An Empirical Study on feature selection for Data Classification

S.Rajarajeswari[1] , K.Somasundaram[2]

Department of Computer Science ,M.S.Ramaiah Institute of Technology,Bangalore,India.[1]
Department of Computer Science ,Jaya Engineering College, Chennai, India. [2]
gouthamraji8@gmail.com[1]
soms72@yahoo.com[2]

Abstract

In the task of pattern classification features play a very important role. Hence, the selection of suitable features is necessary as most of the raw data might be redundant or irrelevant to the recognition of patterns. In some cases, the classifier cannot perform well because of the large number of redundant features. This paper investigates the performance of different feature selection algorithms for the task of data classification. Different features play different roles in classifying datasets. Unwanted features will result in error information during classification which will reduce classification precision. The most of traditional feature selection can remove these distractions to improve classification performance. As shown in the experimental results, after feature selection using the traditional methods to control false discovery rate, the classification performance of DT's and NB classifiers were significantly improved.

Keywords

Data mining, Feature Selection, Classification, Support Vector Machine

1. Introduction

Feature selection is one of the important and frequently used techniques in data pre-processing for data mining. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications, speeding up a data mining algorithms and improving the mining performance. Selecting the right set of features for classification is one of the most important problems in designing a good classifier [7]. Very often we don't know a-priori what the relevant features are for a particular classification tasks. One popular approach to address this issue is to collect as many features as we can prior to the learning and data-modeling phase. However, irrelevant or correlated features, if present, may degrade the performance of the classifier. In the emerging area of data mining applications, users of data mining tools are faced with the problem of datasets that are comprised of large number of features and instances. Such kinds of datasets are not easy to handle for mining. The mining process can be made easier to perform by focusing on a set of relevant features while ignoring the other ones.

In this paper, we present our study on features subset selection and classification with the DT's and NB algorithm. In Section 2, we briefly describe the survey on feature selection methods. The Problem statement is given in Section 3.The Proposed method is described in Section 4. In Section 5, we describe our results. The conclusion is given in Section 6.

2. Literature Review

In general, feature selection techniques can be split into two categories - filter methods [3] and wrapper methods [8]. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, they are generally far too expensive to be used if the number of features is large because each feature set considered must be evaluated with the trained classifier. Filter methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets. Diverse feature ranking and feature selection techniques have been proposed in the machine learning literature, Such as: Correlation- based Feature Selection [6], Principal Component Analysis [6], Information Gain attribute evaluation [6], Gain Ratio attribute evaluation [6], Chi-Square Feature Evaluation [6] and Support Vector Machine feature elimination [5]. Some of these methods does not perform feature selection but only feature ranking, they are usually combined with another method when one needs to find out the appropriate number of features. Forward selection, backward elimination, bi-directional search, best-first search [12], genetic search [4], and other methods are often used on this task. The most often used criteria for feature selection is information theoretic based such as the Shannon entropy measure *I* for a dataset. The main drawback of the entropy measure is its sensitivity to the number of attribute values [11].Therefore C4.5 uses gain ratio. However, this measure suffers the drawback that it may choose attributes with very low information content [9].A comprehensive discussion on Bayes theorem for feature selection is available in [1].

3. Problem statement

Feature selection is the process of removing features from the dataset that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy.This paper presents an empirical study on different features methods for the task of data classification. Firstly, the dependency information among features is identified. In addition to that, feature selection has been improvised by ranking methods which selects features based on their importance. The objective is to improve the classification accuracy such that prediction of the class variable is improved over that of the original data with initial attribute set and also reduces the computational time.

4. Proposed method

We are concerned with the problem of feature selection. The main idea provided is to find out the dependent features and remove the redundant ones among them. The technology to obtain the dependency needed is based on traditional feature selection methods. The purpose of this study is to reduce the computational complexity and increase the classification accuracy of the selected feature subsets.. A traditional feature is implemented and evaluated through extensive experiments, comparing with traditional feature selection algorithms over fifteen datasets from UCI machine learning repository databases [2].

5. Experimental results and Discussion

The traditional feature selection methods are applied to many datasets, and the performance evaluation is done using a software package called WEKA [10]. We presented the performance evaluation on fifteen dataset such as Contact lenses, Shuttle landing, DNAPrometer, Tic-tac-toe, Parity, Nursery, Adult, Chess, Monk, Weather, Splice, Spec heart, King-Rook vs. King-Pawn, Car-evaluation and Balloon. All these datasets are recommended by UCI repository databases [2]. A summary of dataset is presented in Table 2.

For each dataset, we run several feature selection methods such as CFS subset evaluation, Chi-square attribute evaluation, Gain ratio, Information Gain, one attribute evaluation and symmetrical uncertain attribute evaluation respectively, and record the running time and the number of selected features for each algorithm. We then apply ID3, C4.5 and NB on the original dataset (See Table 3) as well as each newly obtained dataset containing only the selected features from each algorithm (See Table 4 and 8) and recorded the overall accuracy using 10 fold cross-validation (See Table 5-8).A traditional feature selection method is implemented and evaluated through extensive experiments comparing with related feature selection algorithms. Our findings can be summarized as follows:

(i) We have studied a performance of each traditional feature selection method and found that it will improve the performance of classifiers such as NB, C4.5 and ID3.

(ii)Improved NB classifier performance from 82.37 to 84.68%, where 84.68 indicate the performance using Gain ratio.

(iii) Improved C4.5 classifier performance from 73.87 to 80.44 %, where 80.44 indicates the performance using OneR attributes evaluation method.

(iv) Reduced the number of selected features when compared to the original features in the datasets such as Contact lenses, Nursery, Monk, Parity, Tic-Tac and Weather.

(v) Improved performance when compared to the performance of conventional algorithms, With C4.5 classifier in datasets such as Contact lenses, Shuttle, Tic-Tac, Nursery, Monk and Car. With ID3 classifier in the dataset such as Contact lenses, Tic-Tac, Nursery, Monk, Weather and Car. With NB classifier in the dataset such as Contact lenses,Shuttle,Tic-Tac,Nursery,Monk,Weather and Car.

Table 1.A Sample Dataset used in the experiment

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Over	Mild	High	False	Yes
Rainy	Cool	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	Yes
Over	Cool	Normal	True	Yes
Sunny	Hot	High	False	No
Sunny	Hot	Normal	False	No
Rainy	Cool	Normal	False	Yes

Table 2.Details description of dataset used in the experiment

Datasets	Instances	Attributes
Contactlense	24	5
Shuttlelanding	15	7
DNAprometer	106	58
TicTocToe	958	10
Parity	100	11
Nursery	12960	9
Adult	20	5
Chess	2128	37
Monk	124	6
Weather	14	5
Splice	3190	61
Spectheart	267	23
King-Rook vs King-Pawn	3196	37
Car-evaluation	1728	7
Balloon	20	5

Table 3.Accuracy of Classifiers on full feature set

Datasets	NB	ID3	J48
Contactlense	70.83	70.83	83.33
Shuttlelanding	80.00	60.00	53.33
DNAprometer	90.57	76.42	81.13
TicTocToe	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
Spectheart	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 4.Number of selected features by each feature selection algorithm

Datasets	CFS subset eval	Chi-square attr eval	Gain ratio
Contactlense	1	2	2
Shuttlelanding	2	6	3
DNAprometer	6	6	6
TicTocToe	5	1	1
Parity	3	6	6
Nursery	1	1	1
Adult	2	2	2
Chess	6	7	5
Monk	2	2	2
Weather	2	2	2
Splice	22	7	8
Spectheart	12	8	10
King-Rook vs King-Pawn	7	11	15
Car-evaluation	1	6	6
Balloon	2	2	2

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

Datasets	**CFS subset eval**	**Chisquare attr eval**	**Gain ratio**	**Information gain**	**One attr eval**	**Symmetrical uncert attr eval**
Contactlense	70.83	87.50	87.50	87.50	54.17	70.83
Shuttlelanding	80.00	73.33	80.00	73.33	73.33	73.33
DNAprometer	95.28	95.28	95.28	95.28	95.28	95.34
TicTocToe	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
Spectheart	82.02	76.78	80.15	79.03	79.03	79.03
King-Rook vs King-Pawn	91.99	88.17	89.86	89.11	88.11	88.67
Car-evaluation	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 6.Accuracy of ID3 on selected features for each feature selection algorithms

Datasets	**CFS subset eval**	**Chisquare attr eval**	**Gain ratio**	**Information gain**	**One attr eval**	**Symmetrical uncert attr eval**
Contactlense	70.83	87.50	87.50	87.50	50.00	70.83
Shuttlelanding	46.67	60.00	66.67	66.67	66.67	66.67
DNAprometer	84.91	84.91	84.91	84.91	84.91	84.91
TicTocToe	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
Spectheart	81.65	79.40	75.6554	75.66	79.03	75.66
King-Rook vs King-Pawn	94.24	96.09	94.7434	94.34	96.18	94.24
Car-evaluation	70.02	89.35	89.3519	89.35	89.35	89.35
Balloon	100.00	100.00	100.0000	100.00	100.00	100.00
Mean	**80.97**	**83.37**	**83.32**	**83.43**	**80.55**	**82.31**

Table 7.Accuracy of C4.5 on selected features for each feature selection algorithms

Datasets	**CFS subset eval**	**Chisquare attr eval**	**Gain ratio**	**Information gain**	**One attr eval**	**Symmetrical uncert attr eval**
Contactlense	70.83	87.50	87.50	87.50	58.33	70.83
Shuttlelanding	53.33	53.33	60.00	53.33	60.00	53.33
DNAprometer	83.02	83.02	83.02	83.02	83.02	83.96
TicTocToe	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
Spectheart	81.65	79.40	75.66	75.66	79.03	75.66
King-Rook vs King-Pawn	94.06	96.50	94.71	94.49	96.81	94.06
Car-evaluation	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**

6. Conclusion

Selecting the right set of features for classification is one of the most important problems in designing a good classifier. Decision Tree induction algorithms such as C4.5 have incorporated in their learning phase an automatic feature selection strategy while some other statistical classification algorithm require the feature subset to be selected in a pre-processing phase. It is well known that correlated and irrelevant features may degrade the performance of the classification algorithms. In our study, we evaluated the influence of feature pre-selection on the predictive accuracy of DT's and NB classifiers using the real world dataset. We observed that the accuracy of the C4.5 and NB classifiers could be improved with an appropriate feature pre-selection phase for the learning algorithm. Beyond that, the number of features used for classification can also be reduced since feature selection is a time consuming process.

References

[1] Balamurugan S.A. and Rajaram R., "Effective and Efficient Feature Selection for Large Scale Data using Bayes Theorem," Journal of Automation and Computing, vol. 6,no. 1, pp. 62-71, 2009.

[2]Blake C.L and Merz C.J., "UCI Repository of Machine Learning Databases," http:// www.ics.uci.edu/~mlearn/mlrepository.html, 2008.

[3]Cover T.M., "On the possible ordering on the measurement selection problem,".IEEE Transactions, SMC, vol.7, no.9, pp.657-661, 1977.

[4]Goldberg D.E., Genetic algorithms in search, optimization and machine learning, Addison-Wesley,1989.5 Guyon I, Weston J, Barnhill S and Vapnik V.

[5]"Gene selection for cancer classification using supportvector machines," Machine Learning, vol.46, pp.389–422, 2002.

[6] Hall M.A and Smith L.A., "Practical feature subsetselection for machine learning", In Proceedings of the21st Australian Computer Science Conference, pp.181–191, 1998.

[7] Jiawei Han and Micheline Kamber, Data mining Concepts and Techniques, Morgan Kaufmann Publishers, 2006.

[8] Kohavi R and John G.H., The Wrapper approach.In,ed.Lui H and Matoda H ,Feature Extraction Construction and Selection, Kluweracademic publishers,pp.30-47,1998.

[9] Lopez de Mantaras R., "A Distance- based attribute selection measure for decision tree induction," Machine Learning, vol.6, pp.81-92, 1991.

[10] WEKA, "Open Source Collection of Machine Learning Algorithms,".

[11] White A.P and Lui W.Z., "Bias in the information-based measure in decision tree induction," Machine Learning, vol.15, pp.321-329, 1994.

[12] Witten H and Frank E., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2nd Edition, 2005.

[13] You-Shyang Chen, Classifying Credit Ratings For Asian Banks Using Integrating Feature Selection And CPDA – Based Rough Sets Approach, Knowledge Based Systems, 26, 259 – 270, August 2011

[14] Zhi- Gang Su, Pei- Hong Wang , Minimizing Neighborhood Evidential Decision Error For Feature Evaluation And Selection Based On Evidence Theory , 39 , 527 -540 ,2011.