Analysis of Information Gain Ranking Feature Selection Algorithm Using UCI Machine Learning Datasets

[1]M.Jeyanthi,[2]Dr.C.Velayutham
[1] Department of computer science, Aditanar College of Arts and Science, Tiruchendur, affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India,
[2] Department of computer science, Aditanar College of Arts and Science, Tiruchendur, affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India,
[1]jeyanthieral88@gmail.com , [2]cvsir22@gmail.com

Abstract - Data mining is a process to extract usable data from a larger set . Feature selection is one of the important pre-processing steps in data mining. Feature selection (FS) is a process to select features which are more informative. This paper analyses five feature selection algorithms such as Information Gain, Correlation Attribute , Relief-F, One-R, Symmetrical Uncertainty using five UCI machine Learning data sets. The classification performance of the reduced data is measured by WEKA classifiers JRIP and J48. Based on the classification accuracy we propose Information Gain feature selection algorithm.

Index Terms- Data Mining, Feature Selection, Classifiers J48, JRIP.

I. INTRODUCTION

Classification is the one of the main technique used for discovering pattern from known classes [1]. In real word, dataset contains hundreds of attributes. But not all the attributes are needed to complete the mining task[2]. In order to find the importance of attributes, feature selections algorithms are utilized. Instead of processing all the attributes, only relevant attributes are involved in the mining process. This will reduce processing time as well as increase the performance of mining task. Therefore attribute selection algorithms are applied before applying data mining tasks such as classification, clustering, outlier analysis and so on.

Attribute selection is a two step process. one is subset generation and another one is ranking. Subset generation is a searching process which is used to compare the candidate subset to the subset already determined [3]. If the new candidate subset returns better results in terms of certain evaluation then the new subset is termed as the best one. This process is continued until termination condition is reached.

The next one is Ranking of attributes which is used to find the importance of attributes [4]. There are many ranking methods such as which are mostly based on statistics or information theory. There are two varieties of attributes selection algorithms. i) Filter approach ii) Wrapper approach. The learning algorithms itself uses the attribute selection task then it is called wrapper approach [5]. In filter approach the attributes are evaluated on the basis of evaluation metrics with respect to the characteristics of the dataset [6].

The organisation of the paper as follows : Section II shortly describes the Literature Review , Section III describes the UCI data set, Section IV describes the feature selection methods, Section V describes the classification Algorithms , Section VI describes the Results and Discussion, Section VII Concludes the paper.

II. LITERATURE REVIEW

Sunita Beniwal et al.,(2012) [7] presents the introduction about the various classification and feature selection techniques frequently used data mining which also states the importance of filter and wrapper approaches of feature selection methods. But this study does not contribute to any experimental study.

Mital Doshi et al.,(2014) [8] determines to predict students performance. For that purpose feature selection techniques such as Chi-square, InfoGain, and GainRatio are utilized. Then classification task is carried out by the use of NBTree, Multilayer Perceptron, NaiveBayes and Instance based K- nearest neighbor classifiers. The result concludes that the accuracy of the prediction is improved because of the applied filter techniques.

M. Ramaswami et al.,(2009)[9] determines the most relevant subset of attributes based minimum cardinality. In order to find the goodness of features, the six feature selection algorithms are involved in this study. It can be measured in terms of F-measure and ROC value. The result assures that the computational time and cost is decreased with minimum number of features.

Dr.C.Velayutham et al.,(2011) [10] proposed a new rough set-based unsupervised feature selection using relative dependency measures. The method employs a backward elimination-type search to remove features from the complete set of original

features. As with the WEKA tool is used to classify the data and the classification performance is evaluated using classification accuracy and mean absolute error, the method is compared with an existing supervised method and it demonstrates that it can effectively remove redundant features.

Dr.C.Velayutham et al.,(2011) [11] proposed unsupervised feature selection method using rough set theory. The K-Means, FCM, and NN-SOM algorithms are used to cluster the data. The classification performance is evaluated using confusion matrix with positive and negative class values. This method is compared with existing supervised methods and it demonstrates that it effectively remove the redundant features.

Dr.C.Velayutham et al.,(2011) [12] proposed the unsupervised feature selection in mammogram image, using rough set based entropy measure. The K-Means, and FCM algorithms are used to cluster the data. The classification performance is evaluated using confusion matrix with positive and negative class values. The proposed method is compared with existing supervised methods and it demonstrates that it can effectively remove redundant features.

Arpita Nagpal et al., (2018)[13] develops a new algorithm for feature subset selection on cancer microarray data based on the concept of Qualitative Mutual Information (QMI). This algorithm removes irrelevant and redundant features so that the dimensionality of data gets reduced and it can produce better classification results.

Hossam Faris et al.,(2017) [14] proposed a robust approach based on a recent nature-inspired meta heuristic called multi-verse optimizer (MVO) for selecting optimal features and optimizing the parameters of SVM simultaneously. Two system architectures are implemented for the proposed approach: the first architecture is commonly used in the literature while the second is proposed in this work to increase the credibility of the SVM prediction results. The developed approach is assessed and benchmarked with four well-regarded meta heuristic algorithms (GA, PSO, BAT and Firefly) and the grid search. Experiments show that MVO was able to optimize SVM achieving the highest accuracy compared with the other optimizers based on the two investigated architectures.

K.Sutha et al.,(2015)[15] presents the benefits and drawbacks of the some feature selection algorithms in terms of efficiency. Nearly 12 feature selection algorithms are involved in this study.

III. DATASET DESCRIPTION

For experiments, data sets are taken from Data Mining Repository of University of California Irvine (UCI) [16]. These datasets are given in Table1.

Table 1: Characteristics of Datasets

No.	Datasets	Features	Instances	Classes
1	Breast cancer Wisconsin	10	699	2
2	Hypothyroid	30	3772	4
3	Dermatology	34	366	6
4	Soybean	35	683	2
5	Autoprice	15	159	2

IV. FEATURE SELECTION METHODS

Feature selection is a process that aims to identify a small subset of features from a large number of features collected in the data set. Various feature selection methods are available in WEKA (Waikato Environment for Knowledge Analysis) such as Information Gain (IG),Correlation Attribute (CA), Relief-F(RA), One R(OR) and Symmetrical Uncertainty (SU).

A. Information Gain (IG)

Information Gain is an important measure used for ranking features. Given the entropy is a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(\frac{Y}{X}) = H(X) - H(\frac{X}{Y}) \quad (1)$$

IG is a symmetrical measure. The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.[17]

B. Correlation Attribute (CA)

Correlation shows how two item sets are closely related to each other which can be used for generation of association rule. It shows the dependence of two itemsetsor correlation of two itemsets. If P(XUY)=P(X)P(Y) then two item sets X & Y are independent otherwise X and Y are considered as correlated.

$$Corr(X,Y) = \frac{P(XUY)}{P(X)P(Y)} \quad (2)$$

If Corr(X,Y) is greater than 1 then those attributes are closely related to each other otherwise X & Y are not correlated or independent attributes. With these terms and using association rule generation algorithm the features can be selected.[18]

C. Relief-F (RA)

The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature f.[17]

$$wf=P(\frac{different\ value\ of\ f}{different\ class})-P(\frac{different\ value\ of\ f}{same\ class}) \quad (3)$$

D. One-R

One R is a simple algorithm proposed by Holte. It builds one rule for each attributes in the training data and then selects the rule with the smallest error. It treats all numerically valued features as continuous and uses a straightforward method to divide the range of values into several disjoint intervals. It handles missing values by treating "missing" as a legitimate value. This is one of the most primitive schemes. It produces simple rules based on one feature only. Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes.[19]

E. Symmetrical Uncertainty (SU)

Symmetric Uncertainty is one of the best feature selection methods and the most feature selection systems based on mutual information uses this measure. SU is a correlation measure between the features and the class.

$$SU=\frac{(H(X)+H(Y)-H(\frac{X}{Y}))}{(H(X)+H(Y))} \quad (4)$$

where H(X) and H(Y) are the entropies based on the probability associated with each feature and class value respectively and H(X,Y), the joint probabilities of all combinations of values of X and Y.[19]

V. CLASSIFICATION ALGORITHM USING WEKA

Weka is written in java and can run on any of the platform. We can say that Weka is a collection of algorithms with the help of which real world problems can be solved. Algorithms can be applied either directly or to a dataset called from own java code. The techniques like Data processing, classification, clustering, visualization regression and feature selection are supported by Weka. In Weka data is considered as an instances and features as attributes . In this main user interface is the explorer but essential functionality can be attained by component based on knowledge flow interface and command line whenever simulation is done than the result is divided into several sub items for easy analysis and evolution. One part in correctly or correctly classified instances partitioned into percentage value and numeric value and subsequently kappa statistics mean absolute error and root mean squared error which will in numeric value[20].

Classification is used to find out in which group each data instance is related within a given dataset. In this study we used two classification algorithm and evaluate the classification performance[20].

A. JRIP rule classifiers

Jrip (RIPPER) is one of the most popular algorithms; it has classes that are examined in increasing size. It also includes a set of rules for class is generated using reduced error Jrip (RIPPER). Proceeded by treating examples of judgments made in training data as a class, and finding rules that cover all the members of the class. Then it proceeds to the next class and repeats the same action, repetition is done until all classes have been covered[20].

B. J48

J48 is the classifier according to which we classify our classes. It is also known as free classifier which accepts nominal classes only. In this prior knowledge should be there while classifying instances. It is used in the construction of decision tree from a set of labeled training data using the information entropy. Attributes which we use for helps in building decision tree by splitting it into subset and normalization information gained can be calculated. Splitting process comes to an end when all instances in a subset belong to the same class. Leaf node is being presented or being created to choose that class a possibility which can also be there that none of the feature provides information gain.J48 creates decision nodes up higher in the tree using expected value of the class.J48 can use both discrete and continuous attributes, attributes with differencing lost and training data with missing attribute values[20].

VI RESULTS AND DISCUSSION

The experiment is performed using the Machine Learning UCI Dataset. In this study we used five dataset. To compare the performance of the classification algorithms with feature selection methods, WEKA data mining tool was used, the default parameters were used for each classification algorithms [21]. All experiments are carried out using a ten-fold cross validation approach.

Five feature selection algorithms such as Information Gain (IG),Correlation Attribute (CA), Relief-F(RA), One R(OR) and Symmetrical Uncertainty (SU) are used to select features before passing the data sets to the classifiers. Each datasets is separately classified by two classification algorithms and analysed the results. The datasets are classified using JRIP and J48 classifier. In Table 3 and Table 4 shows the effects of feature selection and classification results of the JRIP and J48 classifier. Figure 1and Figure 2 shows the classification Performance of the JRIP and J48 classifier respectively. The accuracy values of two classification algorithms such as JRIP and J48 is based on the measures Information Gain algorithm shows the highest performance for all the datasets.

Table 2: Features Derived by Various feature selection Methods

No.	Data Sets	Selected Features				
		IG	CA	RA	OR	SU
1	Breast cancer wisconsin	2,3,6,7,5,8,1	6,2,8,5,3,9,4	6,2,3,1,7,8,5	2,3,7,6,8,5,4	2,6,3,5,8,7,4
2	Hypo Thyroid	18,26,22,20,17,3,29	18,26,22,20,17,10,3	28,4,10,6,2,8,11	18,22,26,9,29,10,7	18,26,22,17,3,20,1
3	Derma Tology	21,20,22,33,34,29,27,12,25,6,16,8,28,9,15,10,24,14,5,26,3,19,31	22,20,21,12,27,6,29,25,33,8,24,9,10,15,14,28,23,26,16,1,5,3,19	21,22,20,33,16,27,28,25,29,12,6,8,14,15,9,5,4,10,3,19,24,2,26	21,29,25,12,33,20,27,6,22,8,16,28,15,9,10,14,24,19,2,4,36,7,23	21,22,20,33,27,29,12,25,6,8,15,9,28,16,10,24,14,31,26,5,7,30,34
4	Soy Bean	15,13,14,22,1,29,28,16,4,30	16,15,13,14,23,29,4,28,11,30	22,15,13,14,29,28,1,11,19,24	35,34,11,12,13,14,15,10,9,8	15,13,14,16,22,29,30,28,4,1
5	Auto Price	7,4,12,3,8,15,5,14,2,9	14,15,1,5,8,3,12,4,10,11	3,10,5,8,9,12,4,15,14,6	15,3,14,8,4,12,5,7,9,10	3,15,14,8,4,12,5,7,9,10

Table 3: Effects of feature selection and Classification results on JRIP classifier.

CLASSIFICATION ACCURACY %								
No	Datasets	Features	ALL	IG %	CA %	RA %	OR %	SU %
1	Breast cancer Wisconsin	10	94.4	94.4	93.5	93.8	93.1	93.1
2	Hypothyroid	30	92.3	97.1	93.1	92.2	95.1	93.5
3	Dermatology	34	88.4	94.1	92.1	90.1	88.8	93.4
4	Soybean	35	95.6	98.7	94.5	96.6	86.5	97.7
5	Auto_price	15	88.7	96.9	90.3	93.7	93.7	94.9

Table 4: Effects of feature selection and Classification results on J48 classifier.

CLASSIFICATION ACCURACY %								
No	Datasets	Features	ALL	IG %	CA %	RA %	OR %	SU %
1	Breast cancer Wisconsin	10	94.4	94.4	94.4	94.4	94.2	94.2
2	Hypothyroid	30	92.3	97.9	92.2	92.2	95.1	92.2
3	Dermatology	34	94.5	95.4	91.8	91.8	90.1	93.7
4	Soybean	35	96.9	97.9	94.4	96.3	86.9	96.9
5	Auto price	15	91.8	97.5	91.8	97.5	97.5	95.9

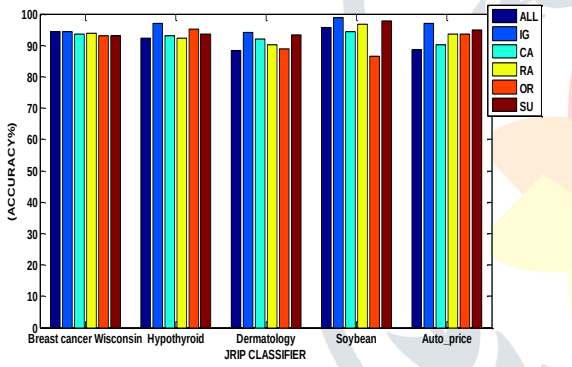

Figure 1: Comparative Analysis of JRIP Classifier Algorithm using UCI Machine Learning Dataset

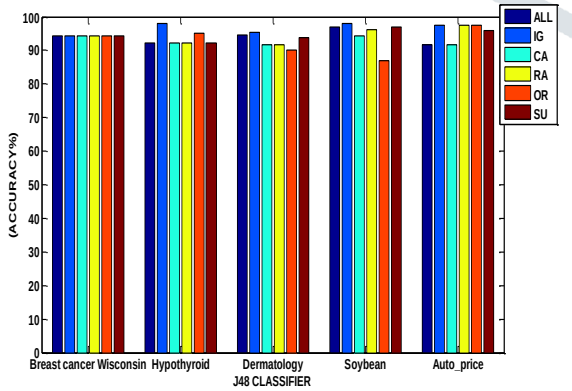

Figure 2: Comparative Analysis of J48 Classifier Algorithm using UCI Machine Learning Dataset

VII CONCLUSION

In this paper, we mainly focused on the performance of five feature selection algorithms such as Information Gain, Correlation Based , Relief-F, One-R., Symmetrical Uncertainty using five UCI machine Learning data sets. The classification performance of the reduced data is measured by WEKA classifiers such as JRIP and J48. Compared to the accuracy of feature selection algorithm Information Gain is proved to be better performance. Hence, we propose Information Gain feature selection algorithm.

ACKNOWLEDGEMENT

The author would like to thank to Data Mining Repository of University of California Irvine (UCI) for providing Machine Learning UCI Dataset.

REFERENCES

[1] Meenatchi V.T, Gnanambal S, et.al, Comparative Study and Analysis of Classification Algorithms through Machine Learning, International Journal of Computer Engineering and *Applications*, 9(1),247-252,2018.

[2] Hany M. Harb1, Malaka A. Moustafa, Selecting optimal subset of features for student Performance model, *IJCSI* , 9(5), 2012, 1694-0814

[3] Hwang, Young-Sup, Wrapper-based Feature Selection Using Support Vector Machine,Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro, Korea, *Life Science Journal*,11 (7), 221-70,2014.

[4] Wang Liping, Feature Selection Algorithm Based On Conditional Dynamic Mutual Information, International Journal Of Smart Sensing and Intelligent Systems,8(1), 2015.

[5] Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data, *IEEE Transactions on Knowledge and* Data Engineering, 25(1), 2013.

[6] Z.Zhao, H.Liu, On Similarity Preserving Feature Selection, *IEEE Transactions on* Knowledge and Data Engineering, 25(3), 2013.

[7] Sunita Beniwal and Jitender Arora, Classification and Feature Selection Techniques in Data Mining, International Journal of Engineering Research & Technology (IJERT), 1(6), 2012.

[8] Mital Doshi and Setu K Chaturvedi, "Correlation Based Feature Selection (Cfs) Technique to Predict Student Performance", International Journal of Computer Networks Communications *(IJCNC*),6(3),2014.

[9] M. Ramaswami and R. Bhaskaran,"A Study on Feature Selection Techniques in Educational Data Mining", *Journal Of Computing*,1(1),December 2009.

[10] Dr.C.Velayutham, Dr.K.Thangavel " Rough Set Based Unsupervised Feature Selection Using Relative dependency Measures", *International Journal of Computational* Intelligence and Informatics, Vol. 1 : No. 1, April - June 2011.

[11] Dr.C.Velayutham, Dr.K.Thangavel," Unsupervised Feature Selection Based on the Measures of Degree of Dependency using Rough Set Theory in Digital Mammogram Image Classification", *IEEE-ICoAC 2011*.

[12] Dr.C.Velayutham, Dr.K.Thangavel," Unsupervised Feature Selection in Digital Mammogram Image Using Rough Set Based Entropy Measure", *World Congress on* Information and Communication Technologies 2011.

[13] Arpita Nagpal, Vijendra Singh," A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data" *International Conference on* Computational Intelligence and Data Science (ICCIDS 2018).

[14] Hossam Faris, Mohammad A. Hassonah Ala, M. Al-Zoubi, Seyedali Mirjalili, Ibrahim Aljarah" A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture", *Published online 2 January 2017*.

[15] K.Sutha and J. Jebamalar Tamilselv, A Review of Feature Selection Algorithms for Data Mining Techniques, International Journal on Computer Science and Engineering (IJCSE), 7(6) ,2015.

[16] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, UCI Repository of machine learning databases, University California Irvine, Department of Information and Computer Science,1998.

[17] Thu Zar Phyu , Nyein Nyein Oo," Performance Comparison of Feature Selection Methods", MATEC Web of Conferences Owned by the authors, published by EDP Sciences, 2016.

[18] K.Rajeswari , Dr.V.Vaithiyanathan and Shailaja V.Pede," Feature Selection for Classification in Medical Data Mining, Volume 2, Issue 2, March – April 2013 ISSN 2278- 6856.

[19] Jasmina NOVAKOVIĆ, Perica STRBAC, Dusan BULATOVIĆ," TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS", Yugoslav Journal of Operations Research 21 (2011), Number 1, 119-135.

[20] Meenakshi, Geetika," Survey on Classification Methods using WEKA", International Journal of Computer Applications (0975 – 8887) Volume 86 – No 18, January 2014.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl., 11 (2009) 10-18.