Journal of Engineering Science and Technology
Vol. 11, No. 11 (2016) 1639 – 1646

REVIEW ON FEATURE SELECTION TECHNIQUES AND ITS IMPACT FOR EFFECTIVE DATA CLASSIFICATION USING UCI MACHINE LEARNING REPOSITORY DATASET

AMARNATH B.[1,*], S. APPAVU ALIAS BALAMURUGAN[2]

[1]Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India
[2]Department of I.T, K.L.N College of Information Technology, Tamilnadu, India
*Corresponding Author: amars_88@yahoo.co.in

Abstract

Feature selection goal is to get rid of redundant and irrelevant features. The problem of feature subset selection is that of finding a subset of the original features of a dataset, such that an induction algorithm run on data containing only selected features makes a classifier to generate with the highest possible accuracy. High dimensional data can contain a high degree of irrelevant and redundant features which may greatly degrade the performance of learning algorithms. The performance of different feature selectors such as CFS, Chi-Square, Information Gain, Gain Ratio, One R and Symmetrical Uncertainty were evaluated on two different popular classification algorithms namely Decision Tree and Naive Bayesian method. A significant improvement in the performance of DT and NB classifier was shown after reducing the number of both irrelevant and redundant features by the use of different feature ranking methods.

Keywords: Data mining, Feature selection, Classification.

1. Introduction

Selection of most relevant features by eliminating irrelevant and redundant features is known as feature selection. Feature subset selection and Feature ranking methods are two broad classifications of feature selection. Wrapper, embedded and hybrid methods are further classifications in feature selection with respected to different supervised learning algorithm used for the experimentation. All the three methods use supervised learning algorithm thereby they are computationally expensive and have lesser generality than others.

The feature ranking and subset based feature selection methods are analysed

Nomenclatures

C_i	Class
I	Information before the split
I_i	Information of node i
$P\ (C_i)$	Probability of any object belonging to class C_i
$P\ (X)$	Probability of obtaining attributes value X
S	Object and Training data
t_i	Number of objects in node i
X	Object

Abbreviations

ACO	Ant Colony Optimization
ANNs	Artificial Neural Networks
BPN	Back Propagation Network
CFS	Correlation Based Feature Selection
C4.5	Classification Tree 4.5
DT	Decision Tree
FS	Feature Selection
GA	Genetic Algorithm
GR	Gain Ratio
ID3	Iterative Dichotomiser
IG	Information Gain
KNN	K-Nearest Neighbour
NB	Naive Bayesian
PSO	Particle Swarm Optimization
SA	Simulated Annealing
SU	Symmetrical Uncertainty
TS	Tabu Searching
UCI	University of California, Irvine
WEKA	Waikato Environment for Knowledge Analysis

in terms of classification accuracy in this paper. In order to choose suitable feature subset selection or ranking method for redundancy analysis, a pragmatic investigation on different feature subset selection and feature ranking methods is carried out by analysing their performance in terms of accuracy and time taken to build a model.

Several feature selection methods that are proposed by various researchers. Wattana and Nachirat [1] proposed a comparative study of feature selection techniques for classify student performance. Aldy et al. [2] compared the rule based method and statistical based method on emotion classification for Indonesian twitter text. Raoof and Ali [3] proposed LDA-based discrimination of left and right hand motor imagery: Outperforming the winner of BCI Competition II.

Ashin et al. [4] proposed identification of gene signatures for classifying of breast cancer subtypes using protein interaction database and support vector machines. Yan et al. [5] proposed unsupervised discovery of subspace trends. Yutong et al. [6] proposed study of test classification algorithm based on domain knowledge. Seyyed et al. [7] proposed automatic MRI image threshold using fuzzy support vector machines. Hossein et al. [8] proposed low-rate false alarm intrustion detection system with genetic algorithm approach. Parth et al. [9]

proposed application of data mining in fault diagnosis of induction motor. Yuan et al. [10] proposed attribute reduction for Chinese question classification.

The proposed method is described in Section 3 and in Section 4 the results are discussed. The conclusion is given in Section 5.

2. Problem Statement

The accuracy of classifier highly depends on the training dataset which is used to learn and develop the classifier. Developing the classifier with high dimensional training dataset takes more time to build the classifier and reduces its classification accuracy since the high dimensional data contains irrelevant and redundant features. The irrelevant features contain similar information both of them can lead to misclassification. Therefore, the feature ranking and subset selection methods are analyzed in terms of classification accuracy and time taken to build the model.

3. Comparative Analysis on Feature Selection Methods

The empirical study on different feature subset selection and ranking methods using the real world dataset from UCI machine learning repository is the main focus of this research work. The performance of different feature subset and ranking methods including CFS, Chi-Square, GR, IG, One R and SU on Contact lens, Shuttle landing, DNA promoter, Tic Toc Toe, Parity, Nursery, Adult, Chess, Monk, Weather, Splice, Spect heart, King-Rook vs King-Pawn, Car-Evaluation and Balloon is thoroughly carried out in terms of number features selected, time taken to reduce the model generation and detection performance of classifiers.

Filter approach can able to deals high dimensional data more effectively than wrappers and having less computational complexity as shown in Fig. 1. Supervised learning algorithms are mainly used in wrapper methods for the purpose of validating the selected feature subsets as shown in Fig. 2.

Hybrid method is a combination of wrapper and filter approach and it produces better detection performance than filter and wrapper approaches and computationally expensive than the filter and cheaper than the wrapper method.

The different combinations of feature subsets by the use of various searching approaches are the main focus of feature subset selection approaches. This approaches usually, having higher computational complexity due to the generation of various possible subsets.

Ranking various features by the use of different rankers such as CFS, Chi-Square, GR, IG, One R and SU was empirically done with respected to feature selection criteria and finally, top ranked features are considered as best feature than others

The performance assessment of the feature subset based feature selection methods namely correlation based and consistency based feature subset selection, feature ranker based feature selection namely Chi-Square based feature selection and IG based feature selection. The performance of these methods is analyzed with the focus on reducing the number of features and improving the accuracy of classification.

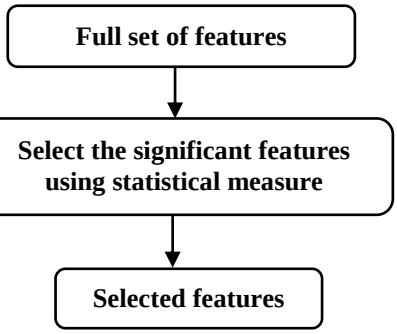

Fig. 1.Filter approach for feature selection.

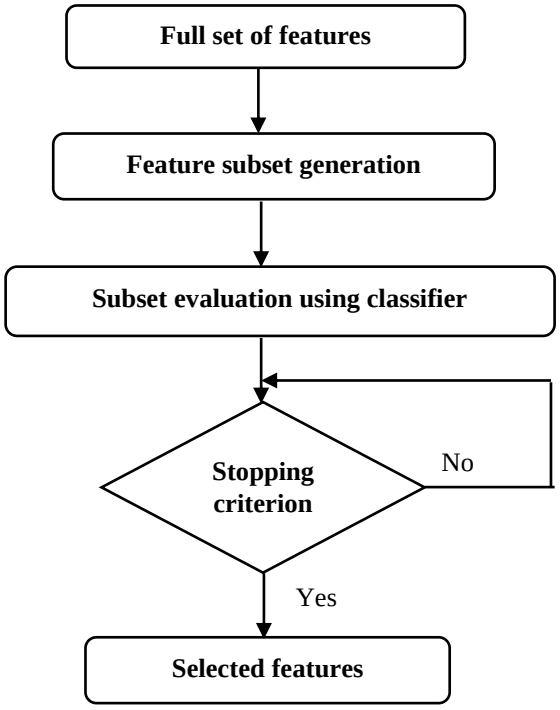

Fig. 2. Wrapper approach for feature selection.

4. Experimental Results and Discussion

In order to conduct the experimental study, the datasets from various domains are collected from WEKA dataset repository and UCI machine learning repository. Further, the experiment is carried out with the WEKA data mining tool. In this experiment, both feature subset selection and feature ranking method performance is evaluated using the classification algorithm namely NB and DT. In the DT, the amount of information is computed using

$$I = -(n/s)\log(n/s) - (p/s)\log(p/s) \quad (1)$$

Information Gain of each attribute is computed using

$$Gain(S, A) = I - \sum_{i\varepsilon \text{ values(A)}} (t_i / s) I_i \quad (2)$$

Likewise in the NB classifier the conditional probability is computed by Bayes theorem.

$$P(C_i / X) = [P(X) / (C_i) P(C_i)] / P(X) \quad (3)$$

The number of selected features by the feature subset selection and feature ranking method is given in Table 1.The accuracy of NB and DT classifiers for different feature selection methods on the dataset are presented in Table 2.

Feature subset selection and Feature ranking method namely CFS subset evaluation, Chi-Square attribute evaluation, GR, IG, One R attribute evaluation and SU attribute evaluation respectively are applied on an original dataset and the running time and the number of selected features for each algorithm are recorded. The classification algorithm namely ID3, C4.5 and NB are applied in the original dataset with original features (See Table 2) as well as the datasets with reduced features (See Tables 3 to 5) and the overall performance is recorded.

Based on the performance analysis of different feature selectors and classifiers from Tables 1 to table 5, the following observations are made

- Symmetrical Uncertainty feature selector significantly reduced the number of initial features comparing with all other feature selection methods in the literature.
- Gain Ratio is best feature selector for the NB classifier than other feature selectors in the literature.
- When Information Gain is used as a feature selector the performance the ID3 classifier will get improved.
- C4.5 classifier with One R attribute evaluator as a feature selector will provide the better classification accuracy.

Table 1. Number of feature selected by each feature selection algorithm.

Datasets	Instances	Att.	CFS	Chi Square	GR	IG	One R	SU
C lens	24	5	1	2	2	2	3	1
S landing	15	7	2	6	3	5	4	4
DNA P	106	58	6	6	6	6	5	5
TTT	958	10	5	1	1	1	1	1
Parity	100	11	3	6	6	6	6	6
Nursery	12960	9	1	1	1	1	5	5
Adult	20	5	2	2	2	2	2	2
Chess	2128	37	6	7	5	7	5	9
Monk	124	6	2	2	2	2	2	2
Weather	14	5	2	2	2	2	3	2
Splice	3190	61	22	7	8	8	9	2
S heart	267	23	12	8	10	9	16	9
KR vs. KP	3196	37	7	11	15	11	19	10
Car-Eval.	1728	7	1	6	6	6	6	6
Balloon	20	5	2	2	2	2	2	2

Table 2. Accuracy of classifiers on full feature set.

Datasets	NB	ID3	J48
Contact lens	70.83	70.83	83.33
Shuttle landing	80.00	60.00	53.33
DNA promoter	90.57	76.42	81.13
Tic Toc Toe	69.62	85.07	83.40
Parity	40.00	45.00	44.00
Nursery	90.32	98.18	97.05
Adult	100.00	100.00	100.00
Chess	89.94	99.20	98.91
Monk	99.36	89.53	94.36
Weather	57.14	85.71	50.00
Splice	95.36	89.53	94.36
Spect heart	79.03	70.04	80.90
King-Rook vs. King-Pawn	87.89	99.68	99.40
Car-Evaluation	85.53	89.35	92.36
Balloon	100.00	100.00	100.00
Mean	**82.37**	**83.90**	**73.87**

Table 3. Accuracy of NB on selected features for each feature selection algorithm.

Dataset	CFS	Chi Square	GR	IG	One R	SU
C lens	70.83	87.50	87.50	87.50	54.17	70.83
S landing	80.00	73.33	80.00	73.33	73.33	73.33
DNA P	95.28	95.28	95.28	95.28	95.28	95.34
TTT	72.44	69.94	69.94	69.94	69.94	69.94
Parity	50.00	46.00	46.00	46.00	47.00	46.00
Nursery	70.97	70.97	70.97	70.97	88.84	70.97
Adult	100.00	100.00	100.00	100.00	100.00	100.00
Chess	94.45	89.61	92.34	89.61	86.33	90.23
Monk	100.00	100.00	100.00	100.00	100.00	100.00
Weather	78.57	78.57	78.57	78.57	71.43	78.57
Splice	96.14	93.89	94.17	94.17	94.29	94.17
S heart	82.02	76.78	80.15	79.03	79.03	79.03
KR vs. KP	91.99	88.17	89.86	89.11	88.11	88.67
Car-Eval.	70.02	85.53	85.53	85.53	85.53	85.53
Balloon	100.00	100.00	100.00	100.00	100.00	100.00
Mean	**83.51**	**83.70**	**84.68**	**83.93**	**82.21**	**82.84**

Table 4. Accuracy of ID3 on selected features for feature selection algorithm.

Datasets	CFS	Chi Square	GR	IG	One R	SU
C lens	70.83	87.50	87.50	87.50	50.00	70.83
S landing	46.67	60.00	66.67	66.67	66.67	66.67
DNA P	84.91	84.91	84.91	84.91	84.91	84.91
TTT	82.78	69.94	69.93	69.94	69.94	69.94
Parity	53.00	53.00	53.0000	53.00	48.00	53.00
Nursery	70.97	70.97	70.9722	70.97	91.70	70.97
Adult	100.00	100.00	100.0000	100.00	100.00	100.00
Chess	94.36	94.36	92.3402	94.36	90.60	94.36
Monk	95.97	95.97	95.9677	95.97	95.97	95.97
Weather	78.57	78.57	78.5714	78.57	57.14	78.57
Splice	90.66	90.60	90.3135	90.31	88.87	90.31
S heart	81.65	79.40	75.6554	75.66	79.03	75.66
KR vs. KP	94.24	96.09	94.7434	94.34	96.18	94.24
Car-Eval.	70.02	89.35	89.3519	89.35	89.35	89.35
Balloon	100.00	100.00	100.000	100.00	100.00	100.00
Mean	**80.97**	**83.37**	**83.32**	**83.43**	**80.55**	**82.31**

Table 5. Accuracy of C4.5 on selected features for feature selection algorithm.

Datasets	CFS	Chi Square	GR	IG	One R	SU
C lens	70.83	87.50	87.50	87.50	58.33	70.83
S landing	53.33	53.33	60.00	53.33	60.00	53.33
DNA P	83.02	83.02	83.02	83.02	83.02	83.96
TTT	79.44	69.94	69.94	69.94	69.94	69.94
Parity	44.00	40.00	40.00	40.00	50.00	40.00
Nursery	70.97	70.97	70.97	70.97	90.74	70.97
Adult	100.00	100.00	100.00	100.00	100.00	100.00
Chess	94.31	94.36	92.34	94.36	90.60	94.31
Monk	91.94	91.94	91.94	91.94	91.94	91.94
Weather	42.86	42.86	42.86	42.86	50.00	42.86
Splice	94.48	93.54	94.01	94.01	93.98	94.01
Spect heart	81.65	79.40	75.66	75.66	79.03	75.66
KR vs KP	94.06	96.50	94.71	94.49	96.81	94.06
Car-Eval.	70.02	92.36	92.36	92.36	92.36	92.36
Balloon	100.00	100.00	100.00	100.00	100.00	100.00
Mean	**78.06**	**79.71**	**79.68**	**79.36**	**80.44**	**78.28**

4. Conclusion

The subset-based feature selection uses search technique for subset generation and induces additional computation cost and space complexity. Therefore, selecting the features from the high-dimensional data with the traditional approaches is crucial in terms of memory space, computation complexity and

classification accuracy. The ranker-based feature selection method weights each feature by any one of the statistical measures without the help of any supervised learning algorithm. Then the features are ranked based on their weight value. Therefore, it is computationally cheaper and maintains high generality since it does not employ any supervised learning algorithm. It takes very less space since the weight of the individual feature is calculated based on the relation between the particular feature and the target class attribute only. Therefore, the ranker method is suitable for high-dimensional data. However, the ranker method selects only the relevant features. It does not deal with the redundancy analysis. It's found that NB classifier with a gain ratio as a feature selection method will provide better classification accuracy than all other methods.

References

1. Wattana, P.; and Nachirat, R. (2105). A comparative study of feature selection techniques for classify student performance. *7th International Conference on Information Technology and Electrical Engineering*, 425-429.
2. Aldy, R.A.; and Ayu, P. (2015). Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text. *International Conference on Information Technology Systems and Innovation*, 1-6.
3. Raoof, M.; and Ali, K. (2015). Enhancing LDA-based discrimination of left and right hand motor imagery: Outperforming the winner of BCI Competition II. *2nd International Conference on Knowledge-Based Engineering and Innovation*. 392-398.
4. Ashin, G.; Mohammad, R.S.; Alireza, V.; and Mohammad, R.M. (2015). Identification of gene signatures for classifying of breast cancer subtypes using protein interaction database and support vector machines. *5th International Conference on Computer and Knowledge Engineering*. 195-200.
5. Yan, X.; Peng, Q.; and Badrinath, R. (2015). Unsupervised discovery of subspace trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10), 2131-2145.
6. Yutong, G.; Feifan, S.; Xiaqing, X.; Qing, S.; and Xu, W. (2015). Study of test classification algorithm based on domain knowledge. *Third International Conference on Cyberspace Technology*, 1-5.
7. Seyyed, M.H.; Javad, V.; and Seyyedeh M.H. (2015). Automatic MRI image threshold using fuzzy support vector machines. *2nd International Conference on Knowledge-Based Engineering and Innovation*, 207-210.
8. Hossein, A.; Ali, A.K.; and Mahdi, G.V. (2015). Low-rate false alarm intrustion detection system with genetic algorithm approach. *2nd International Conference on Knowledge-Based Engineering and Innovation*, 1045-1048.
9. Parth, S.P.; Pratyay, K.; and Paramita, C. (2016). Application of data mining in fault diagnosis of induction motor. *IEEE First International Conference on Control, Measurement and Instrumentation*, 274-278.
10. Yuan, L.; Su, L.; and Shu, Peng. (2016). Attribute reduction for Chinese question classification. *Chinese Control and Decision Conference*, 5488-5492.