International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Effective Feature Selection Strategy using Manova Test

V.Sathya Durga, Thangakumar Jeyaprakash

***Abstract*: Feature selection is the most important step to develop any latest learning model. As the complexity of the leaning models increases day by day there is an increasing demand, in selecting the right features to build the model. There are many methods for feature selection. A new feature selection based on the Manova statistical test is implemented. Using the Manova test, we select attributes from academic datasets. Using the selected attributes, we build a classification model. Accuracy of the model with feature selection is compared with a model with all attributes. Results are discussed. It is proved that the classification model build with features selected by Manova test achieves more accuracy than a model built with all features.**

***Index Terms*: Feature Selection, Manova, Wiki Lambda.**

I. INTRODUCTION

Feature selection is the process of selecting the right attributes to build a learning model. Selecting the right attributes increases the accuracy of the model built. Whereas the wrong selection of attributes decreases performance of the model notably. There are three types of feature selection methods in existence. These methods are Filter, Wrapper and Embedded Method. Filter method uses a statistical test to identify the right features. In the wrapper method, a subset of feature is selected first, then the model is trained and results are obtained. Based on the result it is decided to select or reject particular features from the dataset. Embedded methods combine both the above mentioned method for feature selection.

In this paper, we implement filter based method. The statistical test which we use for feature selection is Manova. Manova is a statistical technique being in existence from the year 1925 [1]. Manova is used in statistical samples where there are many factors which affect the dependent variables. In such a case, Manova is used to determine the most important factor among them. Manova is an extension of Anova. It uses covariance of the output variable to test the statistical difference. Some of the advantages of Manova are, it is easy to study the interaction between factors. It reduces Type 1 errors. Analyzing these properties of Manova, we can determine that the Manova test can be used effectively for feature selection. In this paper, for feature selection, we use Manova statistical technique. There are many types of Manova tests. We use Wilks Lambda test. The base formula for Manova's Wilks Lambda test,

$$\Lambda_{p,h,e} = \frac{|E|}{|E+H|} = \prod_{j=1}^{p}(1-\theta_j)$$

Where E and H are matrices, θj is the Eigen values [2].

In this paper Section I, is the introduction part, Section II contains the review of literature on various feature selection methods used in the literature. Section III contains the methodology followed for feature selection. Section IV is the result section. Section IV.A contains results of Manova tests and IV.B contains results of models build with and without feature selection. Section V contains discussions and conclusion of this research work.

II. REVIEW OF LITERATURE

Pratik Kadam uses a Chi-Square test as a feature selection method for academic datasets [3]. Li et al uses clustering and Chi-Square technique for feature selection from statistical data [4]. Other Machine learning techniques like genetic algorithms, simulated annealing is used for feature selection.

III. METHODOLOGY

The methodology followed in this work is as follows. Students academic datasets are downloaded from the UCI repository. The dataset contains 395 students record from a Portuguese school. In this dataset, students academic details and non academic details are stored. To the above mentioned dataset, Manova test was conducted using NCSS 2019 statistical software and setting the significance level to 0.05. Manova test report was generated.

Manova report contains for each attribute, test values, F Ratio and decision to accept or reject those attribute. We omit the attributes which are rejected in the Manova report and select the attributes accepted in the test. We build two classification model using decision tree classifier. In the first model all features are used to build the classification model and in the second model features selected by the Manova test are used to build the classification model. Accuracy of both the model is compared and results are tabulated.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
V.Sathya Durga*, Research Scholar, Department of CSE, Hindustan Institute of Technology and Science, Padur, India.
Thangakumar Jeyaprakash, Associate Professor, Department of CSE, Hindustan Institute of Technology and Science, Padur, India.

IV. RESULTS

A. Manova Experimental Results

Table. I Manova Test Results

S.No	Attributes	Test Value	DF1	DF2	F Ratio	Prob Level	Decision (0.05)
1.	Sex	0.294679	17	377	1.19	0.270533	Accept
2.	Age	2.237057	17	377	1.40	0.133909	Accept
3.	Address	0.268647	17	377	1.59	0.064734	Accept
4.	Famsize	0.211018	17	377	1.03	0.428073	Accept
5.	Medu	2.247179	17	377	1.95	0.013294	Reject
6.	Fedu	1.538952	17	377	1.32	0.177724	Accept
7.	Mjob	1.889752	17	377	1.27	0.209391	Accept
8.	Fjob	0.672720	17	377	0.90	0.576528	Accept
9.	Reason	2.195557	17	377	1.54	0.078322	Accept
10.	Guardian	0.279550	17	377	0.69	0.815379	Accept
11.	Traveltime	0.575676	17	377	1.19	0.266826	Accept
12.	Study time	0.746122	17	377	1.06	0.389716	Accept
13.	Failures	2.770425	17	377	6.12	0.000000	Reject
14.	Schoolsup	0.215081	17	377	1.99	0.011190	Reject
15.	Famsup	0.191129	17	377	0.80	0.697790	Accept
16.	Paid	0.434775	17	377	1.81	0.025456	Reject
17.	Activities	0.227522	17	377	0.90	0.569151	Accept
18.	Nursery	0.154175	17	377	0.94	0.525533	Accept
19.	Higher	0.068940	17	377	1.46	0.106662	Accept
20.	Internet	0.146942	17	377	1.06	0.396518	Accept
21.	Romantic	0.394910	17	377	1.83	0.022663	Reject
22.	Famrel	0.489717	17	377	0.60	0.893291	Accept
23.	Freetime	0.852533	17	377	0.85	0.635624	Accept
24.	Goout	1.915149	17	377	1.58	0.065410	Accept
25.	Dalc	1.488232	17	377	1.95	0.013250	Reject
26.	Walc	3.833503	17	377	2.46	0.001143	Reject
27.	Health	2.127780	17	377	1.11	0.345589	Accept
28.	Absences	217.664370	17	377	3.81	0.000001	Reject

The features which are mentioned as Accept are picked to build the model. Accuracy of the model without any feature selection and with features selected by using Manova values are as follows.

B. Classification Model Results

Accuracy of classification models build without feature selection and with feature selection is discussed here. We use the decision tree classifier to build the classification model. This part of the research work was carried out in Matlab 2019a.

Table. II Accuracy Comparison

S.No	Method	Accuracy
1.	Classification Model with all Features	83.5 %
2.	Classification Model with Feature Selected by Manova	86.2 %

From the table, we can infer that classification model built with features selected by Manova are more accurate than models built with all features.

V. DISCUSSIONS AND CONCLUSION

In this paper we use Manova test for feature selection of academic datasets. Students academic datasets are downloaded from the UCI repository. We apply Manova's Wiki Lamda test to the academic dataset. Test reports are generated with F-Ratio, Prob level, Decision to accept or reject the attribute. We selected the accepted attribute and build a classification model. We build another classification model with all attributes. Finally, the accuracy of both models is compared. It is found out that model build with features selected with Manova achieves more accuracy than model build with all features. So it is recommended to researchers working on learning models to use Manova for feature selection.

REFERENCES

1. What is manova, Available: https://www.researchoptimus.com/article/what-is-manova.php.
2. Manova. Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Multivariate_Analysis_of_Variance-MANOVA.pdf.
3. Pratik Kadam,"Analysis of Chi-Square Independence test for Naive Bayes feature selection," International research journal of engineering and technology, Dec 2018, Vol.5, No.12, pp.1382-1386.
4. S.Li,H.Wu, D.Wan and J.Zhu. "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine", Knowledge-Based Systems, February 2011, Vol. 24, No.1, pp.40-48.

AUTHORS PROFILE

V.Sathya Durga is a Research Scholar at Hindustan Institute of Technology and Science in the Department of CSE. She is currently pursuing her Ph.D. in Machine Learning and Artificial Intelligence.

Thangakumar Jeyaprakash is an Associate Professor in the Department of CSE, Hindustan Institute of Technology and Science. He has completed his Ph.D. in Network Security. His areas of interest are Machine learning and Artificial intelligence.