengths exceed a user-defined threshold can be considered uninteresting so it is better to be ignored.

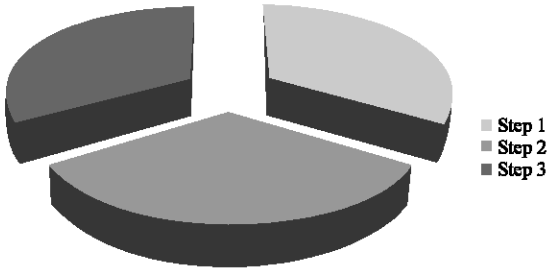

Fig. 4: After processing steps

The model is formed by putting the rules in some readable way which can be used by the user.

Step 2: Testing the findings (the knowledge): Check the accuracy and coverage of the resulted knowledge. Even of high percentage of accuracy and coverage, we still need to test the resulted knowledge to many different normal cases (the testing set) and special cases and then to take action based on the evaluation.

Step 3: Invest the findings: The discovered knowledge is the key point for management levels and decision makers to invest this knowledge and included it in their future investments.

Figure 4 summarizes the after processing steps as three different parts. Each part described by a color represents one of the three main steps of the after processing engine of knowledge discovery.

CONCLUSION

Figure 5 summarizes the eight steps of the whole process of knowledge discovery. It includes the three preprocessing steps, the two processing steps and the three after processing steps. As shown in Fig. 5, the preprocessing engine is of the most important since it controls the findings by its rich, clean and normalized data. The processing engine is of second important since it is a straight forward use of systems or algorithms. After processing engine comes after that and this

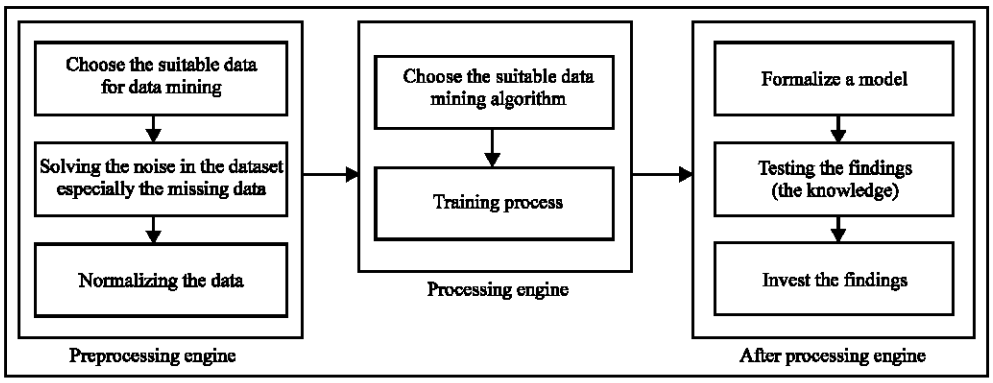

Fig. 5: The three engines of the knowledge discovery process

does not mean that it is not important engine but because it is the presentation of findings which is easily done by any kind of users who have little knowledge about knowledge discovery.

REFERENCES

Agrawal, A. and R. Srikant, 2000. Privacy preserving data mining. ACM SIGMOD Rec., 29: 439-450.

Al-Shalabi, L., 2000. New learning models for generating classification rules based on rough set approach. Ph.D. Thesis, University Putra Malaysia.

Al-Shalabi, L., Z. Shaaban and B. Kasasbeh, 2006a. Data mining: A preprocessing engine. J. Comput. Sci., 2: 735-739.

Al-Shalabi, L., M. Najjar and A. Al-Kayed, 2006b. A framework to deal with missing data in data sets. J. Comput. Sci., 2: 740-745.

Al-Shalabi, L.A., 2009. Improving accuracy and coverage of data mining systems that are built from noisy datasets: A new model. J. Comput. Sci., 5: 131-135.

Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. Morgan-Kaufman Publishers, New York.

Kerdprasop, N., K. Kerdprasop, Y. Saiveaw and P. Pumrungreong, 2003. A comparative study of techniques to handle missing values in the classification task of data mining. Proceedings of the 29th Congress on Science and Technology of Thailand, Oct. 20-22, Khon Kaen University, Thailand, pp: 1-3.

Little, R.J.A. and D.B. Rubin, 1987. Statistical Analysis with Missing Data. 1st Edn., John Wiley and Sons, USA., ISBN-10: 0471802549, pp: 304.

Liu, W.Z., A.P. White, S.G. Thompson and M.A. Bramer, 1997. Techniques for dealing with missing values in classification. Proceedings of the 2nd International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data, Aug. 4-6, Springer-Verlag London, UK., pp: 527-536.

Merz, C.J. and P.M. Murphy, 1996. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Pieter, A. and Z. Dolf, 1996. Data Mining. 1st Edn., Addison-Wesley Professional, Harlow, England, ISBN-10: 0201403803, pp: 176.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn., 1: 81-106.

Quinlan, J.R., 1989. Unknown attribute values in induction. Proceedings of the 6th International Workshop on Machine Learning, (IWML'89), Ithaca, New York, United States, pp: 164-168.

Ragel, A. and B. Cremilleux, 1999. MVC: A preprocessing method to deal with missing values. Knowl. Based Syst. J., 12: 285-291.

White, A.P., 1987. Probabilistic Induction by Dynamic Path Generation in Vertual Trees. In: Research and Development in Expert Systems III, Bramer, M.A. (Eds.). Cambridge University Press, UK., ISBN: 0-521-34145-X, pp: 35-46.