Volume 9, No. 3, March - 2018 (Special Issue)
International Journal of Mathematical Archive-9(3), 2018, 52-57
Available online through www.ijma.info ISSN 2229 – 5046

EXPLORATION OF MATHEMATICAL MODELLING IN DATA MINING

M. KANIMOZHI
Ph.D. Research Scholar (Computer Science)
Sri Sarada College for Women (Autonomous), Salem-16, India.

R. ROSELIN
Associate professor of Computer Science,
Sri Sarada College for Women (Autonomous), Salem16, India.

E-mail: kanismail16@gmail.com and roselinjothi@gmail.com

ABSTRACT

D*ata plays a potential role in making decision. This data can be mined to get some useful information. Here mathematics helps to organize the data for easy interpretation and efficient retrieval of knowledge which are hidden. Data mining process has several stages. In each stage, mathematics the queen of all science helps to do things more efficiently. Mathematical modelling helps in getting accurate results in all stages. This paper mainly covers some mathematical concepts and key ideas in the domain of breast cancer data set taken from UCI (University of California, Irvine) data repository.*

Keywords: *Data Mining, Mathematical Modelling, breast cancer data set.*

1. INTRODUCTION

Data mining is defined as extracting of knowledge from huge amount of data. Data mining involves the raw data, data management system, data pre-processing aspects and finding the pattern within the large amount of unexploring database. There are many other terms carrying a similar or a little different meaning to Data mining, such as knowledge mining from databases, knowledge extraction, data pattern analysis, data archaeology and data dredging. Data Mining is one of the steps in Knowledge Discovery in Databases (KDD). KDD and is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of interest in data [7]. Data mining pre-processing steps include the filling the missing value, removing the noisy data.The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal informationdata.Discretization which converts continuous values into discretized one, normalization which is used to scale the data and mapping the data. Normalization technique is used to find the new value from the existing data and required make them closer.

Feature selection is the process of finding the essential attributes among the given set of attributes.Featureselection tests the possible subset to find the one set which minimizes the error data[1][3]. Finally classification is used to classify the unseen samples. Supervised learning algorithm is used to classify the data[4]. This paper explores the few mathematical models involved in driving a classifier for breast cancer data set. The overall process is given in Figure 1.

CONFERENCE PAPER
National Conference dated 26th Feb. 2018, on "New Trends in Mathematical Modelling" (NTMM - 2018), Organized by Sri Sarada College for Women, Salem, Tamil Nadu, India.

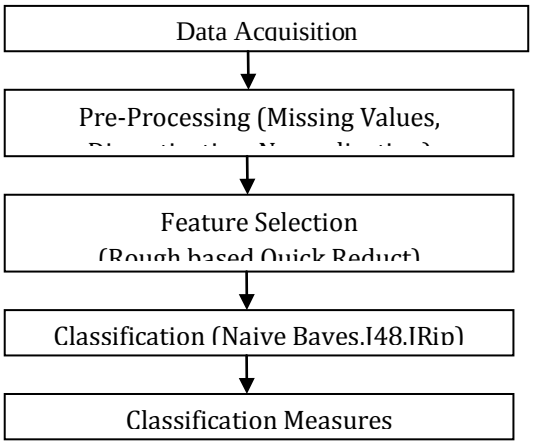

Figure-1: Overall Process

Mathematical models play an important role in identifying the pattern [2] [6]. This paper is organized as follows: Section 1 introduces the concept, Data set description is given in section 2, Data cleaning is discussed in Section 3, Section 4 discusses the process of feature selection with respect to quick reduct algorithm, Section 5 describes the classification models, Experimental results are discussed in Section 6 and Section 7 concludes this with the perspective to further research.

2. DATA SET DESCRIPTION

Breast Cancer Wisconsin (Original) data Set is taken from UCI data repository. It has 699 instances, 10 attributes and it also has missing values. The following are the attribute information:

1. Sample code number (F1) :id number
2. Clump Thickness (F2) : 1 - 10
3. Uniformity of Cell Size (F3) : 1 - 10
4. Uniformity of Cell Shape (F4) : 1 - 10
5. Marginal Adhesion (F5) : 1 - 10
6. Single Epithelial Cell Size (F6) : 1 - 10
7. Bare Nuclei (F7) : 1 - 10
8. Bland Chromatin (F8) : 1 - 10
9. Normal Nucleoli (F9) : 1 - 10
10. Mitoses (F10) : 1 - 10
11. Class : (2 for benign, 4 for malignant)

missing values are represented by '?'. Sample instances from breast cancer data set are given in Table 1.

Table-1: Breast Cancer Data Instances (Sample)

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Class
1183246	1	1	1	1	1	?	2	1	1	2
1183516	3	1	1	1	2	1	1	1	1	2
1183911	2	1	1	1	2	1	1	1	1	2
1183983	9	5	5	4	4	5	4	3	3	4
1184184	1	1	1	1	2	5	1	1	1	2
1184241	2	1	1	1	2	1	2	1	1	2
1184840	1	1	3	1	2	?	2	1	1	2
1185609	3	4	5	2	6	8	4	1	1	4

3. DATA PRE-PROCESSING

This is the important step in data mining. Raw data are needed to be cleaned before mining because it may contain noise. Some of the data may be inconsistent.

3.1 Data cleaning

This process removes noisy and inconsistent data from breast cancer data set

3.2 Smoothing

For smoothing the data, binning technique can be used. Binning methods smooth data value by considering its "neighborhood". The sorted values are distributed into a numberof "buckets" orbins. Smoothing can be done by mean, median and boundaries.

3.3 Outliers

A data set may contain objects that do not realize the general behavior or model of the data. These data are outliers. Many data mining methods reject outliers as noise or exceptions. Outliers may be detected and handled using statistical tests.This assumes a distribution or probability model for the data, or by using distance measures identify the outliers [7].

3.4 Missing values

Data are not always available incomplete form in real data set**.** It may be incomplete and noisy.Some tools ignore missing values; others use some metric to fill in replacements. The metric are:

i. Ignore the tuple.
ii. Fill in the missing value manually.
iii. Use a global constant to fill in the missing value.
iv. Use the attribute mean or median.
v. Use the attribute mean or median for all samples belonging to the same class as the given tuple.
vi. Use the most probable value to fill in the missing value.

To fill the missing value, the mean and median values are used for data replacement. The breast cancer data set has16missing values as 'NaN'. In this study mean value is computed to replace the missing values [8].

$$\text{Mean } (\bar{X}) = \sum_{i=1}^{n} \frac{X_i}{n}$$

Figure 2 shows the Missing values replaced by mean.

```
     2     1     1     1     2     1     1     1     1     2
     5    10    10     3     7     3     8    10     2     4
     4     8     6     4     3     4    10     6     1     4
     4     8     8     5     4     5    10     4     1     4


columnTot =

        3088        2191        2242        1962        2248        2421        2403        2004        1111        1880


colMean =

    4.4177    3.1345    3.2074    2.8069    3.2160    3.5447    3.4378    2.8670    1.5894    2.6896
```

Figure-2: Missing Values Replaced by Mean

3.5 Min-Max Normalization

Data set can also be in varying from, e.g. one attribute varies in range of 100 and other attribute varies in the range of 10000. So, a proper normalization of data set is done in which each attribute come in the range of 100 or whatever user selects. This normalization technique is known as Min-max normalization.

$$\mathbf{B} = \left(\frac{(\text{A}-\text{minimum value of A})}{(\text{maximum value of A}-\text{minimum value of A})}\right) * (\text{D-C})$$

The min-max normalization is pre-processing techniques to scale the data in the range of –1.0 to 1.0, or 0.0 to 1.0.The above formula applying for the min-max function then the original data convert the scale 0 to 1.

```
x =

    0.4444         0         0         0    0.1111    0.1000    0.2222         0         0
    0.4444    0.3333    0.3333    0.4444    0.6667    1.0000    0.2222    0.1111         0
    0.2222         0         0         0    0.1111    0.2000    0.2222         0         0
    0.5556    0.7778    0.7778         0    0.2222    0.4000    0.2222    0.6667         0
    0.3333         0         0    0.2222    0.1111    0.1000    0.2222         0         0
    0.7778    1.0000    1.0000    0.7778    0.6667    1.0000    0.8889    0.6667         0
         0         0         0         0    0.1111    1.0000    0.2222         0         0
    0.1111         0    0.1111         0    0.1111    0.1000    0.2222         0         0
    0.1111         0         0         0    0.1111    0.1000         0         0    0.4444
    0.3333    0.1111         0         0    0.1111    0.1000    0.1111         0         0
         0         0         0         0         0    0.1000    0.2222         0         0
    0.1111         0         0         0    0.1111    0.1000    0.1111         0         0
    0.4444    0.2222    0.2222    0.2222    0.1111    0.3000    0.3333    0.3333         0
         0         0         0         0    0.1111    0.3000    0.2222         0         0
    0.7778    0.6667    0.4444    1.0000    0.6667    0.9000    0.4444    0.4444    0.3333
    0.6667    0.3333    0.5556    0.3333    0.5556    0.1000    0.3333    0.2222         0
    0.3333         0         0         0    0.1111    0.1000    0.1111         0         0
    0.3333         0         0         0    0.1111    0.1000    0.2222         0         0
    1.0000    0.6667    0.6667    0.5556    0.3333    1.0000    0.3333         0    0.1111
    0.5556         0         0         0    0.1111    0.1000    0.2222         0         0
```

Figure-3: ThePartial result of Min-Max Normalization

3.6 Discretization

Data discretization and concept hierarchy generation are also forms of data reduction. The raw data are replaced by a smaller number of interval or concept labels. This simplifies the original data and makes the mining more efficient. For some attributes, if doctors have had existing dividing points, one can adopt it directly. For example, patients' weight can be divided to thin, common and heavy; their age can be divided into underage, youth and the elderly; Medical test results can be said to be normal and abnormal.

3.7 Rough Set based Feature Selection

Feature selection refers to the process of selecting those input attributes that are most predictive of a given outcome. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the features after reduction. Rough set theory (RST) can be used as tool to discover data dependencies and reduce the number of attributes contained in data set using the data along, requiring no additional information [5]. It is possible to find the set of the original attribute using RST that are most informative, all other attributes can be removed from the data set with minimal information loss. The benefits of feature selection are twofold: it considerably decreases the running time of the induction algorithm, and increases the accuracy of the resulting model.

4. QUICK REDUCT ALGORITHM

The quick reduct algorithm is used to calculate a reduct without considering all possible subsets. It starts up with an empty set and adds in turn, one at a timethose attribute result ingreatest increase in the rough set dependency metric, until this process maximum possible value for dataset. This algorithm calculated the dependency for each attribute and chooses the best candidate.

4.1 Reduct

Reduct is a minimal subset R of initial attribute set C(conditional) such that for a given set of decision attribute D [1]. Quick reduct algorithm reduced 10 attributes of breast cancer data set into 4 attributes. Id is removed before applying quick reduct algorithm. They are Clump Thickness (F2), Uniformity of Cell Size (F3), Bare Nuclei (F7) and Normal Nucleoli (F9).

5. CLASSIFICATION

Data classificationis a two-step process, consisting of a learning step and a classification step. In the first step, a classifier is built describing a pre-determined set of data classes or concepts. There aremany classification algorithms. This paper deals only with the following:

i. Decision tree classification
ii. Bayesian classification
iii. Rule based classification

5.1 Naive Bayes Classification

Naive Bayes model is probability based model. It uses prior probability. The method is simple and easy to build. Bayes theorem provides a way of calculating posterior probability, $P(C|x)$, from P(C), P(x) and $P(x|C)$.

5.2 J48 classification

Information gain is the measure used in building a decision tree. The table is split based on the information gain. Leaf nodes bear the class value when the splitting procedure stops if all instances in a subset belong to the same class.The model generated by J48 is given below:

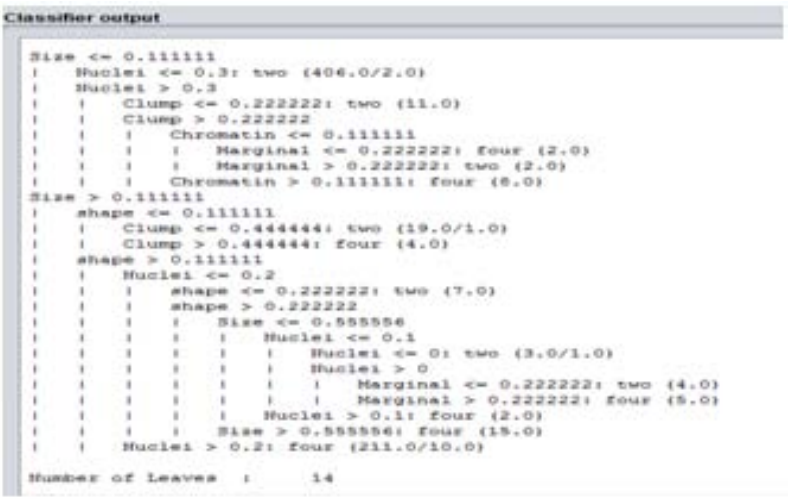

Figure-4: J48 Classifier

5.3 Rule based JRip Algorithm

JRip (RIPPER) is one of famous algorithms has its own popularity. Incremental reduced error is used to generate set of rules for the class.

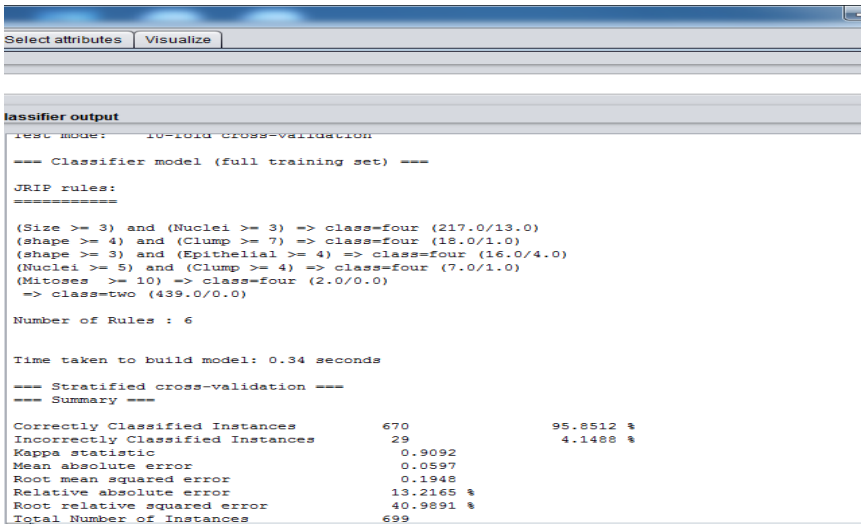

Figure-5: JRip Classifier

6. EXPERIMENTAL RESULTS

Pre-processing is done using MATLAB and the classification is done in WEKA. Table 2 reports the validation measures as per classifier.

Table-2: Validation Measures

Classification	Before QR	After QR	Before QR		After QR	
			Precision	Recall	Precision	Recall
Naivebaye	(436,235)	(441,237)	0.962	0.960	0.971	0.970
J48	(44,225)	(439,227)	0.956	0.956	0.953	0.953
JRip	(438,232)	(441,231)	0.959	0.959	0.962	0.961

The experimental results before and after quick reduct algorithm is illustrated in Figure 6. The experimental result shows that classification produces good result after feature selection for breast cancer data set.

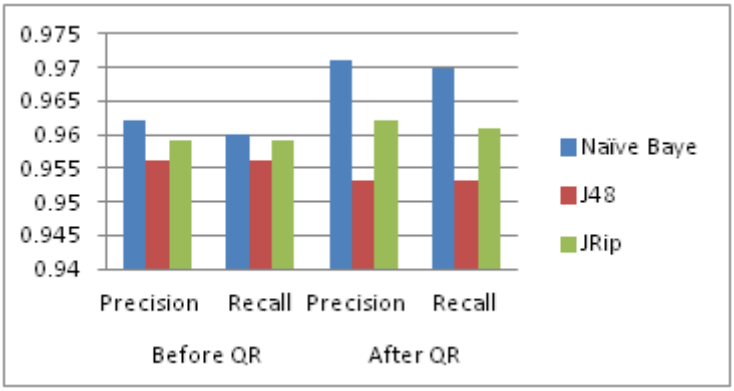

Figure-6: Comparative study of Precision and Recall

7. CONCLUSION

This paper explored only few models in data mining. In the stage of pre-processing missing values are filled and rough set based quick reduct algorithm has been applied to reduce the features. Then classification is done using Naive Bays, J48 and Ripper Classifiers. The resultant model enables one to classify the unseen data. Model is built over 699 instances. Increased number of instances obviously increases the classification accuracy. Without the model the dataset is useless. Constructed model will help in diagnosing the future patient and thereby add meaning to the data.

8. REFERENCE

1. Anitha K., Venkatesan P., "Feature Selection by Rough-Quick Reduct Algorithm**",** International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, pp. 3989-3993, 2013.
2. Elmegård Michael, Starke Jens**,** "Mathematical Modeling and Dimension Reduction in Dynamical Systems"**,** Ph. D. Thesis, Technical University of Denmark, ISSN 0909-3192, 2014.
3. KeerthikaT., Premalatha K., "Rough Set Reduct Algorithm based Feature Selection for Medical domain", Journal of Chemical and Pharmaceutical Sciences, Vol. 9, Issue 2,pp. 896-902, 2016.
4. Mario Rosario Guarracino, "On Classification Methods for Mathematical Models of Learning",High Performance Computing and Networking Institute Italian National Research Council, pp. 2-12, 2005.
5. Pawlak Z, "Rough Sets", International Journal of Computer and InformationSciences, Vol. 11, No. 5, pp. 341–356. 1982.
6. Suresh G., Selvakumar I.A., "An Overview of Data Mining Concepts Applied in Mathematical Techniques", International Journal of Computing Algorithm, Vol. 3, pp. 905-909, 2014.
7. **"**Data Preparation for Data Mining**"**, Dorian Pyle, 1999.
8. "Data Mining: Concepts and Technique", JiaweiHan and Micheline Kamber, 2000.

Source of support: National Conference on "New Trends in Mathematical Modelling" (NTMM - 2018), Organized by Sri Sarada College for Women, Salem, Tamil Nadu, India.

