Comparing Various Classification Techniques Through Weka for Ovarian Cancer

Priyanka Khare, Kavita Burse, Anjana Pandey

Abstract: - In today's world, enormous amount of data is presented in various fields. This data can provide important and helpful information for making important decisions. Data mining is the method of finding valuable information. There are numerous data mining techniques used for extracting information classification is one of them. Classification is the process of classifying various data according to established criteria. In this paper, various classification algorithms are used for classifying the data set before these relevant features are selected by the process of feature selection. The performance of various classifiers is analyzed on the basis of accuracy and time taken to build the model.

Keywords: - feature selection, classification, weka, Interquartile range, navies bayes, instance based learning (IB1) ,k-nearest neighbour (IBK) , K-STAR, logical analysis of data(LAD) Tree

I. INTRODUCTION

A variety of classification tasks need learning of suitable classification function that assigns a known input to one of a fixed set of classes. Feature selection is known to be an important step in the design of pattern classifier for several reasons. Feature selection is used for selecting best feature for improving performance in classification. In many applications like medical diagnosis, text classification there is a practical requirement to decrease the number of measurements without much corrupting the performance of the system [1]. Feature selection methods simple choose subsets of feature from less dimension data comparing with high dimension data. The feature selection is the task of identifying and selection a useful split of features to be used to represent patterns from a larger set of often commonly redundant, possibly irrelevant, features with different associated measurement risks [2]. Statistical model fitting or supervised learning systems normally do not have sufficient labelled training instances to fit accurate models over very large feature spaces, due to finite sample effects [3]. At the same time, in many cases it is not easy to know without training which features are relevant to a given task and which are effectively producing. As a result, the skill to select features from a huge feature set is essential for application of datato compute algorithm. This work proposes a method using genetic algorithm to identify subset of features and for accuracy better classification algorithms are used through weka.

Revised Version Manuscript Received on March 14, 2016.

Priyanka Khare, M.Tech. Scholar, Department of Computer Science and Engineering, Oriental Institute of Science and Technology, Bhopal (M.P). India.

Dr. Kavita Burse, Director, Oriental Institute of Science and Technology, Bhopal (M.P). India.

Dr. Anjana Pandey, Assistant Professor, University Institute of Technology RGPV, Bhopal (M.P). India.

II. PROPOSED FRAMEWORK

2.1 Feature Choice Using Genetic Algorithm

Feature choice is analogous to data pre-processing technique. In this approach subset of features are identify with target model The aim of variable choice is to increase the level of accuracy, reduce dimensionality; shorter training time and enhances generalization by reducing over fitting. Feature selection techniques are a subset of general field of feature extraction. Feature extraction is used to make novel features from functions of the original features, while feature collection proceeds a set of the options. Feature selection techniques return a subset of features. Feature selection is utilized in area where there are few sections and comparatively many features.[4]
A genetic algorithm (GA) is a search and optimization method, which works by mimicking the evolutionary principles and chromosomal processing in natural genetics. A GA begins its search with a random set of solutions usually coded in binary strings. Every solution is assigned a fitness, which is directly related to the objective function of the search and optimization problem. Thereafter, the population of solutions is modified to a new population by applying three operators similar to natural genetic operators. Reproduction choose fine string; crossover merge good strings to attempt to produce better offspring's; mutation modify a string nearby to try to form a superior string.[5]It works iteratively by successively applying these three operators in each generation until a termination criterion is satisfied. Over the past decade and more, GAs have been successfully applied to a wide variety of problems, because of their simplicity, global perspective, and inherent parallel processing.

2.2 Interquartile Range

The Interquartile range (IQR) is a compute of inconsistency, depend on separating a data set in quartiles. Quartiles divide a data set in a well-organised intense on four equivalent parts. The values, which divide all part, are called the first, second, and third quartiles; and they are denote by Q1, Q2, and Q3, respectively.[6]

- Q1 refer as "middle" value in the *first* half of the data set.
- Q2 refer as median/ middle value in the set.
- Q3 refer as "middle" value in the *second* half of the data set.

The Interquartile range is Q3 minus Q1.

2.3 WEKA

The full form of WEKA: Waikato Environment for Knowledge Learning. Weka is a computer program that was developed by the student of the University of Waikato in New Zealand for the purpose of identifying information

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

from raw data gathered from agricultural domains [7]. Data pre-processing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. It is an open source application, which is freely available.

Figure 1. Weka

In Weka datasets should be formatted to the ARFF (Attribute-Relation File Format) format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file. Classify tab in Weka Explorer is used for the classification purpose. A large different number of classifiers are used in weka such as bayes, function, tree etc.[7]

Steps to apply classification techniques on data set and get result in Weka:

Step 1: Take the input dataset.

Step 2: Apply the classifier algorithm on the whole data set.

Step 3: Note the accuracy given by it and time required for execution.

Step 4: Repeat step 2 and 3 for different classification algorithms on different datasets.

Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset.

2.4 Proposed Algorithm

Step 1: Start
Step 2: Study ovarian cancer dataset.
Step 3: Choose best feature.
Step 4: Select the number of preferred features.
Step 5: Put the fitness function.
Step 6: Use the Genetic Algorithm
Step 6.1: Creation of the first generation
Step 6.2: Selection
While stop condition not met, do
Step 6.3: Crossover
Step 6.4: Mutation
Step 6.5: Selection
End
Step 7: Classification through Weka
Step 7.1: Load data
Step 7.2: Use Interquartile Range
Step 7.3: Remove all extreme value
Step 7.4: Apply classify algorithm
Step 7.5: Evaluate result.
Step 8: Evaluate accuracy

2.5 Flow Chart

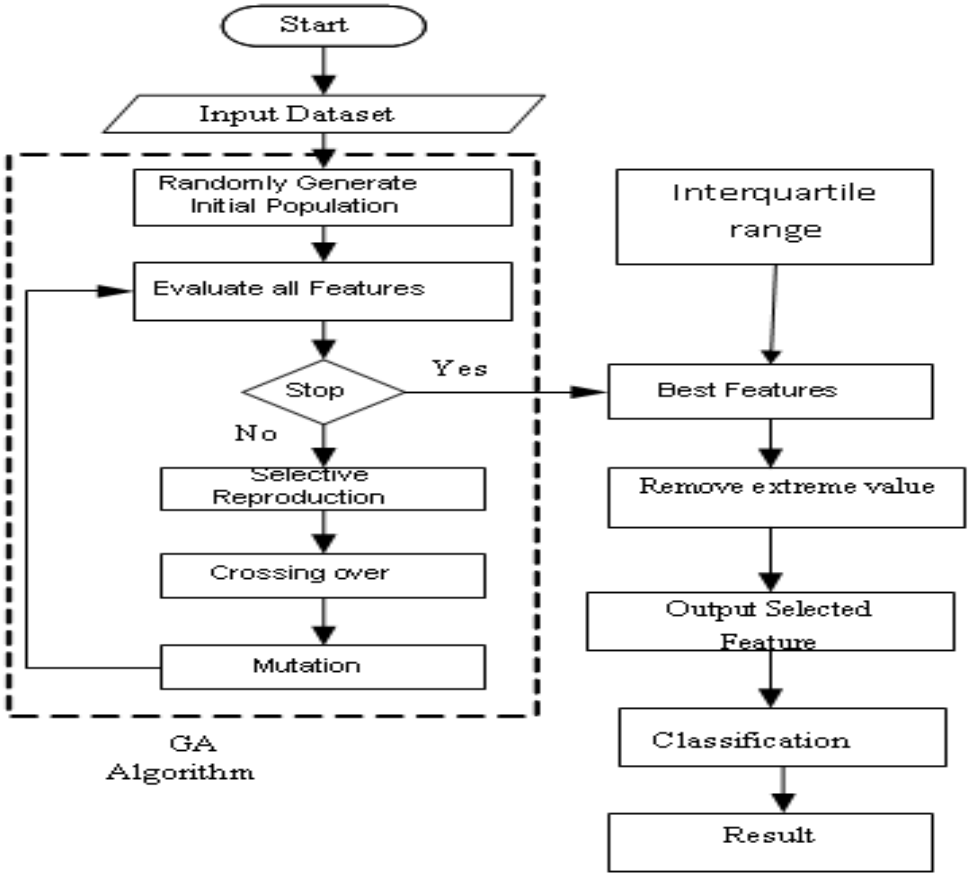

Figure 2 Flow chart of proposed work

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

III. CLASSIFICATION

Classification is an important data mining technique with broad applications. It is used to classify each item of data keen on one of predefined set of classes or groups. Classification algorithm plays an important role in document classification. In this research, we have analysed five classification algorithm named as Naive Bayes, Ib1, IBK, K-Star, LAD Tree.

IV. EXPERIMENT RESULTS

In this research, ovarian cancer dataset is used. Firstly, in Mat Lab best features are selected through genetic algorithm and classify through weka. The results are put into a table1 and table 2. Ovarian cancer data set of data size 15154*216 is reduced to 20*216. In this reduced data, Interquartile range is used so that outlier and extreme value attribute are added to it. Moreover, the data size become to 216*22 from the previous one. Form this all the extreme values are removed from the dataset. After this various classification algorithm are applied to classify data through weka.

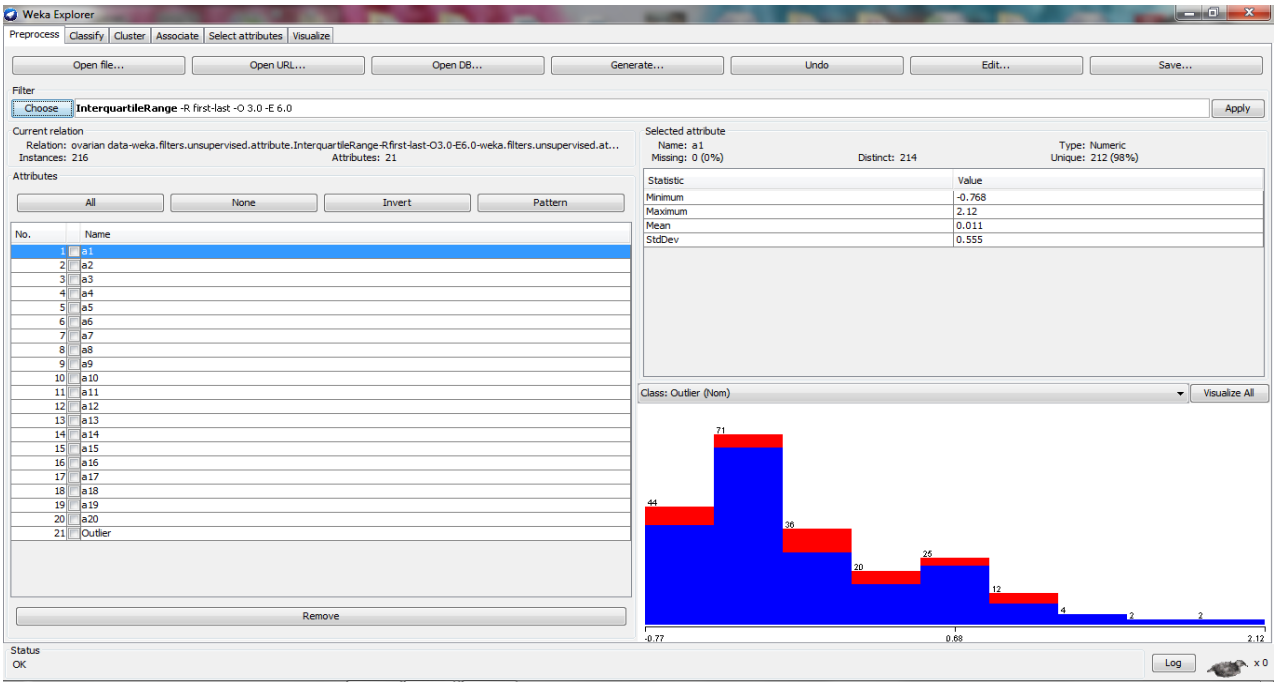

Figure. 3 Dataset in Weka

TABLE 1. DATASET WITH GENETIC ALGORITHM

Dataset	Attributes	Instances	Classes
Ovarian Cancer	15154	216	2 (Benign, Cancer)
Ovarian Cancer (with GA)	20	216	2 (Benign, Cancer)

TABLE 2. APPLY INTERQUARTILE RANGE

Dataset	Attributes	Instances	Classes
Ovarian Cancer (GA)	22	216	2 (Benign, Cancer)
Remove extreme values	21	216	2 (Benign, Cancer)

TABLE-3: CLASSIFICATION RESULTS OF VARIOUS ALGORITHMS

Classifier	Correctly classified instance	TP rate	FT rate	Precion	Recall	F. Measure	Roc area	Time (sec)
Naives Bayes	110 (50.92%)	0.492	0.394	0.874	0.492	0.629	0.635	0.3
IB1	170 (78.70%)	0.847	0.545	0.896	0.847	0.871	0.651	0
IBK	170 (78.70%)	0.847	0.545	0.896	0.847	0.871	0.655	0
K-Star	178 (82.40%)	0.902	0.606	0.892	0.902	0.897	0.626	0
LAD Tree	170 (78.70%)	0.896	0.818	0.859	0.896	0.877	0.574	2

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

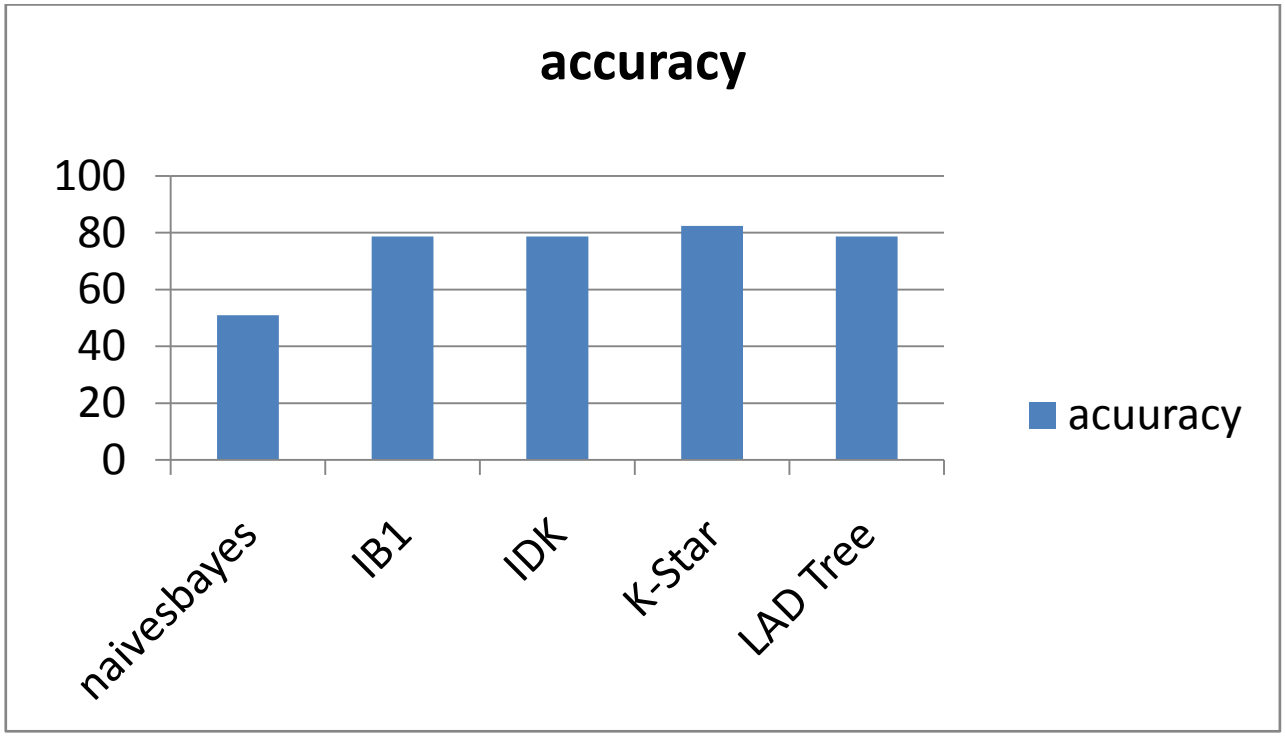

Figure 4. Accuracy measure chart

V. CONCLUSION AND FUTURE WORK

In this work various classification algorithm are used to classify the ovarian cancer data set. Weka tool is used for classification. In this K-Star is the best classification algorithm having 82.40% accuracy and its take less time for classifying the data set. In future we can apply various classification techniques for classifying different medical data in less time.

REFERENCES

1. F. J. Ferri, V. Kadirkamanathan and J. Kittler, "Feature subset search using genetic algorithms", Proceedings of the IEEE Workshop on Natural Algorithms in Signal Processing, vol. 740, 1993.
2. J. Yang and V. Hanover, "Feature subset selection using genetic algorithm", Journal of IEEE Intelligent Systems, vol. 13, pp. 44-49, 1998.
3. K. Jain and B. Chandrasekaran, Dimensionality and Sample Size Considerations in Pattern Recognition Practice, Amsterdam: Handbook of Statistics, vol. 2, 1987.
4. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artijicialintelligence, 97:245- 271" 1997.
5. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.454&rep=rep1&type=pdf
6. http://stattrek.com/statistics/dictionary.aspx?definition=Interquartile%20range.
7. http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

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