IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Classification Techniques in WEKA: A Review

K.H. Wandra[1*], L.P. Gagnani[2]

[1*]Director, Academic Administration, Babaria Institute of Technology, Vadodara, INDIA
[2]Dept. of Computer Engineering, C U Shah University, Wadhwan City, INDIA

[*]*Corresponding Author: khwandra@rediffmail.com*

Available online at: www.ijcseonline.org

Received: 04/Jul/2017, Revised: 17/Jul/2017, Accepted: 19/Aug/2017, Published: 30/Aug/2017

***Abstract*—** Due to the Internet Revolution there has been a data explosion in recent decades. This is due to the easy availability of Internet at any place and time. Therefore it has become very important to extract relevant information from these explosion of data. Data Mining is extraction or mining of useful information from large amount of data. This can be done manually, semi-automatic or automatically. With an enormous of data stored in databases and data warehouse there is need for development of powerful tools to get meaningful data. Data Mining has many tasks such as Classification, Clustering, etc but Classification has gained much importance. Classification is to classify the data into groups based on its characteristics. WEKA is widely used data mining tool. Here a comparison of various algorithms available in WEKA for classification tasks is done. The dataset considered is iris and various parameters considered for evaluation include accuracy, kappa statistics, mean absolute error and root mean square error. 10 mostly used algorithms are compared. Accuracy is given in terms of CCI (Correctly Classified Instances) and ICI (Incorrectly Classified Instances).

Keywords*—Classification, Weka, Data Mining*

I. INTRODUCTION

Data Mining is a branch of computer science which is used to extract meaningful information and knowledge and this extraction can be from any data such as noisy, incomplete data, missing data, etc [1].

Traditional data analysis is different from data mining in that discovery of useful information is not having any clear assumption[2]. Data Mining is commonly referred as "Knowledge Mining" or "Knowledge Discovery in Databases (KDD)". Also data mining is an essential step in KDD process.

Knowledge discovery as a process consists of an iterative sequence of the following steps as shown in Figure 1[3][4]:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge representation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

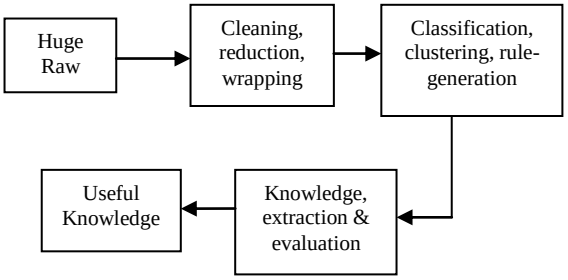

Figure 1. Block diagram for knowledge discovery

Classification is supervised technique. The input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label. The objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data. This model is used to classify test data for which the class descriptions are not known. Classification is performed on IRIS dataset using WEKA with different algorithms

© 2017, IJCSE All Rights Reserved

available in WEKA. The results are compared and depicted[5].

Rest of the paper is organized as follows, Section I contains the introduction of data mining, classification and WEKA toolkit. , Section II contain the related work of various algorithms available in WEKA and other commonly used algorithms, Section III contain the some measures considered for evaluation of classification methods, Section IV describes results and discussion and Section V concludes research work with future directions.

II. RELATED WORK

Brief overview of various commonly used algorithms in classification technique is described in Table 1 as follows. The various algorithms compared are C4.4, ID3, KNA (K-Nearest Algorithm), ANN (Artificial Neural Network), SVM (Support Vector Machine) and Naives Bayes[6][7][8].

Table 1. Review of Classification Algorithms

NO	ALGORITHM	FEATURES	LIMITATIONS
1	C4.5 Algorithm	●Build Models can be easily interpreted ●Easy to implement ●Can use both discrete & continuous values ●Deals with noise	●Small variation in data can lead to different decision trees ●Does not work very well on a small training data set ●Over fitting
2	ID3 Algorithm	●It produces the more accuracy result than C4.5 algorithm ●Detection rate is increase and space consumption is reduced	●Requires large searching time ●Sometimes it may generate very long rules which are very hard to prune ●Requires large amount of memory to store tree
3	K-Nearest Neighbor Algorithm	●Classes need not be linearly separable ●Zero cost of the learning process ●Sometimes it is robust with regard to noisy training data ●Well suited for multimodal classes	●Time to find the nearest neighbours in a large training set can be excessive. ●It is sensitive to noisy or irrelevant attributes ●Performance of algorithm depends on the number of dimensions used
4	Naive Bayes Algorithm	●Simple to implement ●Great computational efficiency & classification rate ●It predicts accurate results for most of the classification and prediction problems	●The precision of algorithm decreases if the amount of data is less ●For obtaining good results it requires a very number of records
5	Support Vector Machine Algorithm	●High Accuracy ●Work well even if data is not linearly separable in the base feature space	●Speed & size requirement both in training and testing is more ●High complexity & extensive memory requirements for classification in many cases ●Learning can be slow
6	Artificial Neural Network Algorithm	●It is easy to use, with few parameters to adjust ●A neural network learns & reprogramming is not needed ●Easy to implement ●Applicable to a wide range of problems in real life	●Requires high processing time if neural network is large ●Difficult to know how many neurons and layers are necessary ●Learning can be slow

III. EVALUATION MEASURES

The comparison in done in WEKA (Waikato Environment for Knowledge Analysis). WEKA is tool of machine learning entirely written in JAVA programming language. It is one of the most widely used data mining software. It is a collection of various algorithms for data mining tasks as classification, clustering, regression, association rule mining, etc. Also it has tools for preprocessing and visualization. WEKA is open source standard issued under the GNU General Public License. It can also be applied to Big Data that has gained much importance these days.

Classification accuracy is based on the confusion matrix or contingency table[9]. Various parameters[10][11][12] in it include:

TP=True Positives (predicted correctly)

FP=False Positives (Predicted as Correct are Incorrect)

TN=True Negatives (Predicted as Incorrect are actually Inorrect)

© 2017, IJCSE All Rights Reserved

FN=False Negatives

The confusion matrix is given as:

	a	b
Actual a=0	TP	FN
Actual b=1	FP	TN

The Kappa Statistic measures the agreement of prediction with the true class. The value 1.0 indicates completed agreement.

The Mean Absolute Error (MAE) measures the average magnitude of errors in a set of forecasts and Root Mean Squared Error (RMSE) is a quadratic scoring rule which measures the average magnitude of error.

IV. SIMULATION RESULTS

The simulation results were carried on Iris.arff dataset with Weka 3.7. The results are given in Table 2.

Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class

Table 2. Simulation Results for Classification on Iris Dataset

Algo	CCI	ICI	KS	MAE	RMSE
FURIA	142 94.6667%	8 5.3333%	0.92	0.0307	0.1636
MLP	146 97.3333%	4 2.6667%	0.96	0.0327	0.1291
RBF network	143 95.3333%	7 4.6667%	0.93	0.034	0.1585
RBF classifier	143 95.3333%	7 4.6667%	0.93	0.1174	0.1771
FLR classifier	137 91.3333%	13 8.6667%	0.87	0.0578	0.2404
BayesNet	139 92.6667%	11 7.3333%	0.89	0.0454	0.1828
Naive Bayes	144 96%	6 4%	0.94	0.0342	0.155
JRip	143 95.3333%	7 4.6667%	0.93	0.0454	0.1727
ZeroR	50 33.3333%	100 66.6667%	0	0.4444	0.4714
J48	144 96%	6 4%	0.94	0.035	0.1586

FURIA-Fuzzy Unordered Rule Induction
MLP-Multi Layer Perceptron
RBF-Radial Basis Function Network
FLR-Fuzzy Lattice Reasoning

CCI- Correctly Classified Instances
ICI- Incorrectly Classified Instances
KP- Kappa statistic
MAE- Mean absolute error
RMSE- Root mean squared error

V. CONCLUSION AND FUTURE WORK

The results conclude that Multi Layer Perceptron (MLP) gave better classification accuracy than other methods like FLR (Fuzzy Lattice Reasoning), RBF (Radial Basis Function Network), FURIA (Fuzzy Unordered Rule Induction), J48, ZeroR, JRip, Bayes algorithm in terms of CCI (Correctly Classified Instances) and KP (Kappa Statistics). This can be depicted from results of Table 1. Further results also depend on dataset chosen and application.

The future scope will consider more methods of Computational Intelligence (CI)/ Swarm Intelligence (SI) as well as hybrid of existing methods. Also the work can be extended for large datasets.

ACKNOWLEDGMENT

We would like to thank the faculties of BITS, Vadodara and C U Shah University for providing a platform to undergo this research.

REFERENCES

[1] E. Hullermeir, "*Fuzzy sets in machine learning and data mining*", Elsevier, pp.1493-1505, 2008.

[2] G. Peter Zhang, "*Neural Network for Data Mining*", Springer, pp.419-444, 2010.

[3] J. Vashishtha, D. Kumar, S. Ratnoo, "*Revisiting Interestingness Measures for Knowledge Discovery in Databases* ", IEEE, pp.72-78, 2012.

[4] K. Lal, N.C. Mahanti, "*Role of soft computing as a tool in data mining*", IJCSIT, Vol.2, Issue.1, pp.526-537, 2011.

[5] L. Gagnani, H. Chhinkaniwala, "*Soft Computing as a Tool in Data Mining:A Review*", In the Proceedings of the 2015 International Conference on Emerging Trends in Scientific Research (ICETSR 2015), Wadhwan, INDIA, pp.148-155, 2015.

[6] M.F. Otham, T.M. Yau, "*Comparison of Different Classification Techniques using WEKA for Breast Cancer*", In the Proceedings of 2007 IFMBE, pp.520-523, 2007.

[7] Marie Fernandes , "*Data Mining: A Comparative Study of its Various Techniques and its Process*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, 2017.

[8] N. Bhargava, S. Dayma, A. Kuar and P. Singh, "*An approach for classification using simple CART algorithm in WEKA*", In Proceedings of 2017 ISCO, Coimbatore, INDIA, pp.212-216, 2017.

[9] P. Shabanzadeh and R. Yusof, "*An Efficient Optimization method for solving Unsupervised data classification problems*",

© 2017, IJCSE All Rights Reserved

Computational and Mathematical Methods in Medicine, Hindawi, 9 pages, 2015.

[10] R. Agrawal, T.L Mielinski, A. Swami, "*Database Mining:A Performance Perspective*", IEEE Transactions on Knowledge and Data Engineering, Vol. 12, pp.914-925, 1993.

[11] S. Radha Priya and M. Devapriya, "*Survey on Attribute Oriented Induction Using Data Mining Techniques*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.5, pp.125-129, 2016.

[12] AR. PonPeriasamy, E. Thenmozhi, "*A Brief Survey of Data Mining Techniques Applied to Agricultural Data*", International Journal of Computer Sciences and Engineering (IJCSE), Vol. 5, Issue. 4, pp.129-132, 2017.

Authors Profile

Mr. Kalpesh H Wandra pursued BE in Electronics & *Mr. Kalpesh H Wandra* pursued BE in Electronics & Communications from NGU, Patan and ME in Microprocessor System Application from MS University, Baroda. Further He pursued Ph.D in Computer Engineeering from Saurashta University, Rajkot in 2010. He is a member in IEEE, CSI and life member of ISTE. He is currently working as Director Academic Administration at BITS, Vadodara. He has published more than 60 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Wireless Communication, Networks and Advanced Microprocessors and Microcontrollers based education. He has 20 years of teaching experience.

Mr. Lokesh P Gagnani pursued Bachelor of Engineering from CCET and Master of Engineering from SSEC, Gujarat. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Information & Technology Engineering at KIRC, Kalol affiliated to GTU since 2008. He is a member of IEEE society since 2015 and a life member of the ISTE since 2013. He has published more than 15 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 9 years of teaching experience.

© 2017, IJCSE All Rights Reserved