iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** II **Month of publication:** February 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

Feature Selection with Data Re-Construction of Standardized Search with Decision Tree

S.Abirami[1], P. Durga[2], J.Vijitha[3], M.S.Vijaykumar[4]
[1, 2, 3, 4] *(Department of Information Technology) Tejaa Shakthi Institute of Technology for Women, Coimbatore.*

Abstract: ***Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. Feature selection in neural network can select essential features and discard derogatory and indifferent features. Such a method may pick up some useful but dependent features, all of which may not be needed. The proposed scheme, named as Feature Selection Multi-layer Perception (FSMLP), can select features with a controlled redundancy both for classification and function approximation/prediction type problems. The effectiveness of the algorithms uses the measure of linear dependency to control the redundancy.***
The use of nonlinear measures of dependency, such as mutual information, is straightforward. These methods can account for possible nonlinear subtle interactions between features, as well as that between features, tools, and the problem being solved. They can also control the level of redundancy in the selected features.
Currently the amount huge of data stored in educational database these database contain the useful information for predict of students performance. The most useful data mining techniques in educational database is classification. In this work work, the classification task is used to predict the final grade of students and as there are many approaches that are used for data classification, the decision tree (ID3) method is used here.
Keywords: Feature Extraction, Multi-layer perception

I. INTRODUCTION

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. *Feature selection* is commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods. There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance.

II. PROBLEM DEFINITION

Feature selection plays an important role in pattern recognition and system identification. It is well known that, for a given problem, all features that characterize a data point are not usually of equal importance; some features may have a derogatory influence on the task at hand. This may be true irrespective of whether the problem is of classification, function approximation, or prediction. Use of more features adds more (flexibility) degrees of freedom to the system, and hence the learning system gets higher freedom to memorize the data, which may result in poor generalization. When we use more features, it may lead to higher design and decision-making cost. Feature selection methods can be classified in different ways. One commonly used classification method groups the feature selection methods into filter methods and wrapper methods. The filter methods do not use/require any feedback from the classifier or the predictor, which finally use the selected features. However, a wrapper method assesses the utility of the features using the classifier (or the predictor), which finally uses the selected features.

III. EXISTING SYSTEM

Finding the optimal subset of features usually requires an exhaustive search considering all possible subsets of features, which becomes computationally prohibitive when the dimension of the data is high. Therefore, even for a wrapper

method, some suboptimal heuristic guided selection methods are used. Use of forward selection or backward elimination schemes or their variants may not be able to exploit the interaction between features. the effectiveness of a feature set depends not only on the problem, but also on the tool that is used to solve the problem.

A. Drawbacks

1) Quality of data is less
2) Redundancy is high
3) Use of forward selection variants may not be able to exploit the interaction between features.
4) Performance is less

IV. PROPOSED SYSTEM

The best way of feature selection should be an integrated approach where the learning system looks at all features simultaneously and picks up useful features while designing the system for solving the given problem. There are a few methods that address the feature selection problem using such an integrated framework. This type of methods has some advantages over other approaches: they do not need to evaluate all possible subsets, they can account for interaction between features, and they can acknowledge the interaction between the features and the tool that is used to solve the problem. Various algorithms and techniques like Classification Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.ID3 Algorithm is implemented to provide better results. The proposed scheme, named as Feature Selection Multi-layer Perception (FSMLP) is to be implemented.

A. Advantages

1) Quality of data is high
2) Redundancy can be reduced
3) Performance can be improved

B. Module Description

The project is divided into six modules.

1) Data Set
2) Pre-processing data
3) Entropy Calculation
4) Information Gain Calculation
5) Tree Formation
6) Feature Selection

C. Data Set

A large dataset of university is taken as for the development of the project. The result based analysis is taken for the feature.

D. Pre-Processing Data

Data pre-processing is an important step in the data mining process. The irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult to process. Data preparation and filtering steps takes considerable amount of processing time.

E. Entrophy Calculation

Entropy is the sum of the probability of each label times the log probability of that same label. Entropy on the other hand is a measure of impurity (the opposite).

F. Information Gain Calculation

This measure of purity is called the information gain. It represents the expected amount of information that would be needed to specify whether a new instance (first-name) should be classified male or female, given the example that reached the node. Calculation is based on the number of male and female classes at the node.

Information Gain = Entropy before - Entropy after

G. Tree Formation

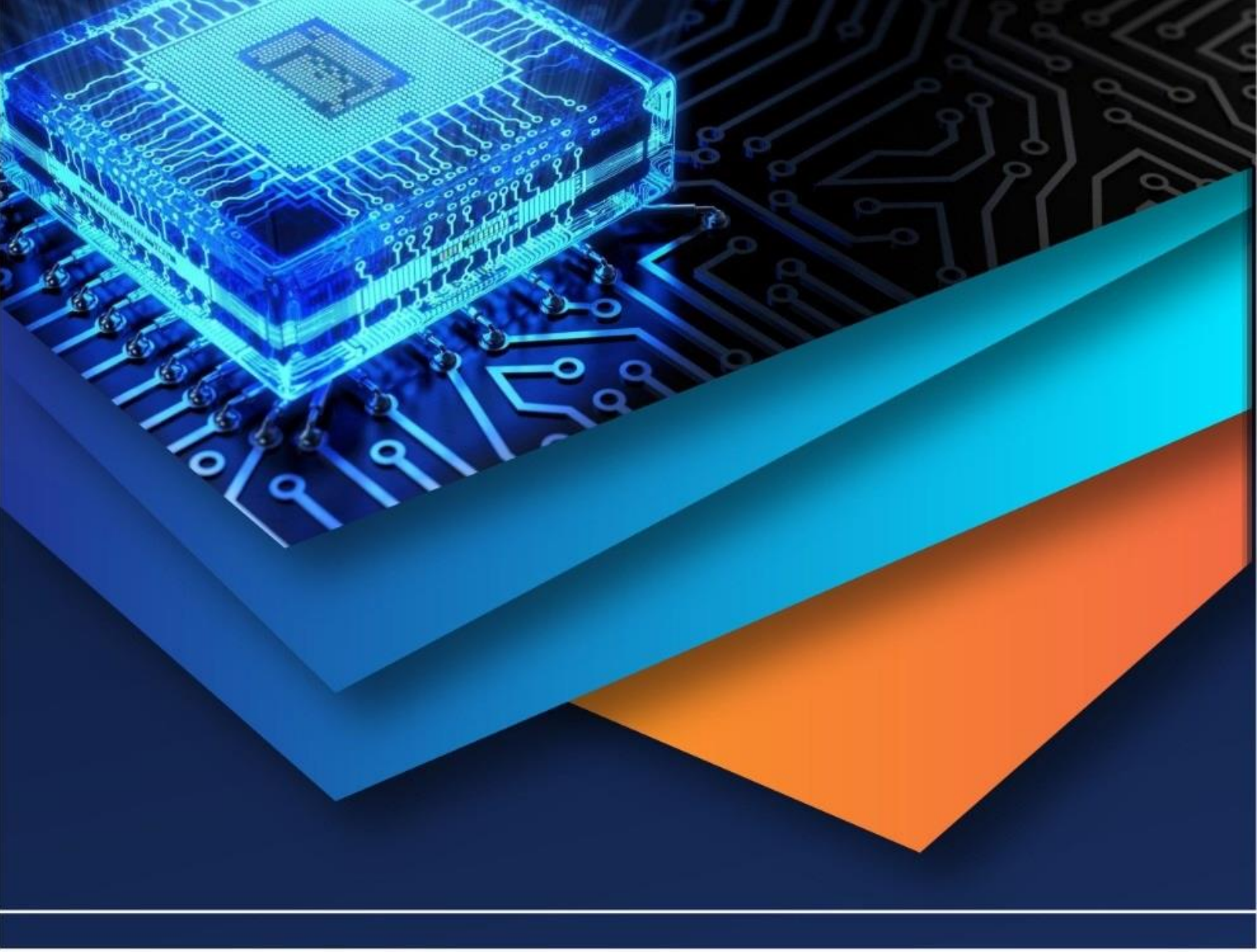
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes

V. CONCLUSION

The feature selection for the semi-supervised data is identified with knowledge discovery. In future the efficiency of the search can improved with increasing the dataset.

REFERENCES

[1] Yu, G. Zhang, Z. Zhang, Z. Yu, and L. Deng, "Semi-supervised classification based on subspace sparse representation," Knowl. Inf. Syst., vol. 43, no. 1, pp. 81–101, 2015.
[2] J. Yu, X. Gao, D. Tao, X. Li, and K. Zhang, "A unified learning framework for single image super-resolution," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 4, pp. 780–792, Apr. 2014.
[3] Q. Zhu, L. Shao, X. Li, and L. Wang, "Targeting accurate object extraction from an image: A comprehensive study of natural image matting," IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 2, pp. 185–207, Feb. 2015.
[4] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint Schatten p-norm and p-norm robust matrix completion for missing value recovery," Knowl. Inf. Syst., vol. 42, no. 3, pp. 525–544, 2013.
[5] X. Lu and X. Li, "Multiresolution imaging," IEEE Trans. Cybern., vol. 44, no. 1, pp. 149–160, Jan. 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY