A Comparative Study of Classification Algorithms on Aliphatic Carboxylic Acids Data Set using WEKA

Kavitha C.R, Mahalekshmi T

Abstract*— Classification is the process of arranging a number of items into groups in such a manner that the members of the group have one or more characteristics in common. In this research paper, we present a comparative study of five different classification algorithms using WEKA, a data mining tool. This article gives an overview about the classification algorithms such as ZeroR, Naïve Bayes, J48, IBK and SMO. The dataset used for conducting the experiment is the toxicity dataset of aliphatic carboxylic acids. The main aim of this paper is to make a comparison of different classification algorithms and to find out the best algorithm out of the five chosen algorithm which gives the most accurate result.

***Index Terms*— classification, ZeroR, Naïve Bayes, J48, IBK, SMO, WEKA**

I. INTRODUCTION

Data mining (DM) is the process of analyzing data from different angles and summarizing it into useful information that can be used for making intelligent business decision. It is not specific to any industry, applied in almost all areas to explore the possibility of hidden knowledge. Now a day, Data mining techniques are also used in the field of Cheminformatics. It is the use of computers and information techniques applied to solve the problems in chemistry. DM involves the analysis of data stored in different chemical databases [1]. A chemical database is a database specifically designed to store chemical information like chemical and crystal structures, spectra, reactions and synthesis, and thermo physical data. The data which is stored in such chemical databases are used for conducting several researches. This work uses the toxicity dataset of aliphatic carboxylic acids [2]. Aliphatic carboxylic acids are a wide range of chemicals that perform a diverse range of industrial functions. Many occur naturally and serve an important function in nutrition, and others are intermediates in normal biochemical processes. Aliphatic carboxylic acids are formed from primary alcohols or aldehydes by reflux with potassium dichromate (VI) acidified with sulphuric acid. Data mining techniques includes classification, clustering and regression. This paper discusses about the classification techniques in detail. Classification is an important data mining method for the analysis of toxicity data that can be used for extracting models describing important data classes. There are many classification methods available which are used by various researchers. The main aim of the paper is to study the performance of the classification algorithms. The remaining paper is organized into 7 sections. Section II gives an overview of the five different classification algorithms used in this paper. The next section presents the different performance evaluation measures for the classifiers. The section IV deals with WEKA. Toxicity dataset has been discussed in section V which is followed by discussion in Section VI and the conclusion is given in section VII followed by the references.

Revised Version Manuscript Received on May 20, 2015.

Kavitha C.R, Research Scholar, R&D, Bharathiar University, Coimbatore, India.

Dr. Mahalekshmi T, Principal, Sree Narayana Institute of Technology, Kollam, India.

II. CLASSIFICATION ALGORITHMS- AN OVERVIEW

A classification technique is an important component of machine learning algorithms in order to extract rules and patterns from data that could be used for prediction. Classification is a method of mapping data records into one of several predefined classes. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [3].

A classifier is built by following two steps namely training and testing. In training phase, a classification model is built. The individual objects or examples are referred collectively as training dataset. Before building the model, this training set should be classified i.e., to attach a class label to each object or example. In testing phase, the model built in the previous step is used for classification. First, the predictive accuracy of the classifier is estimated. A test set which is made up of test tuples and their associated class labels is used to measure it. These tuples are randomly selected from the general data set and are not involved while building the classification model earlier.

Different techniques from machine learning, statistics, information retrieval and data mining are used for classification. They include Bayesian Methods, Bayesian Belief networks, Decision Trees, Neural Networks, Associative Classifiers, Emerging Patterns, and Support Vector Machines (SVM). This study aims to compare the performance of five classification algorithms such as ZeroR from rules sub menu, Naïve Bayes from Bayes sub menu, J48 from Trees sub menu, IBK from lazy sub menu and SMO from function sub menu in WEKA [4]. A brief explanation of each of the techniques applied in this paper is presented below.

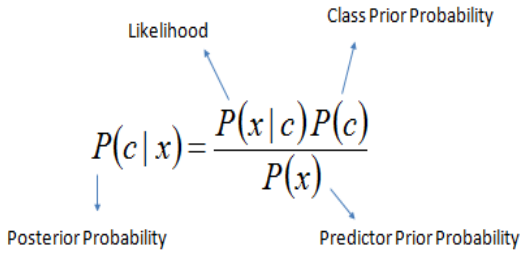
A. Zero R

ZeroR [5] is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier 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. A frequency table is constructed for the target and the most frequent value is selected.

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

B. *Naïve Bayes*

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build and can be used for very large datasets. Naive Bayesian classifier often performs well than more sophisticated classification methods. The posterior probability, $P(c|x)$ is calculated from $P(c)$, $P(x)$, and $P(x|c)$. The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig1: Formula for calculating posterior probability [6]

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ 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*.

One way of classification is by determining the posterior probability for each class and assigning c to the class with the highest probability.

C. *J48*

J48 [7] is a decision tree classifier. Decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes gives the possible values that these attributes can have in the observed samples, while the terminal nodes gives the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

In order to classify a new item, a decision tree is created based on the attribute values of the available training data. Whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that gives the most about the data instances can be classified as the best is said to have the highest information gain. Among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then that branch is terminated and the target value that we have obtained is assigned to it. This is continued in this manner until we either get a clear decision of what combination of attributes gives us a particular target value, or we run out of attributes. In the event of running out of attributes, or if an unambiguous result is obtained from the available information, then this branch is assigned a target value that the majority of the items under this branch possess.

D. *Instance based Learning Algorithm (IBK)[8]*

IBK is a K-NN (K- Nearest Neighbour) classifier, a supervised learning algorithm, where a given data set is partitioned into a user specified number of clusters, K. Predict the same class as the nearest instance in the training set. Training phase of the classifier stores the features and the class label of the training sets. New objects are classified based on the voting criteria. It provides the maximum likelihood estimation of the class. Euclidean distance metrics is used for assigning objects to the most frequently labeled class. Distances are calculated from all training objects to test object using appropriate K value.

It builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class.

E. *Sequential Minimal Optimization (SMO) [9]*

SMO is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines. SMO is an iterative algorithm for solving the optimization problem. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced to:

$$0 \le \alpha_1, \alpha_2 \le C,$$
$$y_1\alpha_1 + y_2\alpha_2 = k,$$

and this reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. k is the negative of the sum over the rest of terms in the equality constraint, which is fixed in each iteration. First, a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem is found out. Then, a second multiplier α_2 is picked and the pair (α_1, α_2) is optimized. Steps 1 and 2 are repeated until convergence. When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved.

III. COMPARISON OF CLASSIFIERS: PERFORMANCE MEASURES

The performance of classifiers can be evaluated by considering certain criteria such as Accuracy, Speed,

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

Robustness, Scalability and Interpretability where Accuracy is the ability of the model to correctly predict the class label, Speed is the computation costs involved in generating and using the model Robustness is the ability of the model to make correct predictions, Scalability is the ability to construct the model efficiently and Interpretability is the ability of the model to provide the insight. [10] The Confusion Matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. It contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of this study: [11]

1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative, and
4. d is the number of correct predictions that an instance is positive.

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Fig 2: Confusion Matrix [12]

Using the values from the confusion matrix, the performance of classifiers are evaluated by using parameter such as TP(True Positive) rate, FP (False Positive) rate, TN (True Negative) rate, FN (False Negative) rate , P (Precision) and Accuracy (AC) where TP is the proportion of positive cases that were correctly identified, as calculated using the equation: TP=d/(c + d) , *FP* is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation: FP=b/(a + b), TN is the proportion of negatives cases that were classified correctly, as calculated using the equation: TN= a/(a + b), FN is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation: FN=c/(c + d), P is the proportion of the predicted positive cases that were correct, as calculated using the equation: P=d/(b + d) and *AC* is the proportion of the total number of predictions that were correct. It is determined using the equation: AC= (a +d) / (a + b + c +d).

IV. WEKA

Waikato Environment for Knowledge Analysis (WEKA) [13] is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API, so WEKA can be embedded like any other library, in other applications. WEKA is freely available under the GNU General Public License. It is portable since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. It contains a comprehensive collection of data preprocessing and modeling techniques and it is easy to use due to its graphical user interfaces. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of WEKA's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

V. DATA SET

In this research, the data set used is the toxicity data of aliphatic carboxylic acids which was downloaded from "http://vincentarelbundock.github.io/Rdatasets/datasets.html " [10]. The characteristics of the data set are summarized in the Table 1. The aim of the data set was to predict the toxicity of carboxylic acids on the basis of several molecular descriptors like toxicity, logKow, pKa, ELUMO, Ecarb, Emet, RM, IR, Ts and P. The problem is to predict whether the given aliphatic acid is toxic or not. This is a two-class problem with class value positive and negative. The data set contains 38 observations and 11 variables with no missing values reported. There are eleven variables, including the class variable, in this data set; all other attributes are numeric-valued. The attributes are given below:

1. Toxicity - defined as *log(IGC50^(-1))*; typically the "response".
2. logKow - the partition coefficient
3. pKa- the dissociation constant
4. ELUMO- **E**nergy of the **l**owest **u**noccupied **m**olecular **o**rbital
5. Ecarb - Electrotopological state of the **carb**oxylic group
6. Emet- Electrotopological state of the **met**hyl group
7. RM - Molar refractivity
8. IR - Refraction index
9. Ts - Surface tension
10. P – Polarizability
11. C – Class variable (positive or Negative)

Table 1. Characteristics of Toxicity data sets+

Data Set	Toxicity
No of Example	38
Input Attributes	10
Output Classes	2
Total No. of Attributes	11
Missing Attributes status	No
Noisy Attributes status	No

VI. DISCUSSION

In this experiment, the classification is done by WEKA, a data mining tool. WEKA accepts data in .CSV or .ARFF file format. Sometimes the data set which we acquire from different sources may not be in the right file format. We

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

cannot apply such data directly to the data mining tool such as WEKA. As a result, file format conversion has to be performed. [10] If the data set is not in the proper format, then we need to convert the file to .CSV or .ARFF file format. In this research, the different classification algorithms taken for study includes ZeroR from rules sub menu, Bayes Naïve from Bayes sub menu, J48 from trees sub menu, IBK from lazy submenu and SMO from function sub menu. In this study, we examine the performance of the above said classification algorithms. The algorithm which has the lowest mean absolute error and higher accuracy is chosen as the best algorithm. To determine the performance on the selected classifiers or algorithms namely ZeroR, Naïve Bayes, J48, IBK and SMO, the simulation results are partitioned into several sub items for easier analysis and evaluation. Firstly, correctly and incorrectly classified instances will be partitioned in numeric as well as in percentage value and subsequently Kappa statistic, mean absolute error and root mean squared error will be found in numeric value only. The relative absolute error and root relative squared error are shown in percentage for references and evaluation. The results of the simulation are shown in Tables 2 and 3 below. Table 1 mainly summarizes the result based on accuracy and time taken for each simulation. Meanwhile, Table 4 shows the result based on error during the simulation. Figures 3 and 4 are the graphical representations of the simulation result. The Confusion Matrix for all Classifiers is given in the Table 5. Figure 5 is the graph that shows the performance of best algorithm. From Table 3 and Figure 5,it is clear that J48 algorithm is the best, Naïve Bayes is the second best than the other algorithms.

Table 2. Experiment Result of each classifier

Algorithm	Correctly classified instances % (value)	Incorrectly Classified instances % (value)	Time taken (seconds)	Kappa Statistic
Zero R	68.4211	31.5789	0	0
Bayes Naïve	89.4737	10.5263	0	0.7564
J48	97.3684	2.6316	0	0.9404
IBK	76.3158	23.6842	0	0.4639
SMO	81.5789	18.4211	0.05	0.543

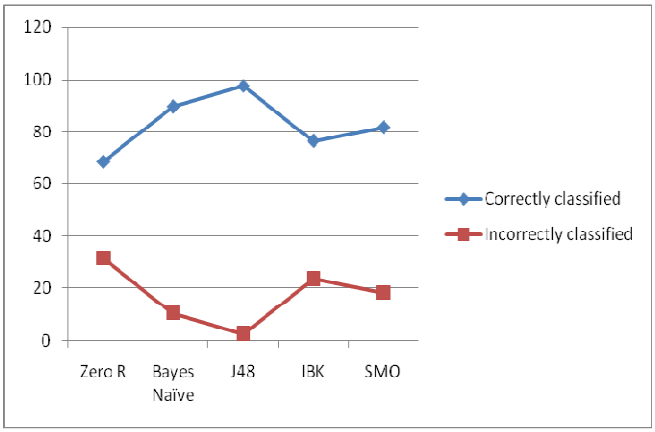


Fig 3. Classification results of toxicity using WEKA

Table 3. Error Comparison

Algorithm	Mean Absolute Error	Root mean squared Error	Relative absolute error (%)	Root relative squared error (%)
Zero R	0.4381	0.4674	100	100
Naïve Bayes	0.1355	0.3398	30.9239	72.6904
J48	0.0263	0.1622	6.0065	34.7043
IBK	0.2514	0.4738	57.3864	101.3704
SMO	0.1842	0.4292	42.0455	91.819

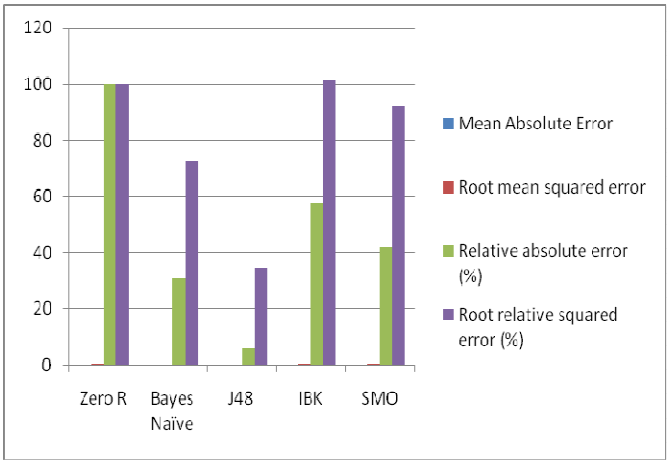


Figure 4. Error Comparison between parameters

Table 5. Comparison of Weighted Avg.

Algorithm	TP-rate	FP-rate	Precision	Recall	ROC Area
Zero R	0.684	0.684	0.468	0.684	0.556
Naïve Bayes	0.895	0.138	0.895	0.895	0.895
J48	0.974	0.012	0.976	0.974	0.974
IBK	0.763	0.289	0.769	0.763	0.766
SMO	0.816	0.309	0.812	0.816	0.808

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

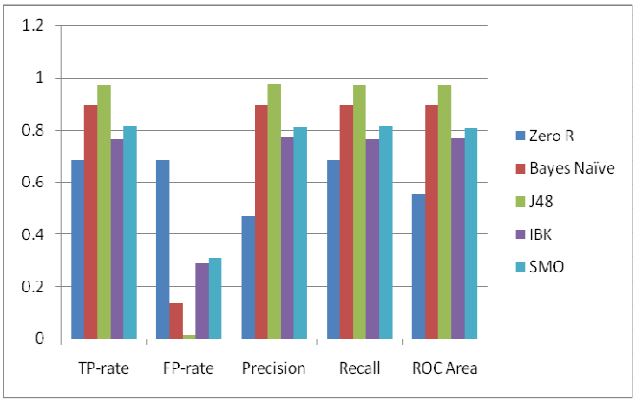


Fig 5. Graph showing the Performance of best Algorithm: J48

Table 5. Confusion Matrix for All Classifiers

Classifier	A	B
ZeroR	26	0
	12	0
Naïve Bayes	24	2
	2	10
J48	25	1
	0	12
IBK	21	5
	4	8
SMO	24	2
	5	7

In this experiment it was found that each classifier shows different accuracy rate. J48 has the highest classification accuracy and the lowest mean absolute error.

VII. CONCLUSION

The main aim of this study is to evaluate and investigate five selected classification algorithms using WEKA. The toxicity data set is used to test the performance of the selected classification algorithms. The algorithm which has the lowest mean absolute error and higher accuracy is chosen as the best algorithm. In this classification experiment, each algorithm shows different accuracy rate for different instances in the data set. By considering different parameters of accuracy and the error rate, it is found out that J48 classification algorithm is the best algorithm with a maximum accuracy of 97.3684.

REFERENCES

[1] Kavitha C.R, Dr. Mahalakshmi, "Chemical Databases: A Brief Walk", International Journal of Emerging Technology and Advanced Engineering (IJETAE), ISSN 2250-2459,ISO 9001:2008Certified Journal Volume 3 Issue 8 August 2013.
[2] 'Aliphatic Carboxylic acids data set', Available at http://vincentarelb undock .github.io/Rdatasets/datasets.html
[3] Jiawei Han, Micheline Kamber, Jian Pei",Data Mining: Concepts and Techniques", third edition, Morgan Kauffman Publishers.
[4] 'Classification', [Online] Available "http//infochemie, u-strasbg.fr /master/ tutochemo/tp8/classification.pdf".
[5] 'ZeroR', [Online] Available at **http://www.saedsayad.com/zeror.htm**.
[6] Posterior probability', [Online] Available at" http://www.saedsayad . com /naive_bayesian.htm.
[7] Foruzan Kiamarzpour, Rouhollah Dianat, Mohammad bahrani, Mehdi Sadeghzadeh, "Improving the methods of email classification based on words ontology",[Online] Available http://arxiv.org /ftp/arxiv /papers /1310/1310.5963.pdf.
[8] Grigoris Antonion, George Potamias, Costas Spyropoulos, Dimitris Plexousakis,"Advances in Aritificial Intelligence", 4th Helenic Conference on AI, SETN 2006, Heraklion crete, Greece, May 2006 Proceedings, Springer.
[9] Olalekan S. Akinola, Adetoun C. Afolabi, "Evaluating classification effectiveness of sequential minimal optimization (smo) algorithm on chemical parametization of granitoids", IJRRAS 13 (2) , November 2012, [Online] Available at http://www. arpapress.com/Volumes/ Vol13Issue2 /IJRRAS_13_2_30.pdf
[10] Performance measures',[Online] Available http://cs.uiuc.edu/class /fa05 /cs412 /chaps/6.pdf,
[11] 'Confusion Matrix', [Online] Available "http://www2.cs.uregina.ca/ ~dbd /cs831 /notes/confusion_matrix/confusion_matrix.html".
[12] 'Confusion matrix',[Online] Available at http://www2.cs.uregina.ca /~dbd/ cs831/notes/ confusion_matrix/confusion _matrix.html".
[13] 'WEKA', Available www.**weka**.net.nz/.
[14] Kavitha C.R, Dr. Mahalekshmi," Chemical File Format Conversion Tools: A n Overview", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 2, February – 2014.

Ms. Kavitha C.R is pursuing PhD from Bharathiar University, Coimbatore, Tamil Nadu. Shee has obtained her Master of Computer Application degree from Bharathidasan University, in April 2001. Her specialization area in research is Data mining in Cheminformatics. She is presently working as an Assistant Professor in the Department of Computer Applications, Sree Narayana Guru Institute of Science and Technology, N Paravur, Kerala from October 2004 to till date. Her total teaching experience is about 12 years. She had attended several national and International seminars and conferences for paper presentation. She also published several articles in the various journals.

Dr. Mahalekshmi T has obtained her Ph.D. degree in Computer Science from University of Kerala, Kerala in March, 2007 and Master of Science in Computer Science from School of Computer Science, University of Minnesota, U. S. A. in August 1985. Her specialization area in research are computational biology and soft computing algorithms.She is working as a Principal and Professor , Department of Computer Applications, Sree Narayana Institute of Technology ,Kollam, Kerala since June 2007. Her writings have appeared in numerous national and international Journals. She was an author for few books such as Data Structures in C " published by PHI Learning Private Ltd, Delhi, 2009 and a text Book "Computer Graphics" based on the 8th Module of 'B' Level – a Master of Computer Application level course, Published in 1999.

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