Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 37A (2011) 4118-4122

ISSN: 2229-712X

A review of feature selection models for classification

B.Kalpana[1], V.Saravanan[2] and K.Vivekanandan[3]

[1]Bharathiar University, A.P, Department of Computer Science & Applns. PSG College of Arts and Science, Coimbatore- 641014

[2]Dr. NGP Institute of Technology, Kalapatti Road, Coimbatore - 641048

[3]School of Management, Bharathiar University, Coimbatore – 641046.

ARTICLE INFO

Article history:

Received: 28 May 2011;

Received in revised form:

22 July 2011;

Accepted: 30 July 2011;

Keywords

Data mining,

Feature selection,

Classification.

ABSTRACT

The success of a machine learning algorithm depends on quality of data .The data given for classification, should not contain irrelevant or redundant attributes. This invariably increases the processing time. The data set, selected for classification should contain the right attributes for accurate results. Feature selection is an essential data processing step, prior to applying a learning algorithm. Here we discuss some basic feature selection models and evaluation function. Experimental results are compared for individual datasets with filter and wrapper model.

Introduction

Feature Selection

There are many potential benefits of variable and feature selection facilitating data visualization and data understanding, reducing the dimensions and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance [1]. Even if resources are not an issue, we want to remove the columns that are not required because they might degrade the quality of discovered patterns, for the following reasons:

• Some columns are noisy or redundant. This noise makes it more difficult to discover meaningful patterns from the data;

• To discover quality patterns, most data mining algorithms require much larger training data set on high-dimensional data set. But the training data is very small in some data mining applications.[2]

Let Y be the original set of features, with cardinality n. Let d represent the desired number of features in the selected subset X, X Í Y. Let the feature selection criterion function for the set X be represented by J(X). Without any loss of generality, let us consider a higher value of J to indicate a better feature subset. Since we are maximizing J(X), one possible criterion function is (1 - pe), where pe denotes the probability of error. The use of probability of error as a criterion function makes feature selection dependent on the specific classifier used and the size of the training and test data sets. Formally, the problem of feature selection is to find a subset X Y such that |X| = d and

$$J(X) = \max_{Z \subseteq Y, |Z|=d} J(Z).$$

Feature selection is mainly based on relevance. John, Kohavi and Pfleger define two notations of relevance [3]

Weak Relevance: An attribute xi is weakly relevant if not strongly relevant and there exists a subset of variables V such that the performance on V [{xi} is better than the performance on V .

Strong Relevance: An attribute xi is strongly relevant if its removal yields a deterioration of the performance of the Bayes Optimum Classifier.

General Characteristics of Feature Selection

The starting point of a feature space: Here initially the dataset has no dimensions and by forward search the dimensions are added. This is called Forward selection. Alternatively, initially the database may contain n dimensions and we can reduce dimensions by backward selection.

The search strategy: A complete search can be exhaustive by searching all 2n combinations or finding a minimum set which acts as a optimal set using branch and bound [14] or beam search [15] can be employed. A sequential strategy is based on hill climbing approach which can sequential forward selection, sequential backward elimination, and bidirectional elimination. A random search starts with random subsets and further search is done by sequential strategy or generate next subset in a random manner.

The evaluation criteria: This can be dependency, distance measure, information gain, probability measure with which the search is steered forward.

Stopping criteria: We can stop the search if there is no distinction between previous subset space and currently chosen one or desired no iterations or dimensions have been reached or we reached a good feature subset.

Available models for feature selection

Existing feature selection methods for machine learning typically fall into three broad categories—those which evaluate the worth of features using the learning algorithm that is to ultimately be applied to the data, and those which evaluate the worth of features by using heuristics based on general characteristics of the data. The former are referred to as wrappers and the latter filters [8], [9].There are three models available for feature selection.

Filter model

This utilizes an independent search criterion to find the appropriate feature subset before a machine learning algorithm is performed, thus it was termed as filter method by John, Kohavi and Pfleger. The advantage of this algorithm is, it need not run the induction algorithm every time an attribute is tested

Tele:

E-mail addresses: vivekbsmed@gmail.com

for relevancy. The algorithm shows high time efficiency. The disadvantage is that it totally ignores the effects of selected feature subset on the performance of the induction algorithm.The generalized filter model algorithm is given [10]

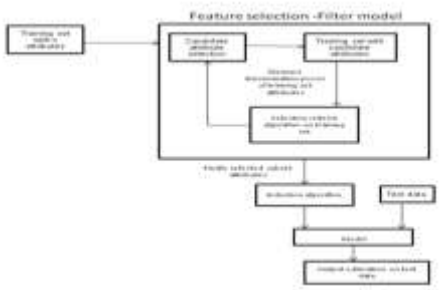

Fig a: Filter model flowchat

Filter Algorithm

input: $D(F_0, F_1, ..., F_{n-1})$ // a training data set with N features
S_0 // a subset from which to start the search
δ // a stopping criterion
output: S_{best} // an optimal subset

```
01 begin
02    initialize: S_best = S_0;
03    γ_best = eval(S_0, D, M); // evaluate S_0 by an independent measure M
04    do begin
05       S = generate(D); // generate a subset for evaluation
06       γ = eval(S, D, M); // evaluate the current subset S by M
07       if (γ is better than γ_best)
08          γ_best = γ;
09          S_best = S;
10    end until (δ is reached);
11    return S_best;
12 end;
```

Fig b: Filter model algorithm

Wrapper model

The strategy of the wrapper model is to use an induction algorithm to estimate the merit of the searched feature subset on the training data and using the estimated accuracy of the resulting classifier as its metric. The wrapper approaches often have better results than the filter approaches because they are tuned to the specific interaction between an induction algorithm and its training data. The wrapper approaches thus take into account the final induction algorithm [11]. Wrapper algorithm as given in [12] is shown in figure d.

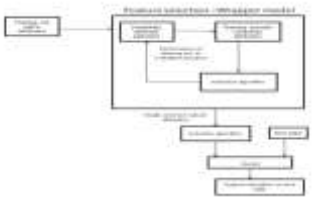

Fig c: Wrapper model flowchart

Wrapper Algorithm

input: $D(F_0, F_1, ..., F_{n-1})$ // a training data set with N features
S_0 // a subset from which to start the search
δ // a stopping criterion
output: S_{best} // an optimal subset

```
01 begin
02    initialize: S_best = S_0;
03    γ_best = eval(S_0, D, A); // evaluate S_0 by a mining algorithm A
04    do begin
05       S = generate(D); // generate a subset for evaluation
06       γ = eval(S, D, A); // evaluate the current subset S by A
07       if (γ is better than γ_best)
08          γ_best = γ;
09          S_best = S;
10    end until (δ is reached);
11    return S_best;
12 end;
```

Fig d: Wrapper algorithm

Hybrid model

The hybrid model does not have a prespecified stopping criterion. A typical hybrid method makes use of both an independent measure and a mining algorithm to select the final best subset among the best subsets across different cardinalities. With an initial empty subset S0 the algorithm tries to add a new subset in each iteration. The current subset with cardinality c, is incremented by searching the subset space c+1 and a new subset S is formed. It is evaluated by an independent measure M and compared with previous best .If S is best it becomes current best subset at level and becomes S'best

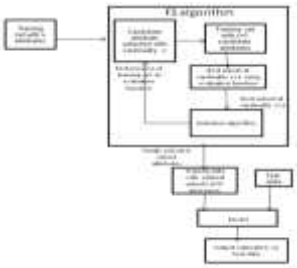

Fig e: Hybrid selection flowchart

At the end of each iteration a mining algorithm A, is applied to s'best and the result ɵ, of the mining algorithm is compared with the result of best subset at level c. If s'best is better, the algorithm tries to find the best subset for next level. Otherwise, the model stops and returns the current best subset as final best subset. The quality of mining algorithm becomes a natural stopping criterion for this model.

Fig f: Hybrid algorithm

Some Evaluation Measures for FS models

Within filter and wrapper categories, algorithms can be further differentiated by the exact nature of their evaluation function, and by how the space of feature subsets is explored. Filter algorithms can be performed on univariate or multivariate attributes. Wrapper algorithms mainly depend on consistency and dependency measures. General measures of evaluation criteria are discussed below.

Chi-square test: Chi-Square ($\chi 2$): based on the statistical theory. It measures the lack of independence between the terms in the category as shown in the equation

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

The primary advantage of the chi square goodness of fit test is that it is quite general. It can be applied for any distribution, either discrete or continuous, for which the cumulative distribution function can be computed [5]. There are two primary disadvantages:

The test is sensitive to how the binning of the data is performed.

It requires sufficient sample size so that the minimum expected frequency is five.

Euclidian Distance: Euclidean distance d, between features Xi and Yi is calculated using the formula

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

T test: The t-test assesses whether the means of two groups are statistically different from each other. This helps to find how far two groups deviate from each other

$$\text{T test} = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{var1}{n} + \frac{var2}{n}}}$$

Information gain : The information gain of a given attribute X with respect to the class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X, I(Y ;X). Entropy is a measure of how "mixed up" an attribute is. It is sometimes equated to the purity or impurity of a variable.

Correlation based computation – A dependency measure

The search evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class. The downside of univariate filters for e.g. information gain is, it does not account for interactions between features, which is overcome by multivariate filters for e.g. CFS. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients are used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search. Correlation is given by [7]

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}$$

Consistency measure: This unlike all mathematical measure relies heavily on class information and depends on Min-Feature bias in selecting the subset. These measures attempt to find a minimum number of features that separate classes as consistently as the full set of features can. An inconsistency is defined as two instances having the same feature values but different class labels [12]

Experiments

Fourteen standard datasets drawn from the UCI collection were used in the experiments .These datasets were chosen because of nominal class features. The number of instances attributes and number of classes vary in the chosen dataset to represent different combinations. The learning algorithm chosen for classifying are NaiveBayes, K-NN (k=10) and C4.5 tree. All datasets were run on Pentium machine on 3 GB RAM and Java 6.

Experiments and discussion

The results in table. b shows that C4.5 performs well in most cases. Naïve Bayes is a true predictor. So the poor performance even though time taken by Naïve Bayes and C4.5 were very small compared to KNN. KNN algorithm had k=10 and has an overall average good performance on all types of datasets. Results in bold indicate the best performance for a dataset between the three chosen algorithm.

The performance of the correlation based feature subset filter model with correlation subset as evaluator is shown in table c. The figures are marked for percent correct with 10 fold cross validation. A '–' in table c indicates feature selection does have a negative performance on the dataset.

Wrapper based methods take longer time to run. The results are shown in table d. The evaluation criteria was correlation based subset evaluation for anneal, contact-lens, iris, soybean and ranker search with information gain as evaluation criteria for splice, letter, lymph, vowel, vehicle, waveform and zoo. A '+' or '-' indicates more than ± 2.0 percent significant increase or decrease with non feature selection classification result. KNN with wrapper approach shows a huge deviation from original results as shown in table d.

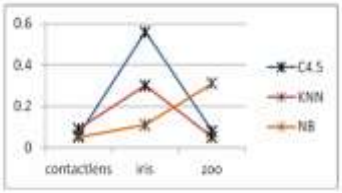

Fig g. Wrapper selection time in sec between three datasets

The iris data set, with 150 instances and 5 attributes is classified and results shown. The classification based on sepal length and sepal width is shown in fig a. The same classification based on petal length and petal width is shown in fig b. The experiment first carried out with Naives Byes classification without any attribute selection criteria. The results in table 1 compare NaiveBayes without the qualifier attributes and only with qualifying attributes and then mixed attribute set. This results are shown with 10 fold cross validation.

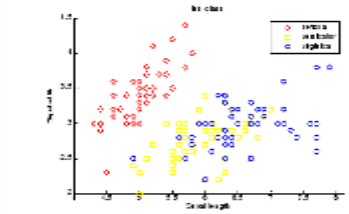

Fig. h. Iris dataset with sepal length and width classification

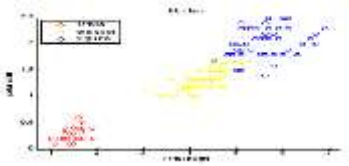

Fig. i. Iris dataset with petal length and width classification

From the figure i we can find that only petal width and petal length are prime attributes that can be used for classification. Table e shows the results recorded using Naïve Bayes classification and c4.5 on iris data set. The figures shows the number of correctly classified instance in each category. NaiveBayes being a probabilistic classifier has the ability to classify with an increase about 6.7% correct classification. When either petal length or petal width is available C4.5 performs a good classification

Comparing the results in soya bean dataset which has 683 instances and 36 attributes the results in table f were achieved. The time taken to build the model was more in NaiveBayes with correlation based feature selection and 10 fold cross

validation.

Conclusion

In this paper we have studied various datasets for classification with filter and wrapper based feature selection models. We propose to combine these two models to produce a genetic approach to feature selection that classifies more accurately with minimum number of attributes.

References

[1] I. Guyon, Andr´e Elisseff., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) pp. 1157-1182

[2] http://msdn.microsoft.com/en-us/library/ms175382.aspx

[3] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In Proceedings of the Eleventh International Conference on Machine learning, pages 121–129, New Brunswick, NJ, 1994. Morgan Kaufmann

[4] P. Saengsiri, P. Meesad, S. Na Wichian and U. Herwig, "Comparison of Hybrid Feature Selection Models on Gene Expression Data," IEEE International Conference on ICT and Knowledge Engineering, 2010, pp.13 -18

[5]http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/chsqgood.htm

[6] A.Jain, D Zongker, Feature selection: evaluation, application and small sample performance, IEEE transactions on pattern analysis and machine intelligence, vol. 19, no. 2,1997

[7] A.Gowda karegowda1, A.S.Manjunath, M.A.Jayaram, Comparative study of attribute selection using gain ratio and correlation based feature selection, Vol 2,Dec,2010

[8] R. Kohavi. Wrappers for Performance Enhancement and Oblivious Decision Graphs, PhD thesis, Stanford University, 1995.

[9] R. Kohavi and G. John. Wrappers for feature subset selection, Artificial Intelligence, special issue on relevance, 97(1–2):273–324, 1996.

[10] H.Liu, L.Yu, Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions on knowledge and data engineering, vol. 17, no. 4, April 2005.

[11] S. Yu, "Feature Selection and Classifier Ensembles :A Study on Hyper spectral Remote Sensing Data", Ph. d thesis, The University of Antwerp,2003.

[12] H. Almuallim and T.G. Dietterich, Learning Boolean Concepts in the Presence of Many Irrelevant Features,Artificial Intelligence, vol. 69, nos. 1-2, pp. 279-305, 1994

[14] P.M. Narendra and K. Fukunaga, A Branch and Bound Algorithm for Feature Subset Selection, IEEE Trans. Computer, vol. 26, no. 9, pp. 917-922, Sept. 1977

[15] J. Doak, An Evaluation of Feature Selection Methods and Their Application to Computer Security, Technical report, Univ. Of California at Davis, Dept. Computer Science, 1992

Table a: Datasets taken for study

Dataset	Instances	Attributes	No. of classes
anneal	898	39	5
contact-lens	24	5	3
glass	214	10	6
iris	150	5	3
letter	20000	17	26
lymph	148	19	4
segment-challenge	1500	20	7
Soyabean	683	63	19
Splice	3190	62	3
Vehicle	846	19	4
Vowel	990	14	11
waveform-5000	5000	41	3
Weather	14	5	2
Zoo	101	18	7

Table b: Percent correct-without attribute selection and 10 fold cross validation

Dataset	Naive Bayes	C4.5	KNN
Anneal	86.59	98.57	97.27
contact-lens	76.17	83.50	74.67
Glass	49.45	67.63	66.04
Iris	95.53	94.73	95.73
Letter	64.07	88.03	95.50
Lymph	83.13	75.84	84.18
segment-challenge	80.17	96.79	95.25
Soybean	92.94	91.78	90.12
Splice	95.41	94.03	79.86
Vehicle	44.68	72.28	70.17
Vowel	62.90	80.20	93.39
waveform-5000	80.01	75.25	79.29
Weather	57.50	47.50	71.00
Zoo	94.97	92.61	95.05

Table c: Filter method implemented on datasets

Dataset	Naive Bayes	C4.5	KNN
Anneal	86.04	97.12(-)	92.56
contact-lens	72.83(-)	81.50(-)	65.17
Glass	48.02(-)	69.15	66.15
Iris	95.93	94.80	95.93
Letter	65.52	88.26	94.43
Lymph	78.24(-)	75.49	81.73
segment-challenge	81.90	95.94(-)	92.54
Soybean	92.53	90.31(-)	85.01
Splice	95.84	94.37	84.65
Vehicle	47.61	67.10(-)	62.27
Vowel	62.42	77.97	78.56
waveform-5000	80.08	76.99	84.63
Weather	58.00	56.00	70.00
Zoo	93.29(-)	93.28	85.46

Table d: Wrapper method implemented on datasets with 10 fold CV

Dataset	Naive Bayes	C4.5	KNN
Anneal	80.17(-)	98.22	87.53(-)
contact-lens	83.33(+)	87.50(+)	79.17(+)
Glass	49.40	72.42(+)	65.30
Iris	94.87	92.66(-)	94.00
Letter	65.90	88.12	94.77
Lymph	82.60	77.03	81.94(-)
segment-challenge	81.11	95.73	93.60
Soybean	92.59	91.80	87.85(-)
Splice	95.85	94.10	84.93(+)
Vehicle	44.56	72.34	69.06
Vowel	67.46(+)	81.41	81.55(-)
waveform-5000	80.69	74.84	84.40(+)
Weather	52.00(-)	64.29 (+)	64.28(-)
Zoo	94.47	91.08	88.90(-)

Table e. iris dataset

Classifier	Full Dataset 10 cv	CFS + 10 fold cv	No qualifying attribute selected	Mixed qualifying attribute given
Naïve Bayes	95.53%	96%	79.33%	92%
C4.5	96%	96%	72.6667%	94.666%

Table f. Soybean dataset

Classifier	Full Dataset	CFS+ 10 fold cv
Naïve Bayes	92.94% Time:0.2s	92.5329 Time :8.49s
C4.5	91.78% Time : 0.11s Tree depth: 61 Leaves: 93	90.3367% Time :0.25s Tree depth :113 Leaves :77