A Mutual Information-Based Feature Selection for Multiclassification Detection

Karthik A[1], Sreedevi M[2]

[1]Dept of Computer Science, S V University, Tirupati
[2]Assistant Professor, Dept of Computer Science, S V University, Tirupati

Abstract— *Feature determination technique is utilized for creating an ideal number of features to be utilized for a specific errand like characterization. The goal of the disposing of interaction is to diminish the size of the information include set and simultaneously to hold the class prejudicial data. This paper proposes and assesses another component choice calculation utilizing data hypothesis which is the Mutual Information data (MI) between mixes of information highlights and the class rather than shared data between a solitary info include and the class for both consistent esteemed and discrete-esteemed highlights. Feature class MI has been utilized to choose a subset of highlights dependent on its significance. The investigation is done on the Vehicle dataset taken from the University of California at Irvine Machine Learning Data Repository. The dataset contains a tremendous volume of feature estimations which are diminished using MI based segment assurance procedure. The dataset contains a gigantic rundown of abilities which is reduced using a further developed segment decision procedure named as covering strategy. The chose subset of highlights then, at that point goes through a preprocessing step to present a consistency in the circulation of information. Since Multilayer Perceptron (MLP) is perceived to have the advantage of giving a prominent execution in grouping phase. The order exhibitions have been found promising when contrasted and arrangements performed utilizing typical classifiers and with utilizing shared data.*

I. INTRODUCTION

The search for efficient and effective algorithms of input feature selection about classification problems has a long history within the field of pattern recognition. Reduction of the input features set can be desirable, or essential for a number of reasons, such as reducing the complexity of building and operating a classifier [3]. In addition to the computational-cost saving, feature-space reduction can also re-duce the actual cost of feature collection and pre-processing, and even lead to an improvement in classifier accuracy. Mutual information (MI) which is a measure of the amount of information between two random variables is symmetric and non-negative and is zero if and only if the variables are independent [6]. This paper proposes a powerful feature selection algorithm based on the MI concept to offer significant improvement over these earlier attempts. The improvements are in terms of wider applicability, as good or better performance on the example classification tasks, a reduction in computational complexity and more information produced on the specific feature interaction presented in a classification problem.

1.1 Feature Selection

Feature selection has been widely investigated and used by the machine learning and data mining community. In this context, a feature, also called attribute or variable, represents a property of a process or system than has been measured or constructed from the original input variables. The goal of feature selection is to select the smallest feature subset given a certain generalization error, or alternatively finding the best feature subset with features, that yields the minimum generalization error [1]. Additional objectives of feature selection are as follows: (i) improve the generalization performance with respect to the model built using the whole set of features, (ii) provide a more robust generalization and a faster response with unseen data, and (iii)achieve a better and simpler understanding of the process that generates the data [3]. Feature selection methods are usually classified in three main groups: wrapper, embedded, and filter methods. Wrappers use the induction learning algorithm as part of the function evaluating feature subsets. The performance is usually measured in terms of the classification rate obtained on a testing set, i.e., the classifier is used as a black box for assessing feature subsets [2] [9]. Although these techniques may achieve a good generalization, the computational cost of training the classifier a combinatorial number of times becomes prohibitive for high-dimensional datasets

1.2 Mutual information (MI)

The MI which is a measure of the dependence between the random variables is always symmetric and non-negative. It is zero if and only if the variables are independent. The mutual information [3][4] between two discrete random variables U=(u1, u2,…,uk) and V= (v1, v2,…,vd) is defined as

$$I\ (U,V) = \ \Sigma_u \Sigma_v \mathrm{p(u,v)} \log \frac{p(u,v)}{p(u)p(v)}$$

Where U=(u1, u2,…,uk) and V= (v1, v2,…,vd) are the values of the discrete variables U and V respectively. p(u,v) is a joint density function and p(u) and p(v) are the marginal density functions.

In this method, a mutual information measure is used to calculate the information gain among features as well as between feature and class attributes. Using a greedy manner, each time we pick a feature from the feature set still left one that provides maximum information about the class attribute with minimum redundancy.

II. METHODOLOGY

2.1 Artificial neural organization (ANN)

An ANN is an information getting ready perspective that is impelled by the way where a characteristic tangible framework in human brain works. ANNs are used comprehensively for the course of action of different issues, including portrayal, vision, talk, plan affirmation, control structures, etc. A colossal number of neurons present in the human frontal cortex structures the vital part of the neural framework perspective and go probably as simple taking care of segments [4]. A fake neuron is a little getting ready unit and plays out a clear computation that is fundamental to the action of a neural framework. The model of a neuron contains the essential parts like wellsprings of data, synaptic burdens, inclination, adding crossing point, and incitation work.

2.1.1 Multilayer Perceptron (MLP)

A MLP is a champion among the most generally perceived Neural Network plan that has been used for various applications. The MLP organize is commonly made out of different centers or dealing with units, and it is figured out into a movement of no less than two layers [5]. The essential layer (or the most diminished layer) is named as an information layer where it gets the external information while the last layer (or the most dumbfounding layer) is a yield layer where the response for the issue is gotten. The disguised layer is the widely appealing layer in the data layer and the yield layer, and may frame with somewhere around one layers. The arrangement of MLP could be communicated as a nonlinear improvement issue. The objective of MLP learning is to find the best loads that limit the differentiation between the information and the yield. The most predominant getting ready estimation used in NN is Back propagation (BP), and it has been used in dealing with various issues in model affirmation and portrayal. This computation depends on a couple of boundaries, for instance, different covered center points at the hid layers learning rate, energy rate, enactment work and the quantity of preparing to happen. Besides, these boundaries could change the exhibition on the gaining from awful to great exactness [7].

2.2 Logistic Regression

Logistic Regression is considered as the standard measurable way to deal with displaying double information [4]. The focal numerical idea that underlies logistic regression is the logit—the normal logarithm of an odds ratio. It's anything but a superior option for a straight relapse which appoints a direct model to every one of the classes and predicts inconspicuous cases basing on greater part vote of the models. By and large, calculated relapse is appropriate for describing and testing speculations about connections between an all-out result variable and at least one clear cut or constant indicator factors. During expectation, rather than foreseeing the point gauge of the actual occasion, it's anything but a model to anticipate the chances of its event. In two class issue for instance, when the chances are more noteworthy than half, then, at that point the case is relegated to the class assigned as "1" for YES and "0" for "YES" and "NO" all things considered.

III. EXPERIMENTAL RESULTS

This part gives results and related conversation on information driven analysis of vehicle dataset was gathered from UCI repository [8]. This exploration work was executed utilizing Weka. WEKA is made by analysts at the University of Waikato in New Zealand. The product is written in the Java language and contains a GUI for communicating with information documents. WEKA additionally gives the graphical UI of the client and gives numerous offices. WEKA is a cutting-edge

office for creating ML methods and their application to true information mining issues. These records were arranged into four classes, contains 846 instances and 19 attributes. The analyses were performed considering 593 examples which implies 70% of the complete examples were preparing information and 30% were trying information. The Vehicle dataset details are shown in the figure-1.

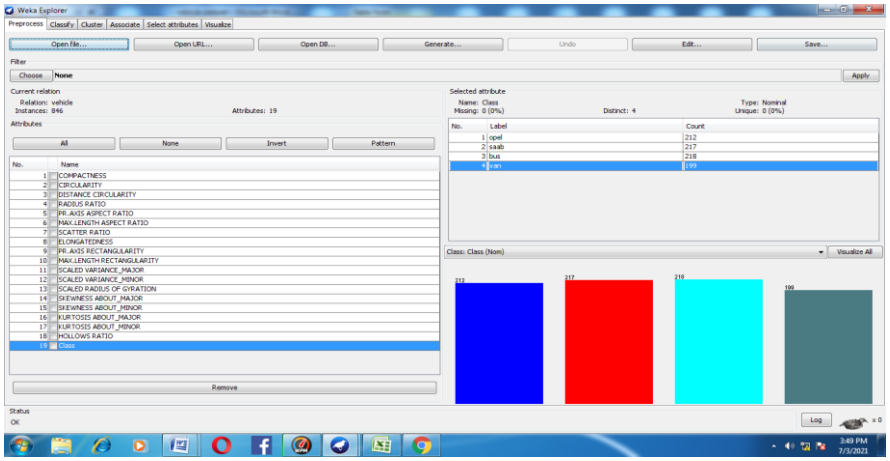

Figure-1: Summary of Vehicle dataset

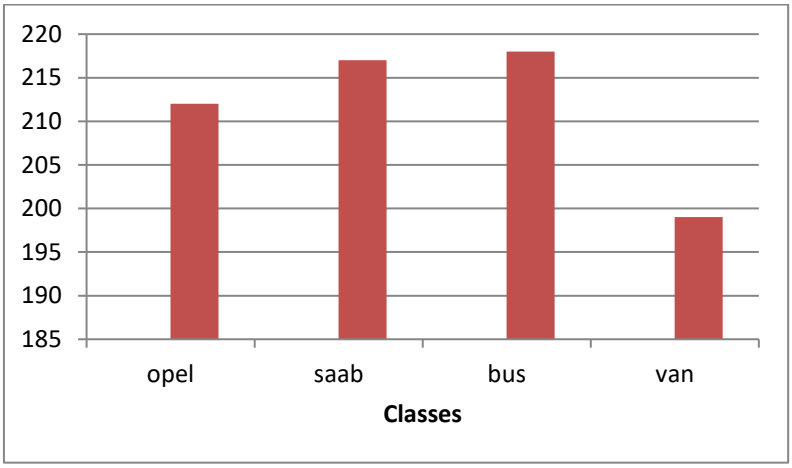

Figure-2: Class wise distribution of labels

The Statistical Summary of the dataset details are shown in the figure-3

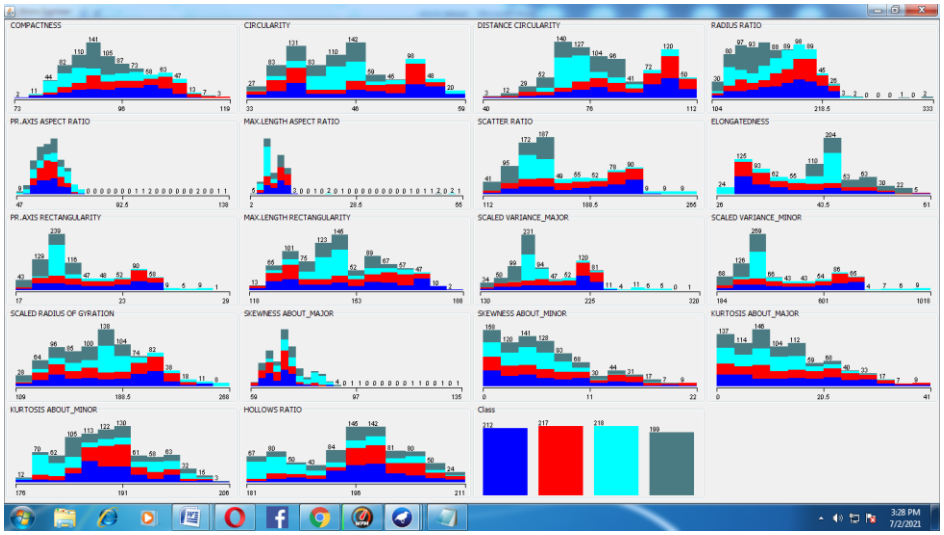

Figure-3: Detailed statistical summary of Vehicle dataset

3.1 Results

In the first stage MLP and Logistic Regression algorithms are trained on the original set of features was used in the experiment. In the second stage we implement a MI algorithm for obtaining the adequate number of features to identify the features selected. The results that we got for MLP and Logistic Regression without feature selection and with feature selection are shown below in the table-1 and same as shown in the figure-4 with their corresponding values.

TABLE-2
PERFORMANCE OF CLASSIFIERS

Algorithm	Accuracy	Precision	Recall
Logistic Regression without MI	79.18	79.6	79.2
Logistic Regression with MI	81.78	81.9	81.8
MLP without MI	86.7	86.4	86.7
MLP with MI	90.8	90.9	90.9

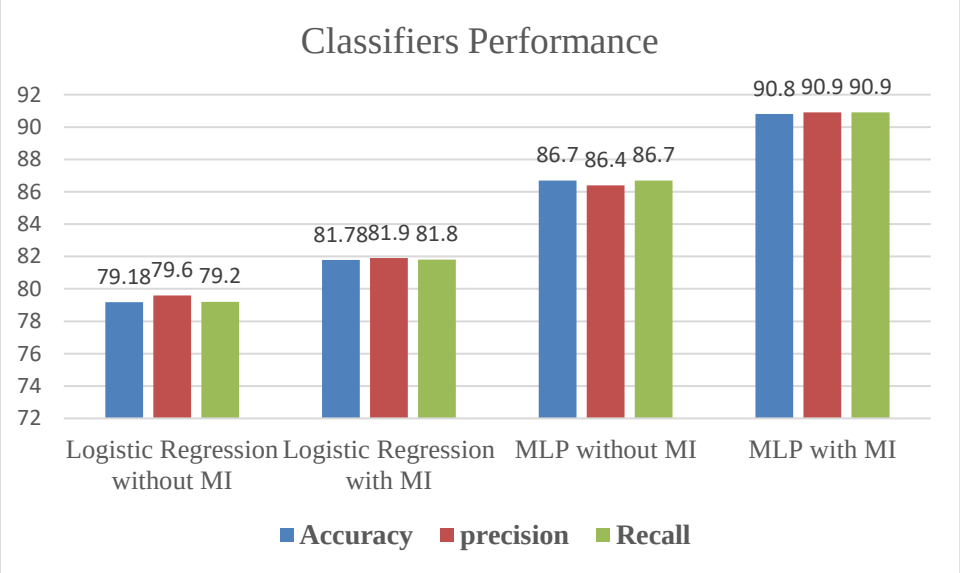

Figure-4: Experimental Results

From the figure-4, we observe the performance of Logistic Regression without MI based on accuracy has got 79.18%, whereas the performance of Logistic Regression with MI feature selection based on accuracy has achieved 81.78%. However, there is an improvement in the accuracy with feature selection. The accuracy rate is increased 2.6% with feature selection.

Also, we observe the performance of MLP without MI based on accuracy has got 86.7%, whereas the performance of MLP with MI feature selection based on accuracy has achieved 90.8%. However, there is an improvement in the accuracy with feature selection. The accuracy rate is increased 4.1% with feature selection.

In our experimental result the MLP with MI feature selection algorithm shows the highest accuracy compared with Logistic Regression with MI. With the improvement the accuracy, the proposed model demonstrated that it performs well after selecting relevant features. This result provided new insight using a classification learning algorithm and reduction technique to selection relevant and important feature in order to improve the accuracy of the system and to identify possible features which may contribute to this improvement. Most of the proposed research system could effectively utilize feature selection process to improve detection rate of their system and minimize considerably the false alarm rate.

IV. CONCLUSION

This paper has investigated the approaches to solve the important classification problem of the feature selection. A presentation and proposition of a feature selection method which consist of a MI feature elimination using a MLP and Logistic Regression classifiers to identify important features have been done. The feature selection, preprocessing, and classification techniques have produced a combination which provides promising results for classification. The evaluation the effectiveness of the method using different classification metric measurement has been made and it has been proved that by reducing the number of features, the accuracy of the model was improved. In order to detect class from large dataset, detection algorithm, and feature selection method have too more efficient.

REFERENCES

[1] G. Ravi Kumar, K. Nagamani and G. Anjan Babu, "A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction", Lecture Notes on Data Engineering and Communications Technologies, ISBN 978-981-15-0977-3, Volume 37, PP:173-180, Springer Nature Singapore Pte Ltd. 2020.
[2] G. Ravi Kumar, Venkata Sheshanna Kongara & Dr. G. A. Ramachandra, "An Efficient Ensemble Based Classification Techniques for Medical Diagnosis", International Journal of Latest Technology in Engineering, Management and Applied Sciences, Volume II, Issue VIII, Pages: 5-9, ISSN-2278-2540, August-2013.
[3] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
[4] H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations", San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2000).
[5] J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann.
[6] Kohavi R, John GH (1997) Wrappers for feature subset selection. ArtifIntell 97(1–2):273–324
[7] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems", 2nd edition, Addison Wesley, 2005.
[8] UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets.html)
[9] Y. pang, Y. Yuan, and X. Li. 2008 "Effective feature extraction in high dimensional space, IEEE Trans. Syst., Man, Cybern. B, Cybern.
[10] Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., Chen, Z.: Effective and efficient dimensionality reduction for large-scale and streaming data.