Journal Homepage: -www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/4285
DOI URL: http://dx.doi.org/10.21474/IJAR01/4285
ISSN NO. 2320-5407

RESEARCH ARTICLE

COMPARATIVE STUDY OF VARIOUS METHODS OF CLASSIFICATION TECHNIQUES USING DIFFERENT DATASETS

Varsha C. Pande[1] and Dr. Abha S. Khandelwal[2].
1. Research Scholar, Department of Electronics and Computer Science, RTMNU, Nagpur.
2. Department of Computer science, Hislop college ,Nagpur Maharashtra, India.

Manuscript Info

Manuscript History

Received: 28 March 2017
Final Accepted: 30 April 2017
Published: May 2017

Key words:-
Data mining algorithms, Weka tool, algorithms, classification methods etc.

Abstract

Data mining or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a data mining tools. It is contain the many machine leaning algorithms. It is provide the facility to classify our data through various algorithms. In this paper we are studying the various classification algorithms. Classification is one of the important features of data mining as a technique for modeling of forecasts. In other words, classification is the process of dividing the data to some groups that can act either dependently or independently. Our main aim to show the comparison of the various classification algorithms with weka on various datasets and find out which algorithm will be most suitable for the users.

Introduction:-

Data mining [1] is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Data mining is a relatively new technology that has not fully matured. Despite this, there are many industries that are previously using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools.

Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions.

In general, data mining contains several techniques and algorithms for finding out interesting patterns from large data sets. Techniques of data mining are classified into two categories ie. Supervised learning and unsupervised learning.

Corresponding Author:-Varsha C. Pande.
Address:-Research Scholar, Department of Electronics and Computer Science, RTMNU, Nagpur.

In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to assess the parameters of the model. Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are examples of supervised learning.

In unsupervised learning, the algorithm is directly applied to the dataset and the results are observed, here we do not create a model or hypothesis prior to the analysis, after that model can be created on the basis of the obtained results. Example of unsupervised learning is Clustering. Various data mining techniques such as Decision Tree, Classification, Bayesian Classification, Clustering, Association Rule Mining, Sequential Pattern, Neural Networks, Prediction, Time Series Analysis, Genetic Algorithm and Nearest Neighbour have been used for knowledge discovery from large data sets.

Classification:-

In data mining [2], Classification is the most commonly applied technique, which occupies a set of pre-classified examples to develop a model that can classify the population of records at large. Credit-risk and Fraud detection applications are particularly well appropriate to this type of analysis. This approach is frequently employs on decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning, the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable, the rules can be applied to the new data tuples. A fraud detection application would include the complete records of both valid activities and fraudulent determined on a record-by-record basis. These pre-classified examples to determine the set of parameters required for proper discrimination is used on the classifier training algorithm. The algorithm then encodes these parameters into a model called a classifier.

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires that the classes be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. This section discusses some of the useful text classification algorithms such as **ZeroR, OneR, J48, and Naïve Bayes.**

OneR Algorithm:-

OneR[3] is a classification algorithm whose main goal is to generate a rule based on single predictor, which is achieved that creates a rule for each predictor and based on the error rate for each predictor, selects one rule that has the lowest error rate. This is a simple algorithm which has its basis on a single level decision tree. One R classifier is fairly accurate in spite of being very simple. To create a rule for a predictor, we have to construct a frequency table for each predictor against the target. We start first by selecting a single predictor attribute. OneR Algorithm for each predictor, for each value of that predictor, make rule as follows:-

- Count how often each value of target (class) appears.
- Find the most frequent class.
- Make the rule assign that class to this value of the predictors.
- Calculate the total error of the rules of each predictor.
- Choose the predictor with the smallest total error.
- Find the best predictor which possesses the smallest total error using OneR algorithm.

ZeroR Algorithm:-

ZeroR [4] is the simplest classification method which relies on the target and ignores all predictors .ZeroR classifier [5] simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. Algorithm Construct a frequency table for the target and select its most frequent value. Predictors Contribution There is nothing to be said about the predictors contribution to the model because ZeroR does not use any of them.

Model Evaluation the ZeroR only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods.

Naive Bayes Algorithm:-
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms [6], a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)}$$

$$P(c/x) = P(x1/c) \times P(x2/c) \times \ldots\ldots \times P(xn/c) \times P(c)$$

Above
- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

J48 Algorithm:-
J48 [7] is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Basic Steps in the Algorithm:-
1. In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labelling with the same class.
2. The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.
3. Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

All the above classification techniques are implemented using WEKA. WEKA stands for **Waikato Environment for Knowledge Analysis.**
Weka[8] is a Java based open source tool data mining tool which is a collection of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering, and association rule extraction

Weka provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support pre-processing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by weka .

Implementation:-
In this study, many classification algorithms have been implemented on various inbuilt data sets and the performance of this algorithm has been analyzed by the data mining tool WEKA.
- In the starting interface of weka, click on the button Explorer.
- Process for applying dataset
- In the Preprocess tab, click on the button Open File.

- ❖ In glass dataset we are applying four methods (ZeroR, OneR, J48, Naive Bayes).
- ❖ In the file selection interface, select the file glass.arff.

ZeroR Method:-

The **ZeroR** method is selected by default. For assessing the predictive performance of all models to be built, the 10-fold cross-validation (by default) and percentage split (60%) has also be specified and result is shown in figure1.

OneR Method:-

The **OneR** method is selected For assessing the predictive performance of all models to be built, the 10-fold cross-validation (by default) and percentage split (60%) has also be specified ,in figure2 shows the results.

J48 Method:-

The **J48** method is selected. For assessing the predictive performance of all models to be built, the 10-fold cross-validation (by default) and percentage split (60%) has also be specified and shws the output window in figure3.

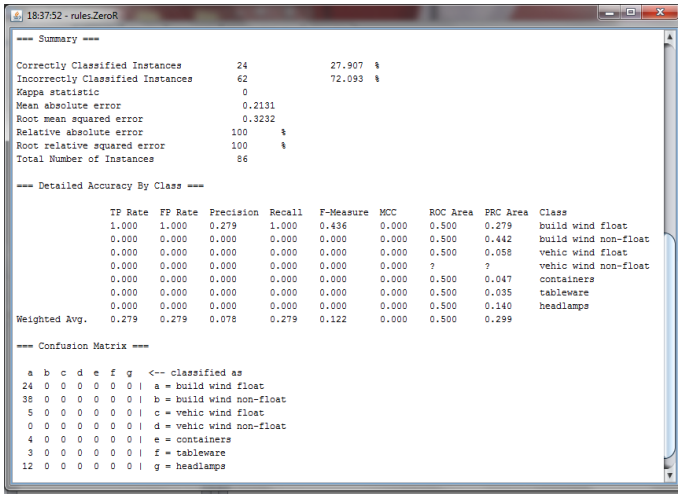

Fig1:- Output of ZeroR method on glass dataset

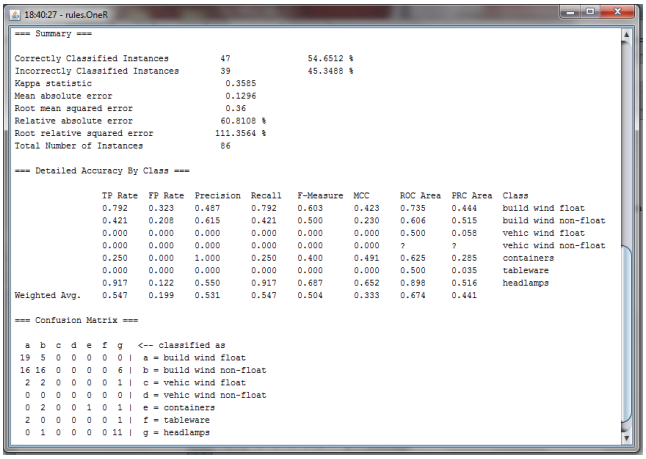

Fig2:- Output of OneR Method on glass dataset.

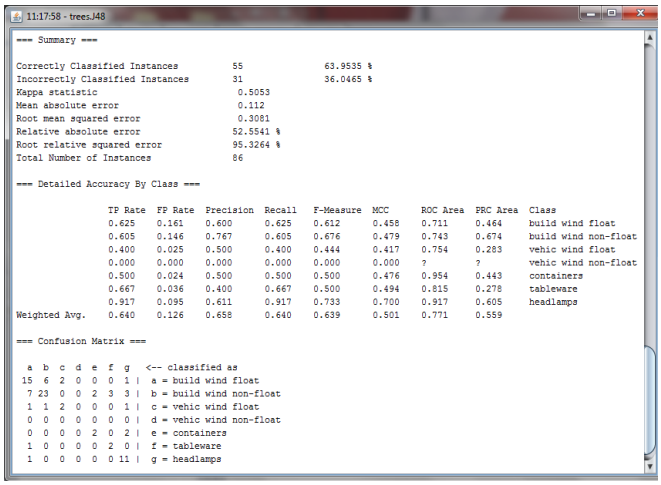

Fig3:- Result of J48 Method on glass dataset.

Naives Bayes Method:-

The **Naives Bayes** method is selected. For assessing the predictive performance of all models to be built, the 10-fold cross-validation (by default) and percentage split (60%) has also be specified. And result shows in following firure 4.

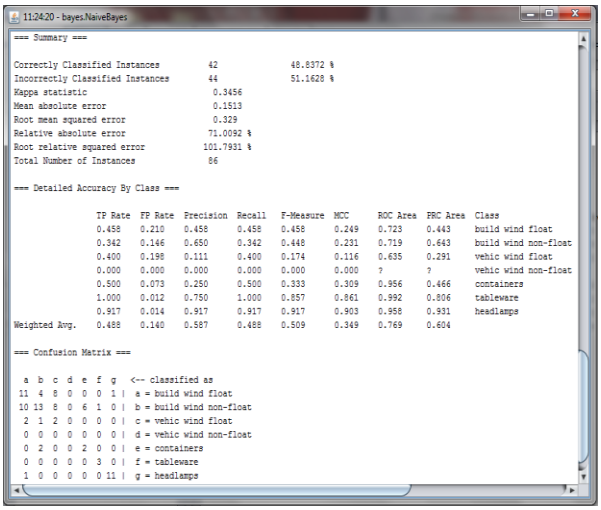

Fig4:-Result of Naives Bayes Method on glass dataset.

Simillarly, we are applying four methods (ZeroR, OneR, J48, Naive Bayes) in **Weather dataset** and the following results are shows in figure5, figure6, figure7 and figure8 respectively.

- In the file selection interface, select the file **weather.numeric.arff .**

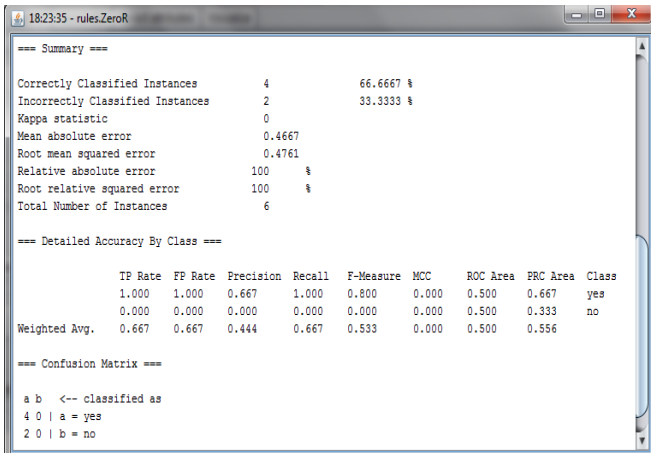

Fig5:- Result of ZeroR Algorithm on Wether dataset

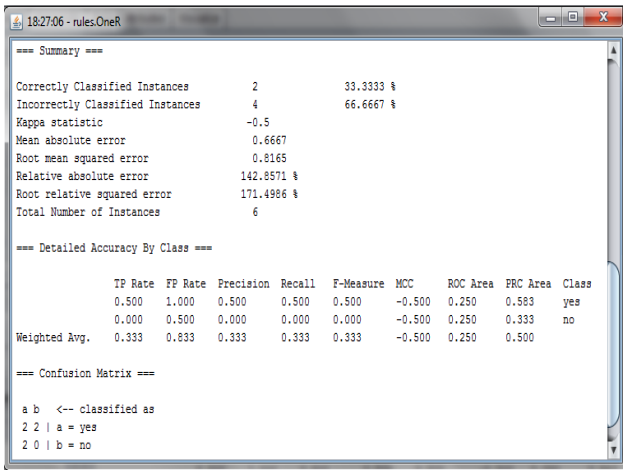

Fig6:-Result of OneR Method on weather dataset.

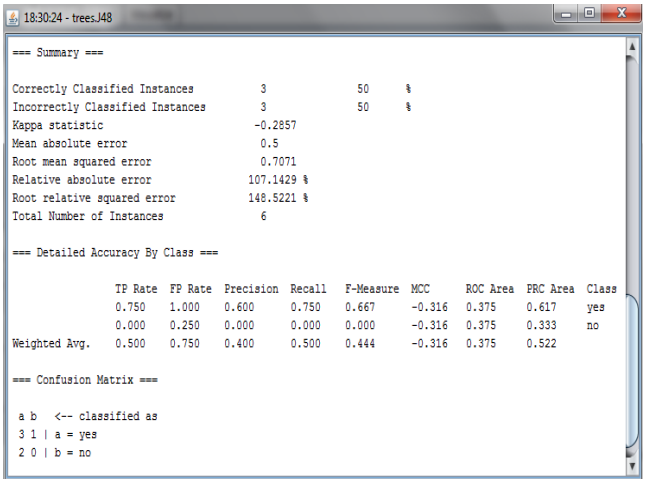

Fig7:-Result of J48 Method on weather dataset.

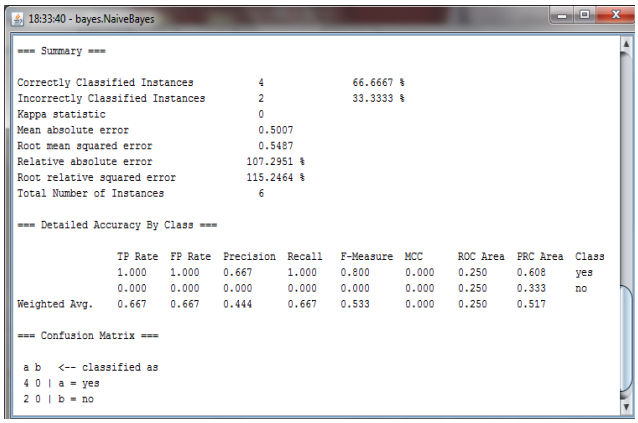

Fig8:-Output of Naives bayes Method on weather dataset.

As same , we are also applying four methods (ZeroR , OneR , J48, Naive Bayes) in the **Unbalanced dataset** and the results are shows in figure9, figure10, figure11 and figure12 respectively.

- In the file selection interface, select the file **unbalanced.arff .**

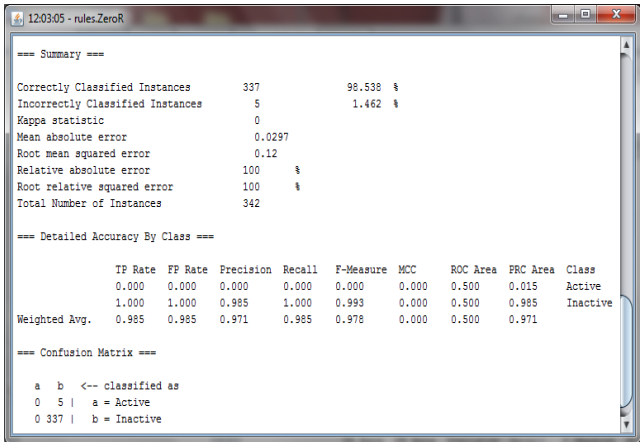

Fig9:-Output of ZeroR Method on unbalanced dataset.

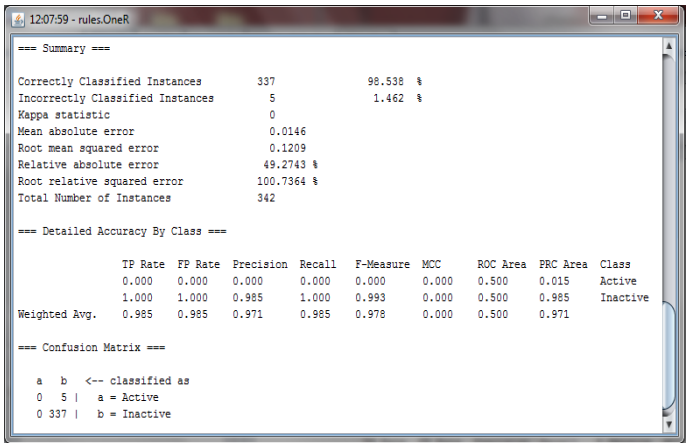

Fig10:- Output of oneR Method on unbalanced dataset.

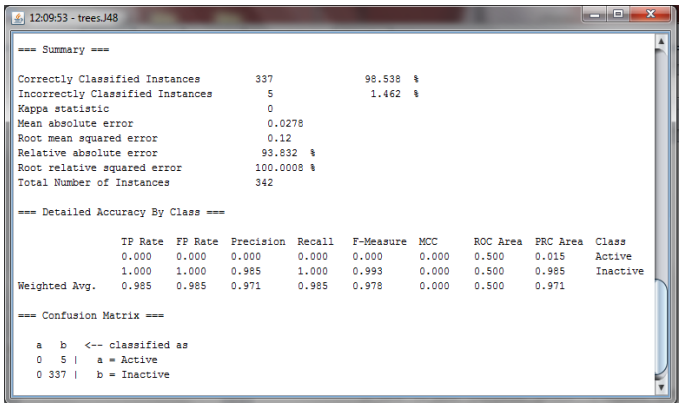

Fig11:-Output of J48 Method on unbalanced dataset.

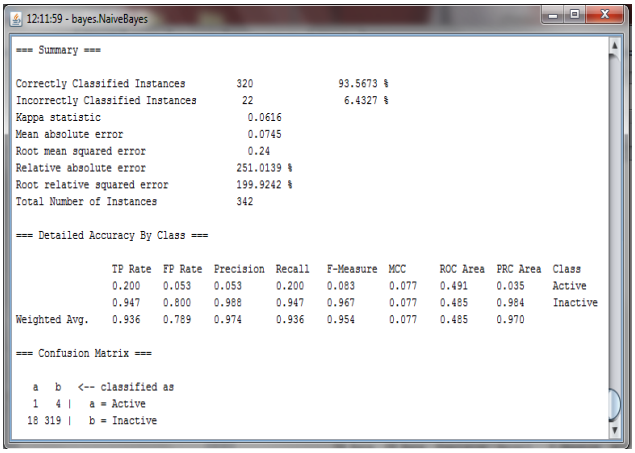

Fig12:- Output of Naive Bayes on unbalanced dataset.

Performance Evaluation:-

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$Precision = TP/(TP+FP)$$

- **Recall:** completeness – what % of positive tuples did the classifier label as positive?

$$recall=TP/(TP+FN)$$

- Perfect score is 1.0
- Inverse relationship between precision & recall
- *F* **measure (***F1* or *F*-**score)**: harmonic mean of precision and recall,

$$F=2\times(precision \times recall)/ (precision + recall)$$

The following table (table 1) shows the results for J48 Algorithm

Table 1:- Results For Zeror Algorithm.

Dataset	Glass	Weather	Unbalanced
Precision	0.09	0.49	0.977
Recall	0.3	0.7	0.988
F-measure	0.138	0.576	0.983

Performance of ZeroR Algorithm based on Datasets (table 1) is shown in figure 18:-

In performance of zeroR algorithm the values of precision (0.09, 0.49, 0.977), Recall (0.3, 0.7, 0.988), F-measure (0.138, 0.576, 0.983) are according to the datasets (Glass, Weather, Unbalanced) respectively. In which recall value is better for all the data sets as compared to other two values (Precision and F-measure) of dataset. An according to Graph, **Unbalance dataset** performance is **better** than other two data sets

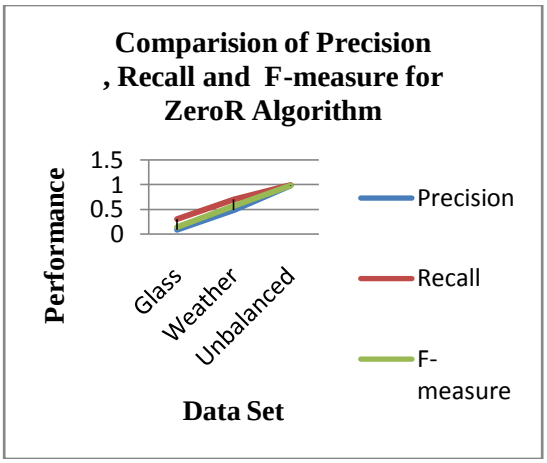

Figure 13: Performance of ZeroR Algorithm

The following table (table 2) shows the results for OneR Algorithm

Table 2:- Results For Oner Algorithm

Dataset	Glass	Weather	Unbalanced
Precision	0.335	0.4	0.977
Recall	0.407	0.4	0.988
F-measure	0.365	0.4	0.983

Performance of OneR Algorithm based on Datasets (table 1) is shown in figure 14:

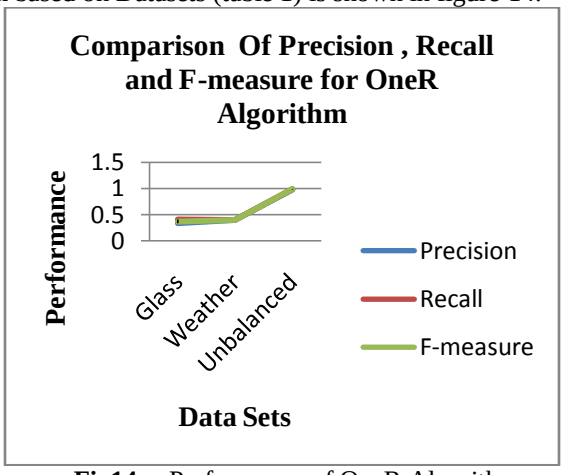

Fig14:- Performance of OneR Algorithm.

In performance of OneR algorithm the values of precision (0.335, 0.4, 0.977), Recall (0.407, 0.4, 0.988), F-measure (0.365, 0.4, 0.983) are according to the datasets (Glass, Weather, Unbalanced) respectively. In which recall value is better for all the data sets as compared to other two values (Precision and F-measure) of dataset. An according to Graph, **Unbalance dataset** performance is **better** than other two data sets.

The following table (table 3) shows the results for J48 Algorithm

Table 3:- Results For J48 Algorithm

Dataset	Glass	Weather	Unbalanced
Precision	0.574	0.49	0.977
Recall	0.56	0.7	0.988
F-measure	0.554	0.576	0.983

Performance of J48 Algorithm based on Datasets (table 1) is shown in figure 15:

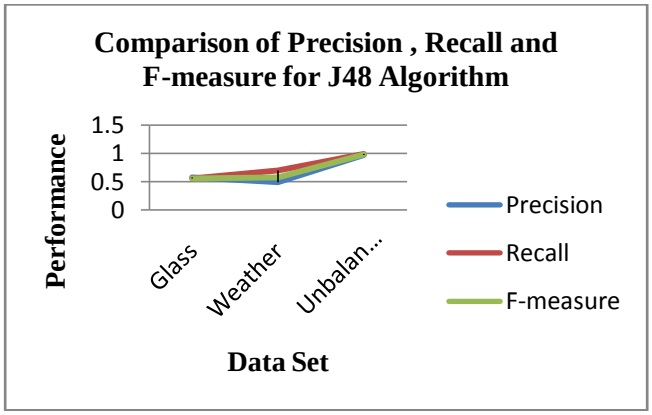

Fig15:- Performance of J48 Algorithm

In performance of J48 algorithm the values of precision (0.574, 0.49, 0.977), Recall (0.56, 0.7, 0.988), F-measure (0.554, 0.576, 0.983) are according to the datasets (Glass, Weather, Unbalanced) respectively. In which recall value is better for all the data sets as compared to other two values (Precision and F-measure) of dataset. An according to Graph, **Unbalance dataset** performance is **better** than other two data sets.
The following table (table 4) shows the results for J48 Algorithm

Table 4:- Results For Naive Bayes Method

Dataset	Glass	Weather	Unbalanced
Precision	0.09	0.49	0.977
Recall	0.3	0.7	0.988
F-measure	0.138	0.576	0.983

Performance of **Naive Bayes** Algorithm based on Datasets (table 4) is shown in figure 16:

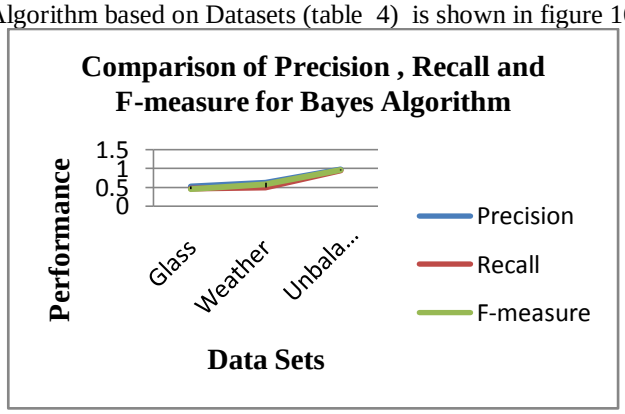

Fig16:- Performance of Naive Bayes Algorithm

In performance of Naives Bayes algorithm the values of precision (0.521, 0.625, 0.979), Recall (0.473, 0.5, 0.953), F-measure (0.453, 0.575, 0.965) are according to the datasets (Glass, Weather, Unbalanced) respectively. In which precision value is better for all the data sets as compared to other two values (Recall and F-measure) of dataset. An according to Graph, **Unbalance dataset** performance is **better** than other two data sets.

Now Finally we find the Accuracy graph of four methods on glass, weather and unbalanced dataset(table 5) shows in figure 17.

Table 5:- Accuracy Table Based On Dataset

Dataset	Glass	Weather	Unbalanced
ZeroR	0.3	0.7	0.98
OneR	0.57	0.4	0.98

J48	0.68	0.7	0.98
Naive Bayes	0.57	0.5	0.95

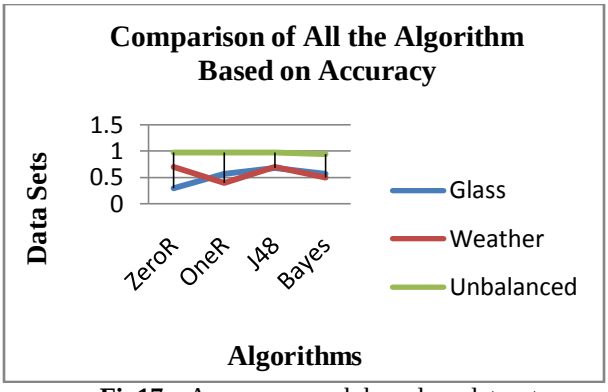

Fig17:- Accuracy graph based on dataset:

Conclusion:-

We are using data mining techniques in mainly in the medical, banking, insurances, education etc. The classification algorithms ZeroR, OneR, J48 and Naïve Bayes have their own importance and we use them on the behavior of the different datasets, but on the basis of this research we found that J48 classification algorithm is simplest algorithm as compared to other algorithms.

The different classification algorithms are studied and implemented using WEKA. The implementation results show the values for Precision, Recall and F-measure. The overall results for the entire algorithms are shown in accuracy table based on dataset.

The Overall Performance of all the algorithms (ZeroR , OneR , J48, Bayes Algorithms) is better for **Unbalanced Dataset** rather than other two datasets.

References:-

1. Mrs. Bharati M. Ramageri, *,"Data Mining Techniques And Applications*" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-3051.
2. Article: "Data Mining Techniques Crucial Concepts in Data Mining" Stat Soft Electronic Book Web Source.
3. http://chem-eng.utoronto.ca/~datamining/dmc/oner.htm .
4. Chitra Nasa and Suman, Evaluation of Different classification
5. Techniques for WEB Data, *International Journal of Computer Applications (0975 – 8887) Volume 52– No.9, August 2012.*
6. http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm
7. http://chem-eng.utoronto.ca/~datamining/dmc/naivebayesian.htm.
8. Korting Thales Sehn,"C4.5algorithm and Multivariate Decision Trees", Image Processing Division, National Institute for Space Research—INPE.
9. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition, Morgan Kaufmann, San Francisco(2005).