Spvryan's International Journal of Engineering Sciences & Technology (SEST)
ISSN : 2394-0905

An Effective Selection approach for Classification Algorithms with proper attribute values to build Classifier model

Ms. Varsha C. Belokar
M.Tech Scholar
Dept. of Information Technology
Institute of Engineering & Technology
Alwar, Rajasthan, India
varsha30belokar@gmail.com

Prof. Rohit Singhal
I/C HOD
CSE and IT Department
Institute of Engineering & Technology
Alwar, Rajasthan, India
mtechrohit@gmail.com

ABSTRACT: **We are dealing with huge amount of data. It may be from any field e.g. Science, Banking, Educational etc. These databases contain hidden information that can be used for intelligent decision making. Data mining can be used in industries and research to extract knowledge from huge amount of data. In banking sector there is need to categorize bank loan application as safe or risky. So bank loan manager needs analysis of his or her data. To solve data mining problem several tools are available. WEKA is one of data mining tool. For solving classification problems, WEKA provides classification algorithms such as decision tree, neural network, lazy classifiers etc. For each algorithm, tool allows user to select specific set of values for different parameters. Classifier performance can be improved by making series of experiments to get best accuracy. Thus for novice user it is difficult to guess proper values for parameters and only option is to try series of experiments which is time consuming. This work aims at developing a database which contains number and type of attributes, presence or absence of missing values, different values for building classifier model and accuracy of classifier. This database then can be made available for novice user to build classifier model based on past experience and then user can classify bank loan data as safe or risky.**

Keywords: Data Mining; classification; classifier; WEKA

I. INTRODUCTION

There is huge amount of information that is hidden in raw data. Data mining is finding previously unknown and potentially useful information from data.[2] Several tools are available for solving data mining problems, both in open source and commercial category.

WEKA is open source data mining tool and widely used in many organizations. [3] WEKA provides variety of classification algorithms such as decision tree, neural network, lazy classifiers etc. For each algorithm, tool allows user to select specific values for different parameters for e.g. in case of neural network user needs to provide values for momentum, learning rate, epochs etc. Although default values are provided for these parameters but to enhance the performance of classifier (accuracy) user needs to perform series of experiments with different values for parameters. For novice users it is difficult to guess proper values for parameters and option is only to try series of experiments so this process is time consuming.

This is our main focus to make the process comfortable with novice user. We are having large amount of data in banking sector for different loan purposes which can be classified in Safe or risky category. A bank loan officer needs analysis of their data to categorize which loan applicants are 'safe' and which are 'risky' for bank. In this, data analysis task is classification, where a model or classifier is constructed to predict categories such as 'safe' or 'risky' for bank loan data.

Classification is the process of finding a model (or function) that describes and distinguishes data classes, for the purpose of being able to use the model to predict class of objects whose class label is unknown. New model is based on analysis of set of training data (i.e. data objects whose class label is known). [4]

This work aims at selecting parameter values for different parameters of classification algorithms to build classifier model with best accuracy. This will be helpful for bank loan officer to further use these parameter values in analysing and classifying loan applicant's data as safe or risky.

II. WEKA DATA MINING TOOL

WEKA (Waikato Environment for Knowledge Analysis) is computer program developed at university of Waikato, New Zealand [3] for the purpose of identifying information from raw data gathered from various domains. WEKA provides implementation of learning algorithms such as decision tree, multilayer perceptron, LWL classifiers, J48 classifiers etc. In WEKA tool each data object is described by fixed number of attributes they are of specific type, normal alpha-numeric or numeric values. [3] After loading database in WEKA, user can apply classification method to that database and starts analysing it by inserting different parameter values required for each classification algorithm. After that it shows analysis result for different parameter values with correctly classified instances and incorrectly classified instances as shown in figure 1.

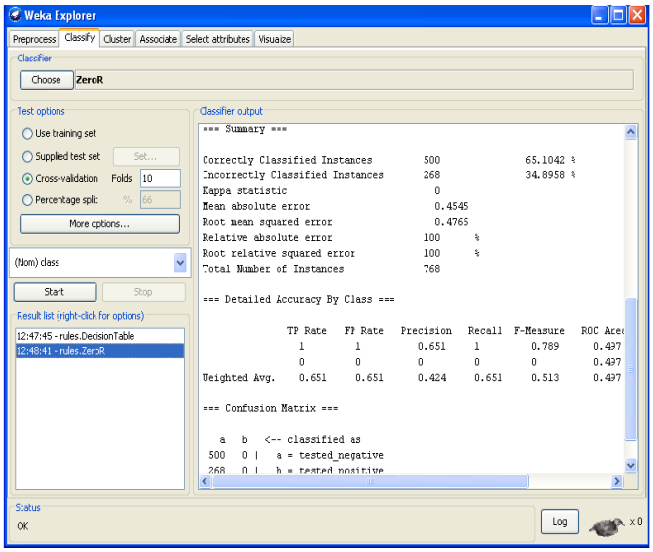

Fig 1: Analysis of dataset in WEKA [5]

But this process is time consuming as user needs to do analysis process for different parameter values for all algorithms till it shows classifier name with sufficient accuracy. So for novice user it is difficult to build model.

III. WORKING METHODOLOGY

The purpose of this work is to make effective selection of classification algorithm with proper parameter values using WEKA which can be used in analysis of loan databases as safe or risky. To achieve this we are developing two database tables as dataset-information and classifier-result. Dataset-information stores details about dataset. Classifier-result stores classification analysis result. Table shows structure of databases.

TABLE 1: DATASET-INFORMATION

Column name	Constraints	Data type
Id	Primary Key	Number
Dataset Name	Not Null	Varchar
Number of attribute	Not Null	Number
Instances	Not Null	Number
Classes	Not Null	Number

TABLE 2: CLASSIFIER-RESULT

Column name	Constraints	Data type
Id	Foreign Key	Number
Classifier	Not Null	Varchar
Parameter 1	Not Null	Varchar
Parameter 2	Not Null	Varchar
…	…	….
Parameter N	Not Null	Varchar
Accuracy	Not Null	Number

Algorithm:

Input- Dataset file in .arff (Attribute Relation File Format) format
Output- Name of Classifier and accuracy
Method: It is divided in two parts as A and B.

A) For finding parameter values to get best accuracy.
1. Read database file (Training Set) on which we want to find parameter values to get best accuracy.
2. Repeat "*p***q*" times i.e. *p* is number of times same algorithm is executed with different parameter values and *q* is number of algorithms.
3. Call algorithm *1* to *q*.
4. Pass values to *parameter 1=x*
   Parameter 2=y
   Parameter 3=z
   . .
   . .
   Parameter N=n
5. Generate Output
6. Read generated output file. Read accuracy value from file.
7. Store *algorithm name, parameter 1....., parameter N* values and *accuracy* value to the database.
8. Jump to step 3 till "*p***q*" time execution completes.

B) Suggesting parameter values and classifier algorithm to user.
1. Read database file (Training Set) on which we want to find parameter values and classifier algorithm with best accuracy.
2. Search *database name* in the result file.
3. If database name is not found, display message "No record found" and again go to step 2 of part A else display parameter values, accuracy value and classifier name from the result file.
4. End.

IV. MODELING AND DESIGN

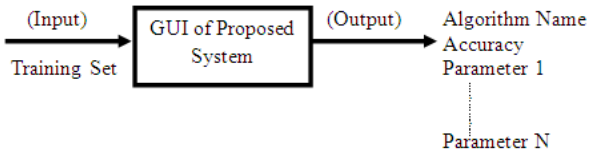

Fig 2: Working of proposed System

As shown in figure 2 proposed system accepts training set (Known records) on which repeated experiments will be performed to predict algorithm and its corresponding parameter values that classifies the given training set with best accuracy result. Training set contains previous loan applications data which can be used for proper selection of classification algorithm.

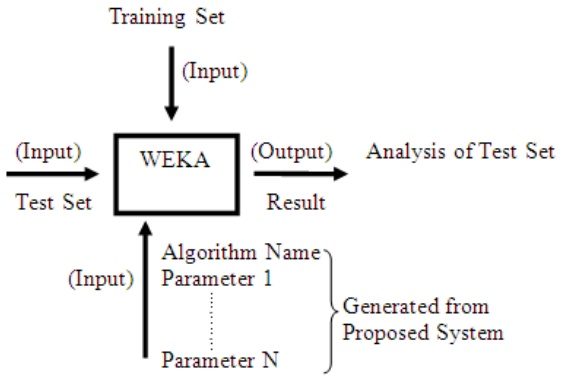

Fig 3: Use of Output Generated from Proposed System

Proposed System predicts suitable algorithm and its corresponding values. This output is useful for the user who works for data mining and analysis of data. As shown in figure 3 loan officer supplies training set to the data mining tool WEKA which contains known loan database that is helpful to approve or reject new loan applicants. Desired classification test set (Unknown records) is also given as input which contains new loan applicants data.

A data mining user may use any algorithm and its default parameter values. This may results in incorrect classification of test set. Output generated by proposed system can be used here for achieving better classification results. These results provide a secure way to deal with new loan applications for their approval or rejection.

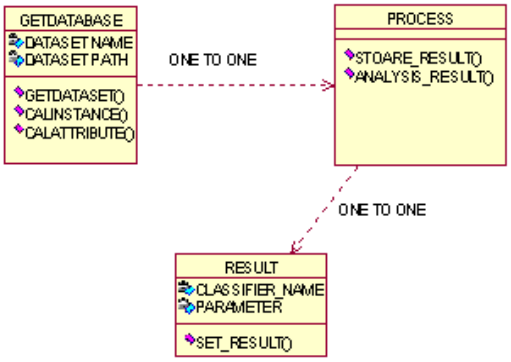

Fig 4: UML Class Diagram

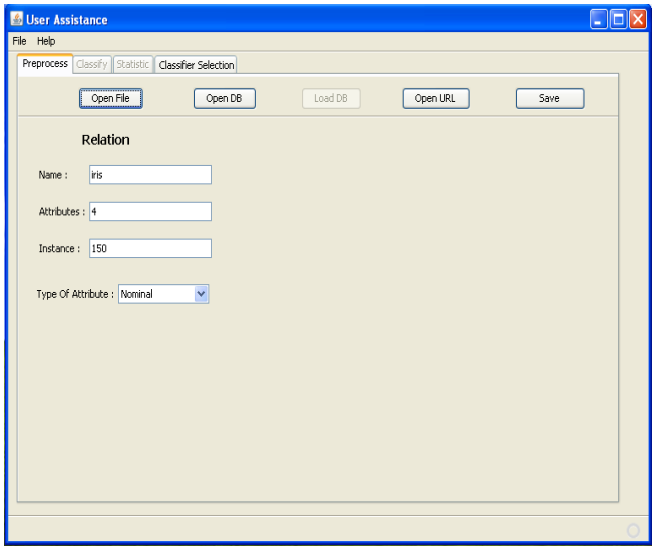

Fig 5: GUI of User Assistance

Figure 5 shows window containing pre-process tab in which bank loan database (.arff) file will be loaded for further processing. After loading it shows Name, Number of attributes, Number of records, Missing value status and type of attribute etc. After saving this information next classify tab will be enabled and according to algorithm user will perform analysis of database.

Figure 6 shows window containing results of comparison of various values for LWL Algorithm. All comparative results are stored in debase file and best known value is only represented to user. It minimizes comparison efforts and provides required information to its user.

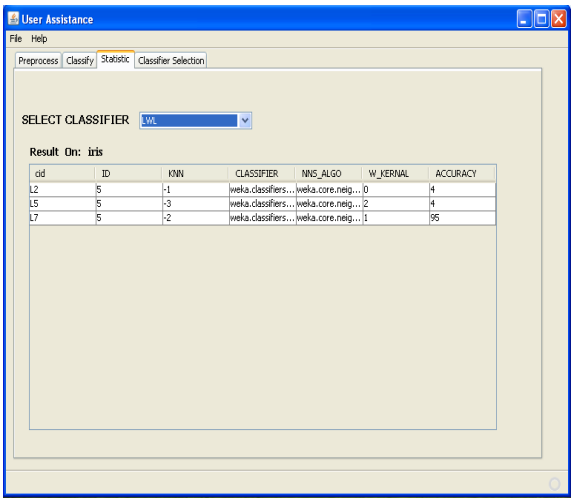

Fig 6: Results of comparison of various values in User Assistance

V. FUTURE SCOPE AND RESULTS

The main purpose of the work is to use in banking sector which deals with different loan areas and also to guide bank loan officer in analysing loan applicants data to get knowledge whether it's safe to approve loan or risky. Results which will be produced using this system can be helpful to intellectually organize information in real life.

REFERENCES

[1] Data Mining: A Knowledge Discovery Approach, K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, Springer, ISBN: 978-0-387-33333-5, 2007.

[2] Data Mining: Concepts, Models, Methods, and Algorithms, Mehmed Kantardzic, ISBN: 0471228524, Wiley-IEEE Press, 2002.

[3] WEKA user manual

[4] Data Mining by Jiawei Han, Micheline Kamber

[5] WEKA 3.6.8jre

[6] WEKA at http://www.cs.waikato.ac.nz/~ml/weka.