Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.359 – 364

REMOVAL OF REDUNDANT AND IRRELEVANT DATA FROM TRAINING DATASETS USING SPEEDY FEATURE SELECTION METHOD

Mrs. Radha R [1], Mr. Muralidhara S [2]

[1]M.Tech Student Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India
[2]Assistant Professor, Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India
[1] radhaprasadnr@gmail.com; [2] dhara.me.uvce@gmail.com

Abstract— Prediction plays important role in decision making system which leads to right decision for saving life, energy effort and cost. Prediction is carried by model called classifier. The prediction accuracy depends on the training data sets used for training the classifier. Hence irrelevant and redundant data should be removed from training data sets to improve the predictive speed, reduce the time taken to build the predictive model and reduce the number of features present in the training dataset. Proposed system works on selecting features for supervised and unsupervised learning. This is classified into wrapper, filter, embedded and hybrid methods. This algorithm selects the most significant features from the dataset and removes the redundant and irrelevant features based on information gain feature selection algorithm. The system achieves better prediction accuracy than the other feature selection algorithms naive Bayes (NB), instance based (IB1) and tree based J48. Proposed System gives better prediction speed for classifiers.

KEYWORDS: Feature selection, classification, clustering, prediction, feature ranking, predictive model.

I. INTRODUCTION

Machine Learning is the study of algorithm which makes the computer programs to improve automatically through experience. Predictive models are used to make decisions that are known as classifier or supervised learner. Decision making systems are used to predict the unknown or unobserved data in applications. The correctness of the decision making system relies on the datasets used to train the System. The irrelevant and redundant datasets results in poor accuracy of the predictive models they should be removed from the training datasets to get good performance of predictive models. Good performance means reducing time taken to build predictive model, reducing the number of features present in the dataset and improving predicting accuracy. Feature selection is a process of choosing the efficient data from the dataset for supervised and unsupervised learning. This feature

selection is classified as wrapper, filter, embedded and hybrid methods. The filter method selects the important features and removes the redundant and irrelevant features from the given dataset using Entropy measure.

The embedded method uses the training phase of the supervised learning algorithm for selecting the important features from the training dataset. Therefore, its performance is based on the classifiers used. The wrapper method uses the classifier to determine important features available in the dataset. The combination of filter and wrapper method is known as hybrid method. Many feature selection algorithms are proposed recently, which concentrate only on removing the irrelevant features. They do not deal with removing the redundant features from the training dataset. These redundant features reduce the accuracy of the classifiers.

This paper proposes a filter based feature selection algorithm named as unsupervised learning with entropy based feature selection. This algorithm selects the most significant feature from the dataset and removes redundant and irrelevant features. The redundant data is removed using entropy measure, creating the graph, constructing minimum spanning tree, tree partitioning and representing attribute selection. The significant features are selected based on threshold function. In the following chapters all the related research works are presented.

II. RELATED WORK

This section discusses various types of feature selection, cluster and classification algorithms.

2.1 Feature selection algorithm

The feature selection algorithm selects the most significant features from the dataset using ranking based techniques, subset based and unsupervised based techniques. In the ranking based technique, the individual features "fi " are ranked by applying one of the mathematical measures such as information gain, gain ratio, on the training dataset "TD". The ranked features are selected as significant features for learning algorithm by a threshold value "TV " calculated by the threshold function. In subset based technique, the features of the training datasets are separated into maximum number of possible feature subsets "S" and each subset is evaluated by evaluation criteria to identify the significance of the features present in the subset. The subset containing most significant features is considered as a selected candidate feature subset. In unsupervised based technique, the cluster analysis is carried out to identify the significant features from the training dataset.

2.2 Feature selection based on correlation (FS-Cor)

In this feature subset selection, the entire feature set F = {f1 , f2 , • • • , fx } of a training dataset "TD " is sub divided into feature subsets "FS i ". Then, two types of the correlation measures are calculated on each feature subset "FS i ". One is the feature-feature correlation that is the correlation measure among the features present in a feature subset "FS i ", and the other one is feature-class correlation that is the correlation measure between the individual feature and the class value of the training dataset. These two correlation measures are computed for all the individual feature subsets of the training dataset. The significant feature subset is identified based on the comparison between the feature-feature correlation and feature−class correlation. If the feature−class correlation value is higher than the feature−feature correlation value, the corresponding feature subset is selected as a significant feature subset of the training dataset.

2.3 Feature selection based on $\chi 2$ (FSChi)

This is a ranking based feature selection technique. The $\chi 2$ statistical measure is applied on the training dataset "TD" to identify the significance level of each feature presents in the training dataset. The $\chi 2$ value "CV" is computed based on the sum of ratio of the difference between the observed (oi,j) and expected (eij) frequencies of the features fi , fj to the expected frequencies (eij) of the features fi , fj of the possible instance value combinations of the features.

2.4 Feature selection based on information gain (FSInfo)

In this feature selection technique, the information gain measure is applied on the training dataset to identify the significant features based on information gain value of the individual features in terms of entropy. The

entropy value of each feature of the training dataset "TD" is calculated and ranked based on the information gain value.

2.5 Feature selection based on gain ratio (FSGai-ra)

In this feature selection technique, the information gain ratio GR(f) is calculated for each feature of the training dataset "TD" to identify the significant feature based on the information present in the features of the"TD".

2.6 ReliefF

This feature selection technique selects the significant features from the training dataset "TD" based on the weighted probabilistic function w(f) with nearest neighbor principle. If the nearest neighbors of a tuple "T " belong to the same class, it is termed as "nearest hit". If the nearest neighbors of a tuple "T" belong to a dfferent class, it is termed as "nearest miss". The probabilistic weight function value w(f) is calculated based on "nearest hit" and "nearest miss" values.

2.7 Feature selection based on symmetric uncertainty (FSUnc)

This technique uses the correlation measure to select the significant feature from the training dataset "TD". In addition to that, the symmetric uncertainty "SU" is calculated using the entropy measure to identify the similarity between the two features f_i and f_j. Since the above mentioned researches work either on irrelevant feature removal or redundant data removal, they still do not work accurately. The proposed system works on removing both irrelevant and redundant data using Entropy value of each feature to feature and feature to class correlation. Irrelevant data is removed by comparing the entropy of each feature to class correlation with threshold.

III. EXISTING SYSTEM

Many studies were proposed for selecting significant datasets from training data, but they either remove irrelevant data or redundant data not both. Presence of duplicate data or irrelevant data in training datasets results in poor performance of the classifiers. It takes more time to build a learning system and results in lower accuracy of prediction.

Disadvantages

The disadvantages of the existing system are listed below:-

- It either works on removal of irrelevant data or redundant data from training datasets not both.
- It may also cause the learning system to become poor in terms of time and accuracy of prediction.
- The performance of the classifiers decreases greatly for the presence of irrelevant and redundant data in training datasets.

IV. PROPOSED SYSTEM

Proposed system works on a new evaluation function to measure the ability of feature subsets in distinguishing between class labels, a solution based on the concept of mutual information. The proposed function is based on the information gain and takes into consideration how features work together. The performance of this function works well compared to that of other measures which evaluate features individually.

We propose a new feature selection method in this project. Specifically, for each feature we use its value to rank the training instances, and define the ranking accuracy in terms of a performance measure or a loss function as the importance of the feature. We also define the correlation between the ranking results of two features as the similarity between them. Based on the definitions, we formulate the feature selection issue as an optimization

problem, for which it is to find the features with maximum total importance scores and minimum total similarity scores.

Advantages

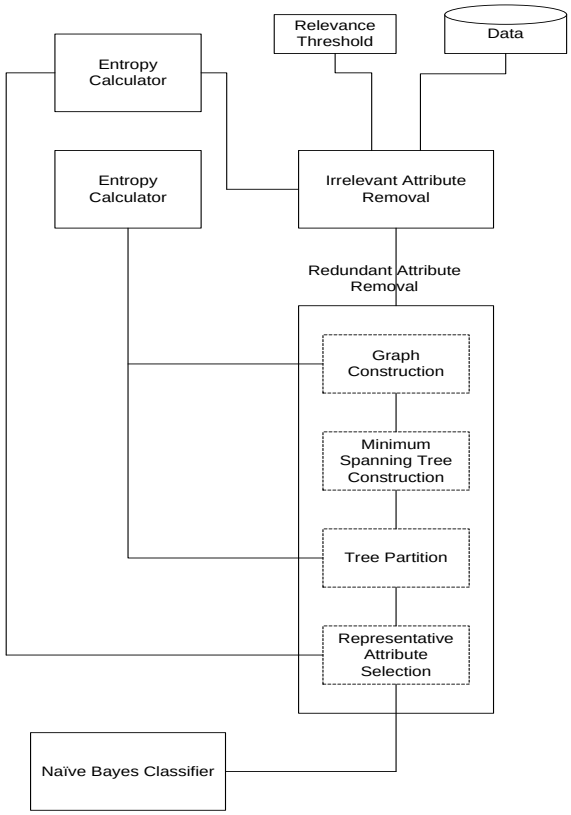
The advantages are listed below:-

- Due to the new structure in the proposed mechanism the redundant and irrelevant data are removed efficiently.
- In the proposed mechanism, works on both supervised and unsupervised learning system.
- Correlation between feature to feature and feature to class is calculated using Entropy, which helps in choosing significant datasets.

V. SYSTEM DESIGN

Design is a creative process; a good design is the key to effective system. The system Design is defined as "The process of applying various techniques and principles for the purpose of defining a process or a system in sufficient detail to permit its physical realization". Various design features are followed to develop the system. The design specification describes the features of the system, the components or elements of the system and their appearance to end-users.

System architecture is the conceptual design that defines the structure and behavior of a system. An architecture description is a formal description of a system, organized in a way that supports reasoning about the structural properties of the system. It defines the system components or building blocks and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system. The proposed architecture for this system is given below. It shows the way this system is designed and brief working of the system.



VI. IMPLEMENTATION

MODULES

1. Irrelevant Attribute Removal

The first point that we should discuss is how to remove irrelevant data from the training dataset. In this case, we use feature-class correlation to check the relevance level of feature with respect to class. Entropy based calculation which comes under Information gain in Machine Learning is used to find the correlation and every correlation is ranked based on a threshold value given as input. Hence the most relevant features are selected whose entropy value is more than a given threshold value.

2. Redundant Attribute Removal

In this module we introduce four efficient steps to remove redundant data out of given datasets. The correlation between feature-feature is calculated using Entropy calculator.

- First step works on Graph construction, where the nodes are the features and the edges carry a weight of their respective entropy values between those features.
- Second step works on creating minimum spanning tree using Prims algorithm.
- Third step works on partitioning the resulted graph into trees. Where the nodes with closer entropy values fall under same tree.
- Fourth step works on choosing a significant node from each constructed tree as a representation of its neighbor nodes. Hence duplicate data / redundant data is removed and most significant dataset is drawn at the end of this module.

3. Naive Bayes Classifier

In this module, we test the performance of proposed system using Naive Bayes Classifier.

CONCLUSION

This paper proposed a feature selection algorithm namely unsupervised learning with entropy based feature selection. The approach achieves better prediction accuracy than the other feature selection algorithms and achieves higher accuracy for IB1 and NB classifiers compared to the other feature selection algorithms. The Speedy Feature Selection method is also considerably good in reducing the time to build model for IB1 compared to FSGai-ra and ReliefF. It reduces the number of features compared to FSCor, FSCon and FSGai-ra. In future, this work can be extended with other statistical measures for ranking the features within the clusters.

REFERENCES

[1] J. Sinno, Q. Y. Pan. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–135, 2010.

[2] M. R. Rashedur, R. M. Fazle. Using and comparing diff erent decision tree classification techniques for mining ICDDR, B Hospital Surveillance data. *Expert Systems with Applica-tions*, vol. 38, no. 9, pp. 11421–11436, 2011.

[3] M. Wasikowski, X. W. Chen. Combating the small sam-ple class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1388–1400, 2010.

[4] Q. B. Song, J. J. Ni, G. T. Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineer-ing*, vol. l5, no. 1, pp. 1–14, 2013.

[5] J. F. Artur, A. T. M´ario. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1794–1804, 2012.

[6] J. Wu, L. Chen, Y. P. Feng, Z. B. Zheng, M. C. Zhou, Z. H. Wu. Predicting quality of service for selection by neighborhood-based collaborative filtering. *IEEE Transac-tions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 2, pp. 428–439, 2012.

[7] C. P. Hou, F. P. Nie, X. Li, D. Yi, Y. Wu. Joint embedding learning and sparse regression: A framework for unsuper-vised feature selection. *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793–804, 2014.

[8] P. Bermejo, L. dela Ossa, J. A. G´amez, J. M. Puerta. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking Original Research

