Wrapper-based Feature Selection Using Support Vector Machine

Hwang, Young-Sup[1]

[1.] Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro 221-70, Korea
young@sunmoon.ac.kr

Abstract: Features are measurable properties and used to classify patterns. When a feature set is large, we have to select a small feature subset, since the large feature set needs much computation time and has the problem of curse-of-dimensionality: when the dimensionality of feature set increases, the required sample size grows exponentially to train a classifier. Feature selection consists of two main procedures: subset generation and evaluation. The number of subsets grows exponentially when the number of set members increases. So, usual heuristic search methods are applied to generate subsets. Evaluation criteria to select a feature subset have been related to the training data. Such characteristics include relevance and redundancy with classes. But classifying new unseen patterns accurately is more important than classifying the trained data, which is called generalization capability. An improved feature selection method is proposed to improve the generalization capability. It uses wrapper-based feature selection and uses support vector machine as the wrapper. The experimental results show that the proposed method can reduce feature size and can improve the generalization capability.

[Young-Sup H. **Wrapper-based Feature Selection Using Support Vector Machine**. *Life Sci J* 2014;11(7):632-636] (ISSN:1097-8135). http://www.lifesciencesite.com. 88

Keywords: Feature selection; wrapper; support vector machine; generalization

1. Introduction

Features are measurable properties such as weight, height, and number of blobs, etc. They can help to discriminate classes or patterns. Good features such as having small within-class variance and large between-class variance have great discriminatory power. But finding or developing such features are very difficult and requires experts. So usually we extract features as many as possible. Then we can select the optimal feature subset. The reason why we have to select a small feature subset is because of the curse-of-dimensionality; when the dimensionality of feature set increases, the required sample size grows exponentially to train a classifier. Large feature set also takes much training time and classification time. To overcome this curse-of-dimensionality, the feature size should be small enough, but finding the optimal feature set is very difficult. So there have been many researches to select a feature subset from large feature set (Mark and Geoffrey, 2009; Luis et al., 2002).

Feature set evaluation methods are divided into filter-based and wrapper-based (Ron, 1997; Wang 1999). Filter-based methods evaluate feature set independent of classifiers. To evaluate a feature subset, they analyze the relationship between the feature subset and a class. For example, when some features don't appear in a class and then the features have no relationship with the class. However the absence of the features doesn't mean that the class is not the wrong class. Ron (1997) defined that a feature is relevant: if we remove the feature, the probability of being the class changes. Filter-based methods inspect this relevance between features and classes. They also remove redundant or irrelevant features. Their execution time is fast since they don't use classifiers but the results of a specific classifier can be not good if the methods are not well suited to the classifier.

As figure 1 shows, wrapper-based methods evaluate feature subset using the target classifier. The classifier is trained using the training data. To evaluate the feature subset reasonably, *k*-fold cross validation is used usually. Since they use the target classifiers, the classification results of the test data are fine, but they take much time for the training and execution of the classifiers. This research uses the wrapper-based method and studies a method to relieve the training time problem.

Evaluation criteria to select a feature subset have been related to the training data (Qinghua, 2010). But classifying new un-trained data accurately is more important than classifying the trained data, which is called generalization capability. A feature selection method is proposed to improve the generalization capability

2. Wrapper-based Feature Selection

Feature selection starts by generating a feature subset. When the cardinality of the feature set is d, the number of the feature subsets is 2^d. So searching all subsets requires exponential time as d increases. Thus various heuristic search methods have been suggested such as sequential forward selection, sequential backward selection, plus-L minus-R selection, bidirectional search, sequential floating selection (Isabelle and Andre, 2003).

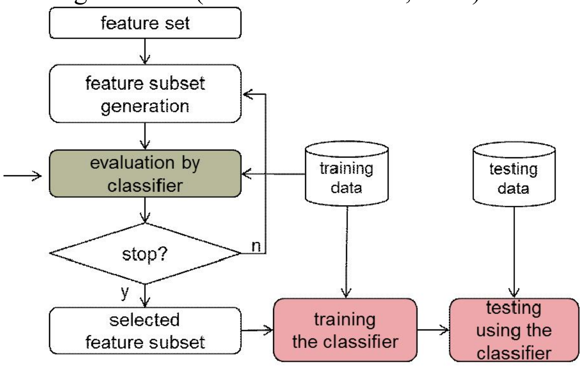

Figure 1. Wrapper-based feature subset selection

Wrapper-based feature selection method uses cross-validation to get an evaluation results on the feature subset from a classifier. A training data (with the feature subset) is divided into k subsets and the classifier is trained using k–1 subsets. The remaining 1 subset is classified by the trained classifier and the result is collected. This process is repeated k times and the averaged value is the evaluation result of k-fold cross-validation. Figure 2 shows 3-fold cross-validation.

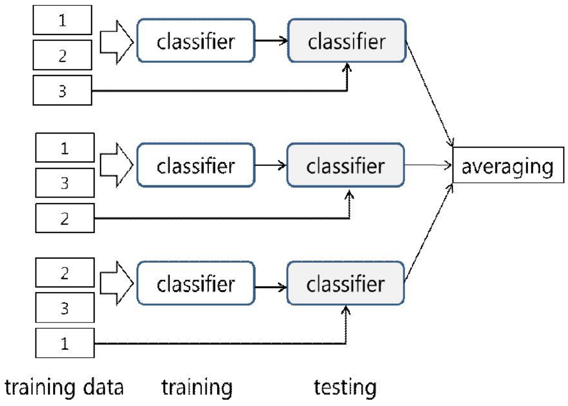

Figure 2. 3-fold cross validation

When the largest evaluation point is acquired, the feature subset is selected as the final feature subset. The classifier is trained using all the training data with the selected feature subset. The trained classifier is used to classify unseen data.

3. Sequential Forward Selection by Support Vector Machine

To overcome the exponential search space, a sequential forward selection is applied. Let the given feature set as $F = \{f_1, f_2, \dots f_n\}$ and the selected feature subset as S. Figure 3 shows the proposed selection algorithm.

The algorithm starts from an empty selected subset. At step 2, each feature in the current remaining feature set F are selected sequentially and it is added to the selected subset S and then the united subset S is evaluated using 5-fold cross-validation. If the evaluation point is highest, the subset S is maintained and the feature is removed from F. If the evaluation point is increased, repeat the procedure to add another feature.

```
(1) $S = \emptyset$
(2) Evaluate each subset $S \cup \{f_i\}$, $f_i \in F$
(3) Select $f_h$ which has the highest
    evaluation point $V$.
        A.  $S = S \cup \{f_h\}$
        B.  $F = F - \{f_h\}$
(4) If $F = \emptyset$, stop.
(5) If $V$ is larger than before, go to step 2.
(6) If $V$ is smaller than before,
```

Figure 3. Sequential Forward Selection

If the evaluation point is decreased, the S is restored by removing the added feature and the procedure stops. If F becomes empty, the procedure stops but the dimension reduction fails in this case. In this algorithm, the evaluation point is the most important. The support vector machine (SVM) is used to evaluate the feature subset (Cristopher, 1998). The SVM is also used to classify the untrained test data. The SVM is a well-known classifier which has the maximum generalization capability. Since a classifier which is optimal to the training data may not be optimal to the test data, a classifier which has a good generalization capability can have better result on the test data. The classification result of the SVM over the feature subset is used as the evaluation point.

SVM can be summarized as follows. Training data is

$$\{\boldsymbol{x}_i, y_i\}, i = 1, \dots, N. \boldsymbol{x}_i \in R^d, y_i \in \{+1, -1\}.$$

SVM computes

$$\boldsymbol{w}^T \varphi(\boldsymbol{x}) + b$$

where $\varphi()$ is a kernel function which maps $\boldsymbol{x}$ to a high dimensional space.

If y_i is +1,

$$\boldsymbol{w}^T \varphi(\boldsymbol{x}_i) + b \geq +1,$$

otherwise

$$\boldsymbol{w}^T \varphi(\boldsymbol{x}_i) + b \leq -1.$$

SVM determines $\boldsymbol{w}$ which is a hyper plane that separates the training data by solving following optimization problem.

$$\min_{\boldsymbol{w},b,\xi}\left[\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+C\sum_{i=1}^{N}\xi_i\right],$$
$$y_i(\boldsymbol{w}^T\varphi(\boldsymbol{x})+b)\geq 1-\xi_i,$$
$$\xi_i\geq 0, i=1,\ldots,N.$$

C is a constant which affects the error rate. The kernel function can make the non-linearly separable problems be solved. The most popular kernel function is the radial basis function (RBF) such as

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=e^{-\gamma\|x_i-x_j\|^2},\gamma>0.$$

There is no general rule to set C and γ, so one should determine them through experiments (Chih-Chung, 2001). By maximizing margin during learning, SVM improves generalization capability. SVM stores support vectors and weights after learning and restores them when testing unseen data. Thus SVM can classify unseen data rapidly after learning. The optimization problem to compute weights in SVM requires high mathematical knowledge but we can easily use publicly available library such as libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) and SVMLight (http://svmlight.joachims.org/).

4. Experimental Results

Experimental data in UCI machine learning repository are used (Bache, 2013). Two kinds of database is used. One is the database with small number of data such as 'anneal', 'ionosphere', 'lymp', 'segment', 'vowel', and 'zoo'. The other is the database with large number of data and large feature size, 'internet-ads'. 'anneal' is a steel annealing data. 'ionosphere' is a radar return data from the ionosphere (Sigillito, 1989). 'lymp' is a lymphography provided by the Oncology Institute (Cestink and Kononenko, 1987). 'segment' is an image data described by high-level numeric-valued attributes. 'vowel' is a vowel recognition data with context sensitive features (Turney, 1993). 'zoo' is a simple database with 7 classes of animals containing 17 attributes. 'internet-ads' is a dataset which represents a set of possible advertisements on Internet pages (Nicholas, 1999). The first two rows in table 1 and 2 describe the characteristics of the databases. Since they are open publicly, it is easy to compare our results with others. You can get the databases at http://repository.seasr.org/Datasets/UCI/arff/. The databases are reformatted to be used in WEKA. WEKA is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering and visualization. Users can add their own Java code to WEKA (Mark, 2009).

Table 1 shows experimental results for six databases. The first row shows the database name of the UCI repository. Row 3 and 4 show that the proposed algorithm can reduce the feature size of each database. Row 5 and 6 show that the reduced feature subset can improve the accuracy for the test data. Since the number of data is not great, 5-fold cross validation is used to select features and the one remaining data subset is used for testing the accuracy. The testing data is not used in the training of SVM. Figure 4 shows that the feature size is reduced about 41%. Figure 5 shows that the accuracies increased about 14 percent point.

For the case of training and testing separation, the large database, internet-ads, was experimented by the proposed algorithm. The number of data in internet-ads is 3,279 and the cardinality of feature set is 1,565. The database was divided into 2,184 training and 1,115 testing data. The proposed algorithm reduced the feature size into 20.

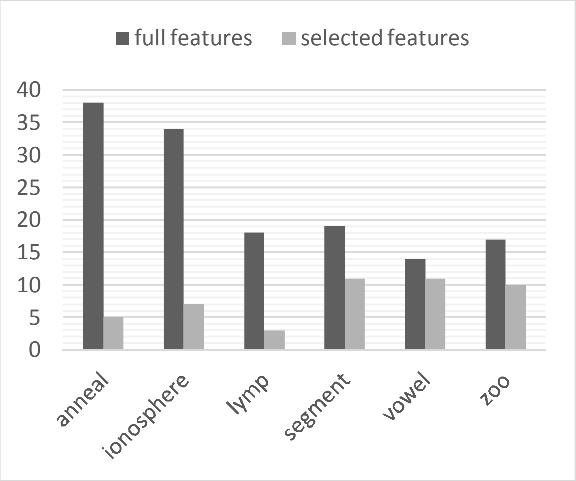

Figure 4. Experimental results of feature subset selection

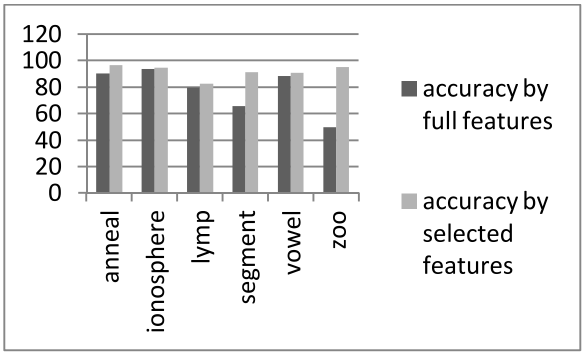

Figure 5. Experimental results of accuracies

Table 1. Experimental results

Database	anneal	ionosphere	lymp	segment	vowel	zoo
Number of data	898	351	148	2310	990	101
Cardinality of full features	38	34	18	19	14	17
Cardinality of selected features	5	7	3	11	11	10
Accuracy with full features (%)	90.1	93.4	79.7	65.4	88.5	49.5
Accuracy with selected features (%)	96.5	94.6	82.4	91.1	90.5	95

Table 2. Experimental results of the large database, *inernet-ads*

Number of data	3,279	Number of training data	2,184
		Number of test data	1,115
Cardinality of full features	1,565	Accuracy on training data	89.8%
		Accuracy on test data	87.9%
Cardinality of reduced features	20	Accuracy on training data	96.8%
		Accuracy on test data	95.7%

An SVM was trained using this 20 feature subset over 2,184 training data. The trained SVM classified 1,067 data accurately in the unseen 1,115 testing data and the accuracy is 95.7%. The normal training and testing an SVM using full feature set returned accuracy of 87.9%. The proposed algorithm improved 7.8 percent point.

From the experimental results we can see that the proposed algorithm can reduce large feature set and improves classification accuracy. The classification accuracy for the unseen data is also improved, which means that the proposed algorithm improved the generalization capability.

To verify the generalization capability of SVM, naïve Bayesian classifier was experimented as the wrapper. Naïve Bayesian classifier is a simple probabilistic model from Bayes' theorem. It can be trained fast and assumes independence of features. The feature subsets were selected from the training data using 5-fold cross validation as before. The experimental results on the test data by the naïve Bayesian classifier which was trained by the selected feature subset is presented at table 3.

Comparison between table 1 and table 3 shows that the size of the selected features increased about two times and the accuracy decreased 5.7 percent point on the average.

Table 3. Experimental results using naive Baysian classifier as the wrapper

Database	anneal	ionosphere	lymp
Cardinality of selected features	21	19	17
Accuracy with full features (%)	86.5	82.5	83.1
Accuracy with selected features (%)	92.5	92.0	83.1
Database	segment	vowel	zoo
Cardinality of selected features	13	10	15
Accuracy with full features (%)	80.2	62.9	15
Accuracy with selected features (%)	85.5	67.7	95.0

For the large database, *internet-ads*, naïve Baysian classifier as the wrapper selected 18 features using the training data and the accuracy for the test data was 93.4%. In this case, the SVM as the wrapper showed 10% better accuracy.

5. Conclusion

When a feature set is large, selection of the optimal subset makes better accuracy and faster computation. Feature selection consists of two main procedures: subset generation and evaluation. In this paper, a sequential forward selection was applied to generate a feature subset and the subset was evaluated by the wrapper-based evaluation method. SVM is used as a wrapper to improve the generalization capability. Experimental results on public database showed that the proposed algorithm can reduce feature size and can improve the generalization capability.

Corresponding Author:
Hwang, Young-Sup
Department of Computer Science and Engineering
Sun Moon University
Asan, Sunmoonro 221-70, Korea
E-mail: young@sunmoon.ac.kr

References

1. Bache, K. Lichman, M. UCI Machine Learning Repository [http://archive.ics. uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.
2. Cestnik G, Konenenko I, Bratko I, Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I. Bratko N.Lavrac (Eds.) Progress in Machine Learning, 1987:31-45.
3. Cristopher B, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 1998:2:121-167.
4. George HJ, Pat L. Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995:338-345.
5. Isabelle G and Adre E, “An Introduction of Variable and Feature Selection,” Journal of Machine Learning Research, 2003:1157-1182.
6. Chih-Chung C and Chih-Jen L, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw /~cjlin/libsvm.
7. Khin MW, Nan SMK. Feature Subset Selection approach based on Maximizing Margin of Support Vector Classifier. Engineering and Technology 2008:18:366-371.
8. Nicholas K, Learning to Remove Internet Advertisements, Proceedings of the third annual conference on Autonomous Agents , 1999:175-181.
9. Luis CM, Luis B, Angela N. Feature Selection Algorithms A survey and Experimental Evaluation. Data Mining, ICDM 2003. Proceedings. 2002 IEEE International Conference on 2002: 306-313
10. Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H. W. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009:11(1):10-18.
11. Qinghua H, Xunjina C, Lei Z, Daren Y. Feature evaluation and selection based on neighborhood soft margin. Neurocomputing 2010:73:2114-2124.
12. Ron K, George H. J. Wrappers for feature subset selection. Artificial Intelligence 1997:97:273-324.
13. Sebastian M, Richard W. A wrapper method for feature selection using Support Vector Machines. Information Sciences 2009:179:2208-2217.
14. Sigillito VG, Wing SP, Hutton LV, Baker KB, Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 1989:10:262-266.
15. Turney P, Robust Classification With Context-Sensitive Features, Proceedings of the Sixth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-93), 1993:268-276.
16. Wang H, Bell D and Murtagh F, Axiomatic approach to feature subset selection based on relevance. IEEE Transactions on pattern Analysis and Machine Intelligence, 1999:21:271-277.

5/24/2014