A NOVEL HYBRID FEATURE SELECTION ALGORITHM:USING RELIEFF ESTIMATION FOR GA-WRAPPER SEARCH

LI-XIN ZHANG, JIA-XIN WANG, YAN-NAN ZHAO, ZE-HONG YANG

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
E-MAIL:zhanglixin99@mails.tsinghua.edu.cn, wjx@s1000e.cs.tsinghua.edu.cn

Abstract:

A new feature selection method named RelifF-GA-Wrapper is proposed to combine the advantages of filter and wrapper. In the RelifF-GA-Wrapper method, the original features are evaluated by the ReliefF method, and the resulting estimation is embedded into the genetic algorithm applied to search optimal feature subset with the train accuracy of induction learning algorithm for the evaluation function. Experiments are carried on handwritten Chinese characters dataset, which is a large-scale dataset, and several other typical datasets with features more than 20. The results show RelifF-GA-Wrapper has better performance than ReliefF and GA-Wrapper, indicating that the proposed RelifF-GA-Wrapper algorithm is competitive and scales well to large datasets.

KeyWords:

Feature selection; Filter; Wrapper; Genetic algorithm; ReliefF

1 Introduction

An abundance of unnecessary features degrade the performance of concept learners either in speed or predictive accuracy. Feature selection is the problem of choosing a small subset of features that are necessary and sufficient to describe the target concept. Many feature selection algorithms have been proposed in recent years [1,2,3,4], most of which fall into filter or wrapper approaches. The difference between the filter and wrapper algorithms is whatever the feature selection is done independently of the induction algorithm. Wrapper methods use the induction algorithm itself as part of the function evaluating feature subsets, while filter methods are generally preprocessing algorithms, which do not rely on any knowledge of the algorithm to be used. Filter methods have the advantage of high speed and ability to scale to large datasets, and the drawback of large bias between the filter evaluation and the induction algorithm. Wrapper method can expect high performance of induction algorithm, but it is difficult to scale to large datasets because of the expensive computation cost [5,6].

One challenge for feature selection is to study on more challenging datasets. As Langley [7] pointed out, few of the domains reported in the literature involved more than 40 features and typical experiments were done with the widely used UCI [8] datasets, which have little irrelevant attributes. Another open problem is how to get a hybrid method that can take the advantages of both filter and wrapper.

This paper focuses on hybrid feature selection methods that can deal with large-scale datasets. We propose a two-phase feature selection method, the idea of which is to use the feature estimation from the filter phase as the heuristic information for the wrapper phase. In the first phase we adopt ReliefF [9] to get feature estimation; in the second phase, genetic algorithm (GA) is adopted as search algorithm for the wrapper selector, which makes use of induction algorithm as part of its evaluation. The feature estimation obtained from the first phase is used for guiding the initialization of the population for genetic algorithm. We experiment on handwritten Chinese characters recognition (HCCR) dataset and several other relatively large-scale datasets.

This paper is organized as follows. In section 2, Relieff and GA-Wrapper methods are described; In section 3, ReliefF-GA-Wrapper method is proposed and described in detail; Results of experiments with ReliefF-GA-Wrapper on various datasets are discussed in section 4. Conclusions and suggestions for the future work are given in section 5.

2 ReliefF and GA-Wrapper

2.1 ReliefF Estimation

The algorithm of ReliefF [9], which is the extension of Relief [10], is shown in Fig. 1. The key idea of ReliefF is to estimate the quality of attributes according to how well their values distinguish between the instances that are near

0-7803-7865-2/03/$17.00 ©2003 IEEE

to each other. For that purpose, given a randomly selected instance *R*, ReliefF searches for *k* nearest neighbors of *R* from the same class, called nearHits, and also *k* nearest neighbors from each of the different classes, called nearMisses. The quality estimation *W*[*A*] for each attribute *A* is updated depending on *R*, nearHits and nearMisses. In the update formula, the contribution of all the hits and misses are averaged. The process is repeated for *m* times to return weights of all features, where *m* is a user-defined parameter.

ReliefF is fast, not limited by data types, fairly noise-tolerant, and unaffected by feature interaction, but it does not deal well with redundant features.

```
ReliefF(D, N, M, K)
/*      D—training set, N—number of features          */
/*   M—iterate times, K—number of nearest neighbors   */
Initialize all weights, W[A], to zero
For i=1 to M do begin
    Randomly choose an instance R in D
    Find its K nearHits and K nearMisses from each class different from which R belongs to
    For A = 1 to N
        W[A] = W[A] – Avg(diff(A, R, nearHit) + Avg(diff(A, R, nearMiss)
    End.
Return all weights W[A]
```

Figure.1. ReliefF algorithm pseudo code

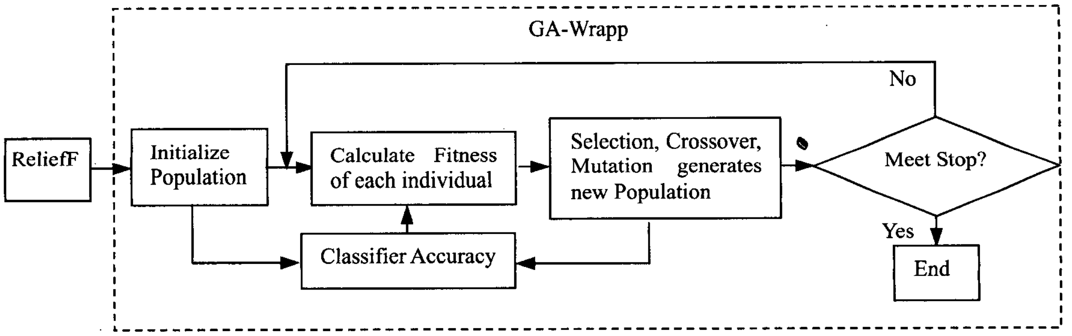
GA-Wrapp
ReliefF
Initialize Population
Calculate Fitness of each individual
Selection, Crossover, Mutation generates new Population
Meet Stop?
No
Yes
End
Classifier Accuracy

Fig.2. GA-Wrapper(circled by dashed line) and ReliefF-GA-Wrapper(the whole figure)

2.2 GA-Wrapper

As shown in Fig.2, GA-Wrapper method uses genetic algorithm [11] to search for the optimal subset of features with high train accuracy gained by the induction algorithm.

For GA-Wrapper, a feature subset is represented by a binary string with length of *n* (*n* denotes the total number of origin features), called a chromosome, with a zero or one in position *i* denoting the absence or presence of feature *i*. Each chromosome is evaluated by the performance of induction algorithm in this paper. A population of chromosomes is maintained and evolved by operators of selection, crossover and mutation until stopping criteria is satisfied.

Genetic algorithms are based on an analogy with biology in which a group solution evolves via natural selection. Though they can't always find the optimum, they would be more robust than gradient-descent algorithms when there are strong interdependencies among features. They make relatively few assumptions about the shape of the search space, and are generally quite effective for rapid global search of large search spaces in optimization problems. Genetic algorithms have demonstrated substantial improvement over a variety of random and local search methods [12].

3 ReliefF-GA-Wrapper

The whole of Fig. 2 illustrates the ReliefF-GA-Wrapper algorithm, namely ReliefF adheres to GA-Wrapper process. The population of GA is initialized based on the rank of the features according to the evaluation result gotten by ReliefF (Rank in the front means the feature has stronger distinguishing quality).

The features in the front rank should have bigger probability to be selected, which means the corresponding bit in the chromosome should have more chance to be one. The individuals are initialized according to the probabilities based on the rank of corresponding features. The fitness function of GA compromises between the size of the feature subset and the classifier accuracy.

3.1 Initialization of GA population

The population of GA is initialized as follows:

1) Get the rank list of the original features according to the evaluation results of ReliefF;

2) Generate the selection probabilities of each featurè: set the probability to be p_1 (p_1>0.5) for the feature ranking first and p_2 (p_2<0.5) for the feature ranking last, and then generate probabilities for the other features according to a specified rules, for example, geometric sequence or arithmetic sequence. Arithmetic sequence is used in this work, and p_1 and p_2 are set to be 0.8 and 0.4, respectively.

3) Individuals are initialized according to the selection probabilities of each feature obtained in step (2).

3.2 Evaluation function of individuals in GA

The evaluation functions of GA are different for different objectives [13]. In this paper, we take the classifier accuracy as a more important factor, and when the classifier accuracies of two individuals are very close, the size of feature subsets work. We propose the following evaluation function:

$$f(X) = \alpha \exp\left(-\frac{|X|}{N}\right) + \exp\left(\frac{A(X) - \beta A_0}{\beta A_0}\right)$$

where X denotes the feature subset, $|X|$ denotes the number of features in X, $A(X)$ denotes the classifier accuracy with X, A_0 denotes the classifier accuracy with the original feature set, and n is the number of original features. The first term in the right hand side means that the evaluation value decreases with an increase in the size of the selected feature subset, and the second term means the evaluation value increases with an increase in the classifier accuracy. This optimization function requires that the answer subset X satisfies $A(X)>\beta A_0$, since the exponential function penalizes heavily for subsets not satisfying $A(X)>\beta A_0$. The constant α is used for controlling the contribution of the first term to the evaluation function. The value of α and β are varied according to the problem. In this work, α and β set to be 0.5 and 0.99, respectively.

4 Empirical Study

4.1 Experimental Setup

To validate the performance of the ReliefF-GA-Wrapper algorithm on large datasets , experiments are carried out on datasets with more than 20 features. We adopt the dataset of HCCR, a relatively large dataset with 100 classes and 100 samples in each class. The other datasets with various feature type and data size are from UCI [8]. A summary of all the datasets used in this paper is presented in Table 1.

Table 1. Summary of datasets. Notations: Nom – nominal features, Num– numerical features, #Train/#Test – size of train set/ size of test set, #C – number of classes.

Dataset	Nom	Num	#Train/#test	#C
HCCR	0	256	5000/5000	100
Anneal-u	32	6	598/300	6
Auto	11	15	136/69	7
Chess	36	0	2130/1066	2
DNA	180	0	2000/1186	3
DNA-NOMINAL	60	0	2000/1186	3
German-Credit	13	7	666/334	2
Horse-colic	15	7	300/68	2
hypothyroid	7	18	2108/1054	2
ionosphere	0	34	234/117	2
nettalk	203	0	7229/7242	324
satimage	36	36	4435/2000	6
sick-euthyroid	18	7	2108/1055	2
soybean-large	35	0	455/228	19
tokyo1	0	44	479/480	2
waveform-21	0	21	300/4700	3
waveform-40	0	40	300/4700	3

The nearest mean classifier [14] is used for the dataset of HCCR. In the nearest mean classifier, the test sample is classified to the class with mean, which is calculated from the train dataset, nearest to the test sample. C4.5 decision tree are adopted for the other datasets. We run 5 trials for each dataset, which is shuffled before each trials, and the results listed are the final average. Each trial was run as follows:

(1) Get the feature evaluation by the ReliefF algorithm. The parameter M is set to be one-third the size of train dataset. K is set to be 10 for the HCCR dataset, and for the other datasets, K is set to be 1.

(2) Get the feature selection results by ReliefF, GA-Wrapper and ReliefF-GA-Wrapper respectively.

Parameters of GA are set as follows: size of population, maximum number of generations, probability of crossover, and probability of mutation are set to be 30, 10, 0.8, and 0.1 respectively. We adopt the keep-best and roulette selection rules. The stopping criteria are that best individual keeps unchanged for 5 times or maximum generations are meet.

(3) Get the test accuracies with the feature subsets selected by step (2) and all the original features.

4.2 Results and Discussion

4.2.1 Results and analysis on HCCR

Experimental results on HCCR by ReliefF-GA-Wrapper and GA-Wrapper and the corresponding ReliefFs are listed in Table 2. ReliefF-GA-Wrapper and GA-Wrapper have obviously higher accuracy compared with corresponding ReliefF (which is just listed for comparison). ReliefF-GA-Wrapper is better than GA-Wrapper in the classification accuracy, though the feature subset size of ReliefF-GA-Wrapper is slightly bigger than that of GA-Wrapper. The only difference between ReliefF-GA-Wrapper and GA-Wrapper lies in whether the feature evaluation of ReliefF is used in initializing the population of GA, suggesting that the introduction of ReliefF evaluation into GA-Wrapper is helpful to improve the accuracy. Compared with using all the original features, ReliefF-GA-Wrapper reduces the number of features in 36.6 percent with a loss of only 1.9 percent in the test accuracy.

Table 2. Feature selection and test accuracies on HCCR. Acc: test accuracy, Nf: size of feature subset, G-W: GA-Wrapper, R-G-W: ReliefF-GA-Wrapper, R1: ReliefF which selects the same number of features with GA-Wrapper, R2: ReliefF which selects the same number of features with ReliefF-GA-Wrapper

	G-W	R1	R-G-W	R2	All
Acc	0.9120	0.8865	**0.9218**	0.9065	0.9396
Nf	135.4	135.4	162.4	162.4	256

4.2.2 Results and analysis on the other datasets

The test accuracy and the size of the selected feature subset of GA-Wrapper, ReliefF-GA- Wrapper and ReliefF on the other datasets are listed in Table 3 and 4, respectively. The results with all the original features are also listed for comparison.

Comparing with GA-Wrapper, ReliefF-GA-Wrapper gets higher accuracy on 11 datasets, equal accuracy on 1 datasets, and lower accuracy on 4 datasets. The size of the feature subset obtained by ReliefF-GA-Wrapper doesn't differ much from that by GA-Wrapper. Comparing with ReliefF, ReliefF-GA-Wrapper gets higher accuracy on all datasets except *horce-colic* and *ionosphere*, and much smaller size of feature subsets on all the datasets except *Nettalk.*

ReliefF almost keep all original features except on *waveform-40* and *Nettalk*, indicating that ReliefF does not deal well with redundant features. From the comparison of accuracy and size of feature subset, it can be concluded that ReliefF-GA-Wrapper has better performance than both ReliefF and GA-Wrapper on relatively large-scale datasets.

Someone may doubt that when the generations of GA increase, the initialization of population may not contribute much to GA for searching the optimum feature subset, then the ReliefF-GA-Wrapper and GA-Wrapper will not differ much. This is validated by our successive experiment. However, it is prohibitive to adopt many generations for GA in the problem of large-scale datasets, due to the computation cost of the wrapper structure and the genetic algorithm, ReliefF-GA-Wrapper can achieve a satisfactory performance with respect to both the accuracy and computation cost

Table 3. Test accuracy with various methods. G-W: GA-Wrapper, R-G-W: ReliefF-GA-Wrapper.

Data	G-W	R-G-W	ReliefF	All
anneal-U	0.983	0.990	0.987	0.987
Auto	0.696	0.812	0.623	0.623
Chess	0.982	0.992	0.989	0.995
DNA	0.900	0.922	0.927	0.927
DNA-nominal	0.917	0.929	0.924	0.924
German	0.704	0.722	0.724	0.733
horse-colic	0.824	0.838	0.853	0.853
hypothyroid	0.992	0.992	0.992	0.992
ionosphere	0.880	0.863	0.880	0.880
Nettalk	0.635	0.723	0.722	0.720
Satimage	0.857	0.849	0.854	0.854
sick-euthyroid	0.969	0.980	0.977	0.980
Soybean-large	0.908	0.890	0.895	0.895
tokyo1	0.908	0.900	0.892	0.906
waveform-21	0.697	0.709	0.702	0.702
waveform-40	0.711	0.721	0.725	0.704

Table 4. Number of features selected by various methods, notations are the same with Table 3.

Data	G-W	R-G-W	ReliefF	All
anneal-U	17	23	30	38
Auto	13	16	25	26
chess	25	28	29	36
DNA	88	110	180	180
DNA-nominal	36	37	60	60
German	14	13	18	20
horse-colic	13	10	20	22
hypothyroid	9	9	24	25
ionosphere	15	22	33	34
Nettalk	113	166	166	203
Satimage	21	25	36	36
sick-euthyroid	10	16	23	25
soybean-large	22	24	34	35
tokyo1	26	26	37	44
waveform-21	12	12	20	21
waveform-40	21	19	28	40

From experiments on various datasets above, we can see that ReliefF-GA-Wrapper gains better or as good as performance compared with ReliefF and GA-Wrapper on relative large-scale datasets.

5 Conclusion

ReliefF-GA-Wrapper is a two-phase algorithm combining the filter with the wrapper, in which the wrapper process acts as the amender for filter and the result of filter provides heuristic information for the search of wrapper procedure. ReliefF-GA-Wrapper has no limit to data type only if the classifier used that has no special restriction, and the experiments on various type and sized datasets validate that ReliefF-GA-Wrapper outperforms ReliefF or GA-Wrapper separately on large-scale datasets.

The spirit of ReliefF-GA-Wrapper can be used to construct other similar feature selection methods. The filter methods used in the first phase and the search algorithms used in the second phase may chose others besides ReliefF and the genetic algorithm, and the combining manner of filter with wrapper needs further investigation as well.

Acknowledgements

We gratefully thank Prof. Ma Shaoping in Tsinghua University for providing datasets of handwritten Chinese characters recognition.

References

[1] Almullim, H., Dietterich, T., Learning with many irrelevant features. In Proceedings of Ninth National Conference on Artificial Intelligence, Vol. 2, AAAI Press, pp. 547-552, 1991.

[2] Pudil, P., Novovicova, J., Kittler, J. M., Floating search methods in feature selection. Pattern Recognition Letters, 15(11), pp. 1119-1125, 1994.

[3] Yang, J., Honavar, V., Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, Vol. 13, pp. 44-49, 1998.

[4] Weston J., et al, Feature selection for SVMS, In Advances in Neural Information Processing Systems, vol. 13, pp. 668-674, 2000.

[5] Kohavi, R., John, G.., Wrappers for feature subset selection. Artificial Intelligence, Vol. 97, pp. 273-324, 1997.

[6] Liu, H., Motoda, H., Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, pp. 62-68, 1998.

[7] Langley, P., Selection of relevant features in machine learning. R.Greiner,editor. In Proc. AAAI Fall Symposium on Relevance. New Orleans: AAAI Press, pp. 140-144, 1994.

[8] Murphy, P. M., Aha, D. W., UCI ML Repository, http://www.ics.uci.edu/~mlearn/ MLRepository.html, 1994.

[9] Kononenko, I., Estimation attributes: analysis and extensions of RELIEF, In proceedings of the 1994 European Conference on Machine Learning, pp. 171-182, 1994.

[10] Kira, K., Rendell, L. A., The feature selection problem: Traditional methods and a new algorithm. In AAAI-92, Proceedings of the Ninth National conference on Artificial Intelligence, AAAI Press, pp. 129-134, 1992.

[11] Goldberg, D.E., Genetic Algorithm in Search, Optimiation & Machine Learning. Addison Wesley, 1989.

[12] Davis, L., Handbook of Genetic Algorithms, Van Nostrand reinhold, 1991.

[13] Kudo, M., Sklansky, J., Comparison of algorithms that select features for pattern classifiers. Pattern Recognition. Vol. 33, pp. 25-41, 2000.

[14] Fukunaga K., Introduction to statistical pattern recognition, Academic Press, pp. 400-407, 1990.