An Integrated Approach for feature selection

[1]C. Sudhakar Reddy, [2]Dr K.venugopal Rao

Associate Professor, Professor,

[1]Dept of Information Technology,

G.Narayanamma Institute of Technology & science, Hyderabads, India.

Abstract: In the current digital world huge voluminous data is getting generated and stored by various stake holder of the system. To improve the performance organizations started utilizing this data for analysis for better decision making and improve the performance. Data mining/machine learning techniques are most widely used techniques for analyzing the data. As the real world data is very huge and may attributes are included it has become mandatory to apply some feature selection algorithm for dimensionality reduction in order to identify the features relevant for given type of analysis and use of feature selection algorithm generates a model with better accuracy and reduced learning time and only such features are used for analysis and so that the model will be created with high accuracy and less learning time. In this paper we propose a measure called relative dependency for identifying redundant attributes. Comparative performance of various classification algorithms is illustrated in the paper on various data set collected from UCI repository.

Keywords: Feature Selection, Attribute Dissimilarity.

Introduction

Feature selection is defined as the process of identifying the most relative attributes from the given set of features. Wrapper method and filter method are the two most widely used feature selection techniques. In case of Wrapper method respective classifier itself chooses a measure for identifying relevant attributes where as in case of filter method approach feature selection will be performed first irrespective of the Classification algorithm then selected features will be used by classification algorithm. Both these methods performs exhaustive search for identifying features relevant for classification which may be time consuming for high dimensional data.

Proposed Method

In this paper we propose an integrated method for feature selection .An integrated approach uses the combination of wrapper method and the measure called relative dependency together to identify the features relevant for the given classification. The procedure for integrated approach is shown in the figure 1

Input: Dataset
Output: Selected List of Features
Steps

1. On the given data set apply any wrapper method to produce a set of attributes A_1
2. Use the given data set and generate relative dependency matrix.
3. Use K-means Clustering Algorithm to form the clusters using matrix constructed in step 2
4. From the set A_1 reduce the no of attributes based on clusters formed to generate Attribute set A_2.
5. Set A_2 represents set of relevant features

figure 1

Dependency between any two attributes namely A_i and A_j is calculated using attribute dissimilarity.

Attribute Dissimilarity

Dissimilarity between the two attributes is calculated as fallows

Given two attributes A_i and A_j then dependency between A_i and A_j is represented as Dep (A_i, A_j)

$$\mathrm{Dep}(A_i, A_j) = \frac{\pi_{A_i}(R)}{\pi_{A_i, A_j}(R)}$$

$\Pi_A(R)$ indicates projection of attribute A over the relation R.

As the $Dep(A_i,A_j)$ is not symmetric we calculate the the Dependency as average of $Dep(A_i,A_j)$ and $Dep(A_j,A_i)$

The distance (dissimilarity) measure for the pair of attributes Ai and Aj is thus proposed as follows

$$Dep(A_i,A_j) = \frac{1}{Avg(Dep(A_i,A_j),Dep(A_j,A_i))}$$

Dissimilarity between various attributes is calculated and represented as a matrix.

After generating the dissimilarity matrix clusters are constructed using simple k means. Membership of the attributes is used to reduce the attributes further in order to increase the accuracy and decrease the learning time.

Example

	inter	btech	tec ev	nteecev	comm	Placed
A	IB	EB	TOK	NOK	GOOD	YES
B	IA	EB	TOK	NBEST	BAD	YES
a	IA	EB	TOK	NOK	OK	YES
A	Dist	EA	TGOOD	NBEST	OK	NO
A	Dist	EA	TGOOD	NGOOD	OK	YES
A	Dist	EA	TGOOD	NGOOD	BAD	YES
A	Dist	EB	TOK	NBEST	OK	YES
A	Dist	EB	TGOOD	NBEST	OK	YES
A	Dist	EB	TOK	NOK	GOOD	YES
A	Dist	EC	TOK	NGOOD	GOOD	NO
B	IA	EC	TOK	NGOOD	BAD	NO
A	Dist	EC	TOK	NBEST	OK	NO

Table 1 (Relation R)

$\Pi_{\text{Placed}}(R)=2$

$\Pi_{\text{Btech,}}(R)=3$

$\Pi_{\text{Btech,Palced}}(R)=4$

so Dependency (B,Tech, Placed)=3/4=0.75

Experimental Results:

To carry out the experiment we collected data sets from UCI repositories and kaggle. The data sets size is varying from few hundreds to thousands. Comparative performance of J48 and Random Forest on various data sets is shown in the table.

Data file	Source	No of Instances	No of attributes	Algorithms	Accuracy	time taken to build model
xmap-edu-data	UCI	480	17	J48	75.83	0.1
				j48(with RD)	76.25	0.02
				RandomForest	76.6	0.33
				RandomForest (With RD)	74.19	0.16
Placementdatafullclass	kaggle	215	15	J48	82.79	0.02
				j48(with RD)	80	0
				RandomForest	85.18	0.23
				RandomForest (With RD)	83	0.04
Student performance data	kaggle	30	20	zeroR	40	0
				zero(with RD)	40	0
Network Intrusion Detection System	kaggle	25000	41	J48	99.5	2.56
				j48(with RD)	99.36	3.6

				RandomForest	99.78	12
				RandomForest (With RD)	99.78	10

Table 2

Conclusion

We performed experiments on data sets of different sizes using the integrated approach by combining wrapper methods with relative dependency between the attributes using classification algorithms J48, Random Forest. The performance of the integrated approach is providing improved results when compared with the results obtained by using cfssubsetselection feature selection, information gain ranking feature selection methods especially learning time is reduced.

References

1.Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni, Marcello Restelli, "Feature Selection via Mutual Information: New Theoretical Insights" , arXiv:1907.07384v1 Jul 2019.

2. Bhavesh Patankar, Dr. Vijay Chavda, "Improving Classifier Performance Using Feature Selection with Ensemble Learning", 2016 IJSRCSEIT | Volume 1 | Issue 1 | ISSN : 2456-3307.

3.Tzung-Pei Hong ·,Yan-Liang Liou , Shyue-Liang Wang , Bay Vo " Feature selection and replacement by clustering attributes" , Vietnam J Computer Science (2014) , DOI 10.1007/s40595.

4. Samuel H. Huang, "Supervised feature selection: A tutorial" Artificial Intelligence Research, 2015, Vol. 4, No. 2, ISSN 1927-6974.

5.N. Hoque, D. K. Bhattacharyya, J. K. Kalita, "A Mutual Information-based Feature Selection Method", Expert systems with applications (Elsevier) April 21, 2014.

6. Jianchao Han, Ricardo Sanchez, Xiaohua Hu, T.Y. Lin, "Selection Based on Relative Attribute Dependency: An Experimental Study", 10th International Conference, RSFDGrC 2005, DOI: 10.1007/11548669_23.

7. Saurabh Pal , "Classification Model of Prediction for Placement of Students", DOI: 10.5815/ijmecs.2013.11.07.

8. Dianhong Wang, Liangxiao Jiang , "An Improved Attribute Selection Measure for Decision Tree Induction" IEEE 2007 (FSKD 2009).

9. https://archive.ics.uci.edu/ml/index.php.

10. https://www.kaggle.com/data.