International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 10
ISSN: 2454-4248
87 – 92

Performance of Lung Carcinoma in Classification Neural Network with Pre Processing Using WEGA

[1]S. Karthigai
[#1] Research Scholar in Computer Science
Erode Arts and Science College – Erode, India.
e-mail:skarthigai@yahoo.com

[*2] Dr. K. Meenakshi Sundaram
[#2] Associate Professor of Computer Science,
Erode Arts And Science College Erode, India.
e-mail:lecturekms@yahoo.com

Abstract— Data pre processing ease the mining procedure by removing the insignificant information and errors that may arise while entering the data manually. The data collection method is not strict so there accompanies missing and incorrect values, irrelevant variables, data with out of range etc. These have significant impact and minimize the accuracy of the mining process. Generally accuracy in the case of medical research must reach to the extent. There are many factors affect the analysis on the given task. The precise representation and quality of the dataset is vital. If there exists more irrelevant and redundant information the meaningful discovery of knowledge is a big question. Pre processing is a prominent way for the data preparation and thus it the earlier stage in mining. It includes many variant procedures according to the problem of the set. The output is taken as the direct training set for further research. This research analyse the Lung cancer dataset with fifteen attributes by applying pre processing method attribute evaluation. This method reduce the dimensionality, file size and time taken for the analysis by considering only on the most relevant variables. The work is carried in the WEKA tool as it has enormous procedures for data preparation. The performance before and after pre processing is discussed with suitable metrics.

Keywords – *Data Mining, Pre processing, Lung cancer dataset, Gain ratio attribute evaluation.*

*****

I. INTRODUCTION

Mining the data is the method to [7] discover information by analyzing massive set from various perception and extracting useful information via procedures. This is the most motivated research area to find of variant patterns. The main goal is to discover the knowledge hidden in data. Due to enormous growth of data, mining is at drastic level in each field. Getting the right information from data is the most challenging task. Many academicians and industry researchers are engaged on the process of knowledge mining due to abundance of data. It is the core step of Knowledge discovery procedure The recent aspects and development promotes the rapid growth of KDD and DM. Plain data is highly liable to noise, inconsistencies and missing values. Data pre-processing is the best solution to improve the quality of data which affects the product if it is not removed before analyzing the set.

Pre-processing is one of the most crucial steps in a mining process which has the concern about preparing and transforming the initial set. The medical data is often incomplete or lacking some entries due to the result of the test taken and it leads in certain errors. This may be inadequate to take right decisions sometimes. To elevate this issue, pre processing techniques gives the accurate solution.

Generally the healthcare organization generates a plenty of data which are in structured, un-structured and semi structured format. The healthcare data are collected from heterogeneous sources like hospitals, clinics, doctor's note, patient records. Thereafter transforming them into a single standardized format is a must and is done by numerous existing pre-processing techniques and methods.

Lung Carcinoma

'Lung cancer or Lung carcinoma', is a malignant tumor identified by [9] uncontrolled cell growth in the lung. This growth can penetrate beyond the lung by the process of 'metastasis' into nearby tissue or other parts. The two main types are Small-Cell (SCLC) and Non-Small-Cell lung carcinoma (NSCLC).

Most common symptoms for both types:

a. Coughing (including coughing up blood),
b. Tobacco smoking ,
c. Weight loss,
d. Shortness of breath and
e. Chest pains.

About ten to fifteen percent of cases occur in people who have never smoking habits. These cases are often caused by a combination of genetic factors ,second-hand smoke and air pollution. The diagnosis of cancer is confirmed by biopsy performed by bronchoscopes. Common treatments include surgery, chemotherapy, and radiotherapy. NSCLC needs surgery, whereas SCLC usually may be cured in chemotherapy and radiotherapy.

In a Worldwide survey that takes place in 2012, lung cancer occurred in 1.8 millions and resulted in 1.6 million life loss.

 ISSN: 2454-4248
87 – 92

II. PRE PROCESSING

The set of techniques used prior to the application of a mining method is coined as Data [8] Pre-Processing and it is known to be one of the most meaningful within the Knowledge Discovery. Handling medical data is a very complicated task since it comprises real of many attributes. The larger amounts of data collection in this area requires more sophisticated mechanisms. Data pre-processing is able to adapt the data as per the requirements posed by each algorithm, that would be unfeasible otherwise.

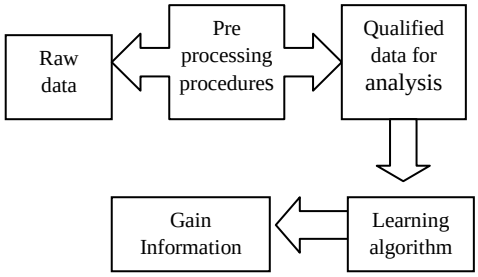


Figure 1. Pre processing

Figure 1. shows the pre processing steps that initialize from raw data to qualified one through pre processing procedures and latterly it is used by any kind of learning algorithm in mining.

Significance of preprocessing

Real world data are generally,

- Incomplete –
  Lacking attribute values, lacking certain attributes of interest, or contains only aggregate data.
- Noisy-
  Contains errors or outliers.
- Inconsistent –
  Contains discrepancies in names.

Categories in pre processing

A. Data Cleaning

The techniques in mining rely on a data set that mostly not complete or noise-free. But the existing data is far from being clean or complete. In data cleaning, the set is employed techniques to either re-moving the noisy data or fill the missing data. In medical records there are many imperfectness like missing certain fields because it must be filled by patients but it is impossible in case of emergency.

Some methods:

Fill in missing value, Rename the attribute.

B. Data integration

The info from variant sources are integrated into a single unit. They are in variant formats in variant sites. The sources may be 'Files, Spread Sheets, Documents, Internet' then on. Integration is a vital as the data are get from variant sources and doesn't match. It is really complicated to confirm whether the entries in variant sources have the same value or not. To reduce errors, integration process is used. The most common issue is redundancy. The same entry may be available in variant dataset or even in variant sources. The integration process tries to reduce redundancy without affecting the reliability.

Some methods:

Physical, Virtual, Manual Integration.

C. Data Transformation

Transformation is a process of transmitting source format into required destination format. Medical information collected from various sources with variant formats like database, XML or Excel sheet. The first step of transformation is mapping. It determines the relationship between the elements of two applications and list out instructions regarding how the data from the source is transformed before it is loaded into the target application. In other words, mapping creates metadata that is needed before the actual conversion takes place.

Some methods:

Smoothing, Aggregation, Generalization, Normalization etc.

D. Dimensionality Reduction

Mmining algorithms face the issue of dimensionality problem as the attribute become large in the number or the number of instances; It obstructs the function of most algorithms as the computational cost rise. Medical data reduction is bit complicated since every instance are plays an important role in some other analysis. The applied algorithm must be highly efficient to select the prominent feature.

Some methods:

Feature selection, Space Transformation, Instance reduction/selection,.

E. Discretization

Most used preprocessing techniques in the scientific community. It transforms quantitative into qualitative data by dividing the numerical features into non-overlapped intervals in a limited number. Using the boundaries, each numeric is mapped to each interval. Some algorithm works better with discretization methods than nominal data as many existing applications generally produce real valued outputs.

Some Methods:

Binning, Entropy based discretization.

 ISSN: 2454-4248
87 – 92

Problem definition

Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant or redundant to the mining task. Although it may be possible for a domain expert to pick out some of the useful attributes but this can be a difficult and time consuming task. Leaving out relevant attributes or keeping irrelevant attributes is mischievous, causing confusion. Medical data reduction is complicated since every instance plays an important role depends on the analysis.

Objectives

- To employ an effective method to know the importance of each attribute.
- To reduce the size of the data set.
- To reduce the analyzing time in mining by retaining only the most significant attribute.

III. LITRATURE REVIEW

Divya Tomar *et al* [1] made a survey on Pre processing and Post processing technique in data mining. The data may include several inconsistencies, missing values and irrelevant data. These all removed with the use of Pre-processing which is carried earlier. This paper elaborates the pre and post processing with various methods. Also it describes three visualization tools as it is vital in exploring the data.

Rattanawadee Panthong *et al* [2] put forth a method for feature selection. This paper use of wrapper feature selection sequential forward and backward selection which is the simplest greedy search algorithm. Thirteen datasets containing variant numbers of attributes and dimensions are obtained from the UCI Machine Learning Repository. This study shows that the search technique using SFS based on the bagging algorithm using Decision Tree obtained better results in average accuracy than other methods. The benefits of the feature subset selection are an increased accuracy rate and a reduced run-time when searching multimedia data consisting of a large number of multidimensional datasets.

Nasution *et al* [4] apply a Principal Component analysis for feature selection. It is a major problem in classification process called over-fitting results from noisy and irrelevant features. It also creates misclassification and imbalance in assessing. This research, proposed a framework for selecting relevant and non-correlated feature. The experiment is done with UCI Cervical cancer data set with thirty two attributes. It is reduced and the proposed framework is robust to enhance classification accuracy.

Priyanka Jindal *et al* [5] analyses existing feature selection and extraction techniques and addresses the benefits and challenges of these algorithms. The two types of dimensionality reduction techniques are Feature Selection and Extraction or Transformation. Feature selection is the process of removing irrelevant and redundant information where Feature extraction creates new feature subsets from original by applying some transformations which has more significant features. The new set have low dimensional than the previous one. Feature extraction or selection methods can be used separately or in combination. It improves the performance. Among the two methods of dimensionality reduction, feature extraction is more general.

Uma K, M.Hanumanthappa *et al* [6] proposed data collection and pre processing methods in health care set. This set contains structured and unstructured data and it combines text and images in a single file. This type of heterogeneous data needs a best pre-processing tool for the cleaning of data for further analysis. This paper discusses the collection method and types of pre processing technique and its need. For handling the missing value this paper employs the imputation method and finally data reduction method is used to eliminate the complexity and cost.

IV. METHODOLOGY

The proposed work employs Gain ratio attribute evaluation method for the reduction of dimensionality by eliminating the insignificant attribute [3] and retaining the best one.

Gain Ratio attributes evaluation

Gain Ratio estimate the ratio of the method Information Gain (IG). [12] Information gain assesses the importance a given attribute. Information gain ratio is a ratio of information gain to the intrinsic value Information gain is the measure of information gained by knowing the value of the attribute, which is calculated by the entropy of the distribution before the split minus the entropy after the split. The largest information gain [11] is equals the smallest entropy. The entropy is the average rate at which information is produced by a hypothetical source. The measure of information entropy is the negative logarithm of the probability mass function. Thus, when the data source has a lower probability value, the event carries more "information" than when the source data has a higher value.

Working Principle

It measures how much information an attribute gives with regarding the class. A common measure is Shannon entropy [10] for the information gain.

Entropy is calculated as,

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \quad \text{--- (1)}$$

Where,

X is a set of training samples, in the form,

$(x, y) = (x_1, x_2, \ldots,..x_k, y)$

where x_a € Val(a) the value of the a^{th} attribute of sample x and y is the class .

B is the base normally it is 2.

The entropy is calculated by taking the probability.

The Information gain is calculated with the above entropy as,

$$IG(X, a) = H(X) - H(X|a) \quad \text{--- (2)}$$

Intrinsic Information for a test set is calculated as,

$$\text{Intrinsic Info}(X, a) = -\sum_{i=1}^{n} \frac{|X_i|}{X} \log_b \frac{|X_i|}{X} \quad \text{--- (3)}$$

Gain ratio:

Information gain ratio is a [5] ratio of information gain to the intrinsic information.

$$\text{Gain Ratio}(X, a) = \frac{IG(X, a)}{\text{Intrinsic Info}(X, a)} \quad \text{--- (4)}$$

***Procedure for** Gain Ratio*

Step 1: Start with full data set.

Step 2: Calculate the entropy as (1) except for the class.

Step 3: Find the information gain as (2) by using entropy measurement.

Step 4: Calculate the Intrinsic value for the test set. as (3).

Step 5: Get the Gain ratio by (4).

Step 6: Select the feature with highest gain ratio.

Advantages

- A Good measure of finding the relevance of attribute.
- Gain ratio biases the decision against considering attributes with a large number of distinct values and thus solve the drawback of information gain.

Disadvantage

- Sometimes the most prominent attribute is discarded in final set for its lowest info gain.

V. EXPERIMENTAL RESULTS

The database is created in Microsoft excel sheet. The results are validated in WEKA 3.8.6. It expands as "Waikato Environment for Knowledge Analysis". Weka support only ARFF files. The file can be easily converted to ARFF format if it is CSV file.

A. Data set in Excel sheet

The Lung cancer dataset are collected from a medical practitioner. It consists of 14 attributes with a class and 3772 instances.

B. Data set before Pre processing

The dataset consists of 14 attributes with a class and 3772 instances.

Attributes:

The fifteen attributes are Patient id, gender, chronic cough, Hemoptysis, Pain in chest, Dysponia, Cachexia, Infection in lungs, Swelling, Wheezing, Dypsnea, Clubbing in nails, Dysphasia, Tumor location and a class label with four classes Adeno carcinoma, Squemous carcinoma, Large cell Carcinoma and Small cell Lung Carcinoma.

C. Dataset after Pre processing

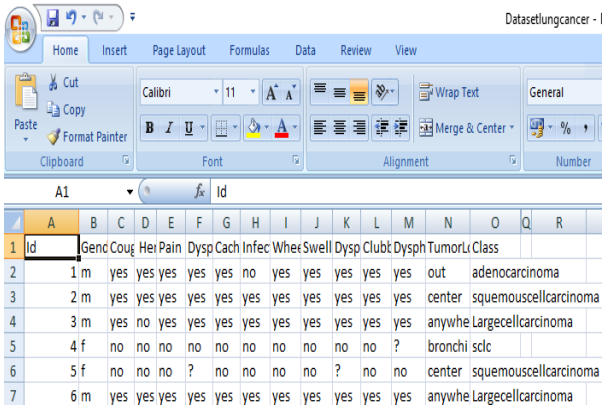
The dataset consists of 9 attributes with a class and 3772 instances.

Attributes:

The nine attributes are Patient id, gender, Hemoptysis, Dysponia, Cachexia, Wheezing, Dypsnea, Dysphasia, Tumor location and a class.

D. Results

Data set in excel



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Q	R
1	Id	Genc	Coug	Hei	Pain	Dysp	Cach	Infec	Whee	Swell	Dysp	Clubb	Dysph	TumorL	Class		
2	1	m	yes	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	out	adenocarcinoma		
3	2	m	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	center	squemouscellcarcinoma		
4	3	m	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	anywhe	Largecellcarcinoma		
5	4	f	no	no	no	no	no	no	no	no	no	no	?	bronchi	sclc		
6	5	f	no	no	no	?	no	no	no	no	?	no	no	center	squemouscellcarcinoma		
7	6	m	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	anywhe	Largecellcarcinoma		

Figure 2. Dataset in excel

Figure -. shows the dataset in excel sheet with 14 attributes and a class with four trials as Adeno carcinoma, Squemous carcinoma, Large cell Carcinoma and Small cell Lung Carcinoma.

D. Performance Evaluation

Table 1. Evaluation

Evaluation Measures	Before Pre processing	After Pre processing
Number of Attributes	14 + 1 class label	9 + 1 class label

ISSN: 2454-4248
87 – 92

Attribute	Id, Gender, CoughChronic, Hemoptysis, Pain, Dysphonia, Cachexia, Infection, Wheezing, Swelling,Dyspnea, Clubbing Nail, Dysphasia, Tumor location, Class.	Id, Gender, Hemoptysis, Dysphonia, Cachexia, Wheezing, Dyspnea, Dysphasia, Tumor location, Class.
File Size	253 Kb	189 Kb

Table 1 - show the evaluation measures before and after pre processing.

Dataset before Pre processing

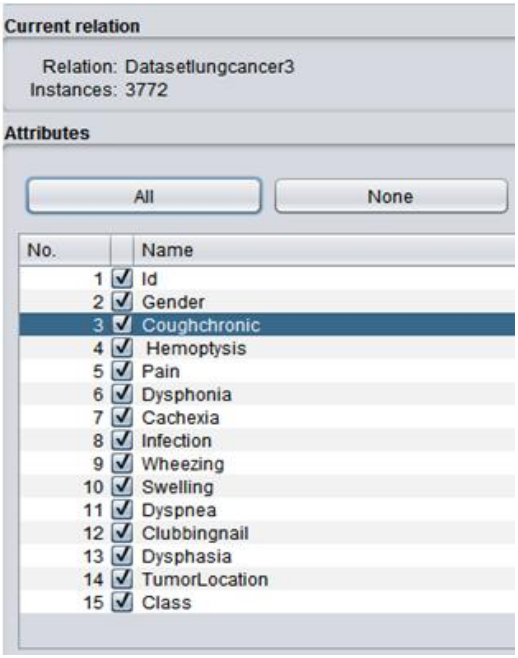


Figure 3. Before Pre processing

Figure 3 - show the data in WEKA tool before pre processing.

Dataset in WEKA after Pre processing

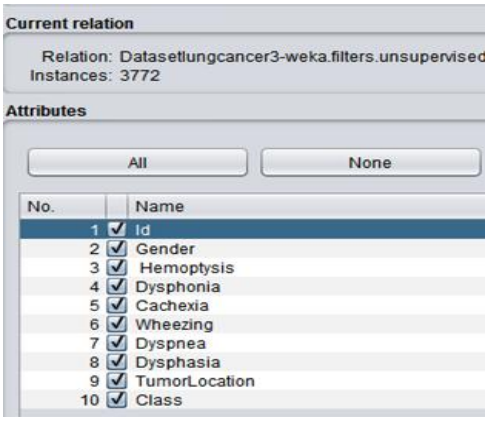


Figure 4. After Pre processing

Figure 4 - shows the nine attributes after pre processing with the method Gain ratio.

E. Represent in the chart

The evaluation result show in the reducing the file size and the attributes at chart representation.

Chart 1

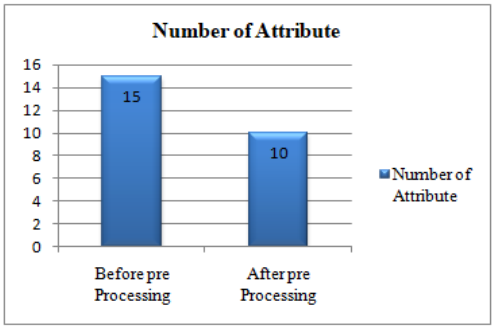


Chart 1. Number of Attributes

Chart 1- shows the comparison of number of attribute and file size before and after pre processing.

Chart 2

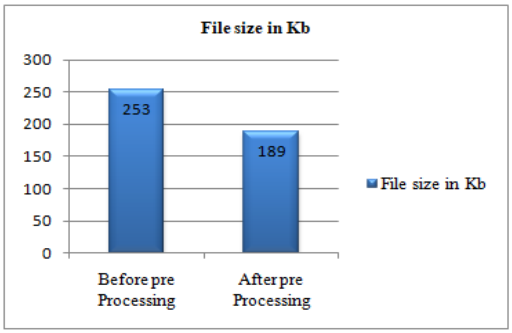


Chart 2. File size

Chart 2 - shows the File size in kilobytes before and after pre processing.

E. Reduce the file size

The appraisal result show in the reducing the file size by the number of attributes at represent in the properties window.

Before pre processing in the property window

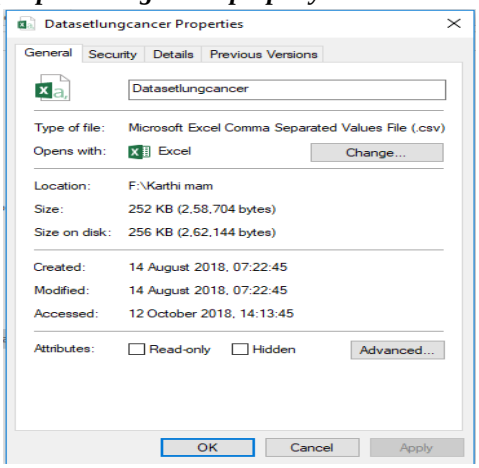


Figure 5- shows the fourteen attributes before pre processing in the property window.

 ISSN: 2454-4248
87 – 92

After pre processing in the property window

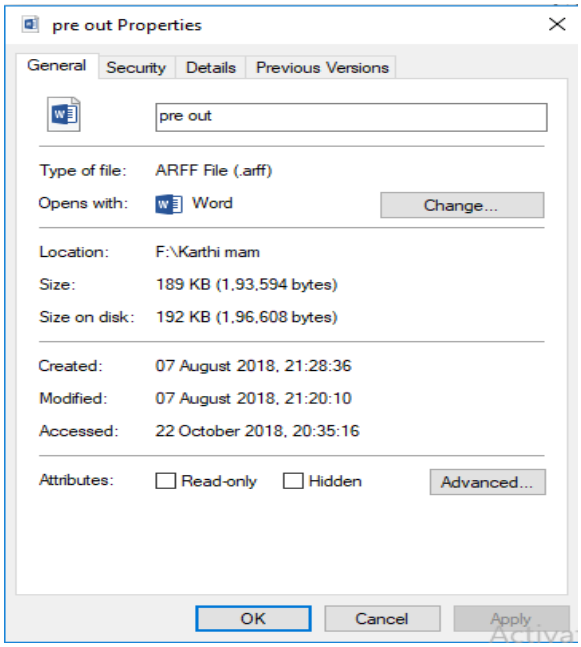


Figure 6- shows the Nine attributes after pre processing in the property window.

VI. CONCLUSION AND FUTURE WORK

Data preprocessing eliminates the irrelevant attribute and eases the mining procedure by various methods. Generally the medical records are entered manually with the attribute that is compatible for the hospital. Not all the information is needed for the analysis and it depends on the area of analysis. This paper categorizes the Lung carcinoma patients with fourteen attributes and a class before pre processing. In pre processing Gain Ration attribute evaluation method is applied on the dataset and the procedure eliminates five attributes which have the least information gain. The work is carried in the WEKA 3.8.2 tool as it has enormous procedures for data pre processing. The performance before and after pre processing is discussed with suitable metrics.

In future this pre processed set can be used for further mining process such as classification, clustering and so on.

REFERENCES

[1]. Divya Tomar, Sonali Agarwal, " A survey on Pre processing and Post Processing technique in data mining", Intrnational Journal of Database theory and applications, Vol 7, No.4, 2014.

[2]. Rattanawadee Panthong, Anongnart Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm", Procedia Computer Science 72 , 2015 .

[3]. A. H. Shahana, V. Preeja, " Survey on Feature Subset Selection for high dimensional data", In proceedings of International Conference on Circuit, Power and Computing Technologies, pp. 1-4, 2016.

[4]. M Z F Nasution , O S Sitompul and M Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification", 2nd International Conference on Computing and Applied Informatics 2017.

[5]. Priyanka Jindal, Dharmender Kumar," A Review on Dimensionality Reduction Techniques", International Journal of Computer Applications Volume 173 – No.2, September 2017.

[6]. Uma K, M. Hanumanthappa, " Data Collection Methods and Data Pre-processing Techniques for Healthcare Data Using Data Mining", International Journal of Scientific & Engineering Research Volume 8, Issue 6, June-2017.

[7]. J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[8]. J.Gama, " Knowledge discovery from data streams", Chapman and Hall /(RC, 2010.

[9]. www. Lungcancer.wikipedia.

[10]. en.wikipedia.org/wiki/Entropy

[11]. en.wikipedia.org/wiki/Mutual_information

[12]. en.wikipedia.org/wiki/Information_gain_ratio.