The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms

Bai-Ning Jiang[1] Xiang-Qian Ding[2] Lin-Tao Ma[2]
Ying He[2] Tao Wang[3] Wei-Wei Xie[1]

[1]Department of Electronic Engineering, Ocean University of China, Qingdao 266071, China
[2]Center of Information Engineering, Ocean University of China, Qingdao 266071, China
[3]Department of Computer Science, Ocean University of China, Qingdao 266071, China

Abstract A hybrid feature selection method called SU-GA-W is proposed to make full use of advantages of filter and wrapper methods. This method falls into two phases. The filter phase removes features with lower SU and guides the initialization of GA population; the wrapper phase searches the final feature subset. The effectiveness of this algorithm is demonstrated on various data sets.

Keywords Feature Selection; Filter; Wrapper; Symmetrical Uncertainty; Genetic algorithms

1 Introduction

Feature selection has been the focus of interest in statistical pattern recognition[1], machine learning[2,3], and data mining[4,5]. Feature selection aims to choose an optimal subset of features that are necessary and sufficient to describe the target concept. It has proven in both theory and practice effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results[6,7].

Optimal feature selection requires an exponentially large search space, where N is the number of features [8]. So it may be too costly and impractical .Many feature selection methods have been proposed in recent years. They can fall into two approaches: filter and wrapper[9]. The difference between the filter model and wrapper model is whether feature selection relies on any learning algorithm. The filter model is independent of any learning algorithm, and its advantages lies in better generality and low computational cost[10]. The wrapper model relies on some learning algorithm, and it can expect high classification performance, but it is computationally expensive especially when dealing with large scale data sets[11].

The paper combines the two models to make use of their advantages. We adopt a two-phase feature selection method. The filter phase removes some features and uses the feature estimation as the heuristic information to guide GA[12]. We adopt symmetrical uncertainty[13]to get feature estimation; the second phase is a genetic algorithm (GA)-based wrapper selector. The feature estimation obtained from the first phase is used for

guiding the initialization of the population for genetic algorithms[12]. The effectiveness of this method is demonstrated through empirical study on UCI data sets.

In section 2, we describe the definition of Symmetrical Uncertainty (SU); in section 3, the combination of SU and GA-Wrapper is proposed and described; in section 4, the effectiveness of the algorithm is evaluated on various data sets. Conclusions are given in section 5.

2 Information Gain and Symmetrical Uncertainty

The symmetrical uncertainty between features and the target concept can be used to evaluate the goodness of features for classification[14]. The feature having larger SU value gets higher weight. The population of GA is initialized based on the weight of the features .In other words, the features that have higher weight should have bigger probability to be selected in initial population of GA.

Information gain, as an extensive correlation measure, is based on the information-theoretical concept of entropy, a measure of the uncertainty of a random variable.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

And the entropy of X after observing values of another variable Y is defined as

$$(X|\mathrm{Y}) = -\sum_{\mathrm{j}} \mathrm{P}(\mathrm{y_j}) \sum_{\mathrm{i}} \mathrm{P}(\mathrm{x_i}|y_j)\log_2(P(x_i|y_j))$$

The decrease of the uncertainty of X after observing Y, namely information gain, is defined as

$$IG(X;Y) = H(X) - H(X|\mathrm{Y})$$

$IG(X;Y)$ is a measure of dependency between variable X and variable Y. Generally, it should be normalized to between 0 to 1; Therefore, we choose symmetrical uncertainty(SU) as a measure of correlation between features and the concept target, then give features corresponding weight by their SU value. The feature having larger SU value gets higher weight.

SU is defined as

$$SU(X,Y) = \frac{2IG(X;Y)}{H(X)+H(Y)}$$

If the data sets have continuous features, the features need to be properly discretized[15] or approximating their densities with non-parametric method such as Parzen Windows[16].

3 The Combination of SU and GA-Wrapper

3.1 Genetic Algorithms

Genetic algorithms (GA), a form of inductive learning strategy, are adaptive search technique initially introduced by Holland[17]. Genetic Algorithms are designed to simulate

the evolutionary processes that occur in nature[18]. The basic idea is derived from the Darwinian theory of survival of the fittest.

There are three fundamental operators in GA: selection, crossover and mutation within chromosomes. As in nature, each operator occurs with a certain probability. There must be a fitness function to evaluate individuals' fitness. The evaluation function is a very important component of the selection process since offspring for the next generation are determined by the fitness values of the present population. Crossover and mutation are used to generate new individuals (offspring) for the next generation. Crossover operates by randomly selecting a point in the two selected parents and exchanging the remaining segments of the parents to create new individuals. Mutation operates by randomly changing one or more components of a selected individual. Figure 1 provides a simple diagram of the iterative nature of genetic algorithms.

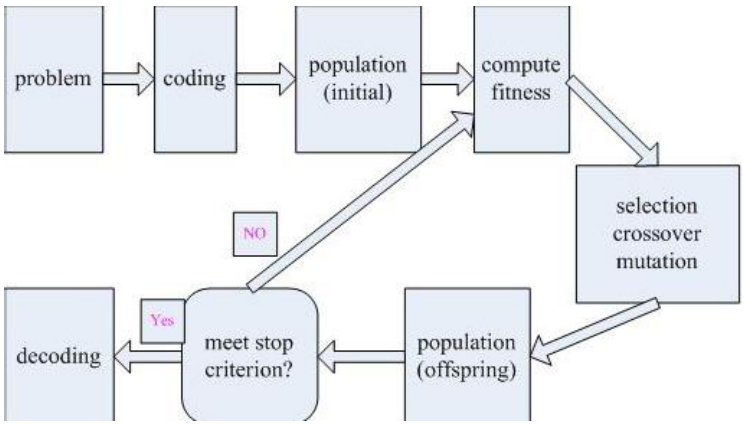

Figure 1: Flow diagram of genetic algorithms.

3.2 The Combination of SU and GA-Wrapper

We name the combination of SU and GA-Wrapper to SU-GA-W. This method will be described in detail next.

3.2.1 Computation of Symmetrical Uncertainty

The first phase, we compute symmetrical uncertainty (SU) between features and the target concept. The SU value has two main functions: it can remove the features with SU lesser than threshold and gets every feature's weight to be used to guide the initialization of the population for genetic algorithms. The feature having larger SU value gets higher weight.

3.2.2 Application of Genetic Algorithms

The second phase is the application of genetic algorithms in optimal feature subset searching. The fitness function of GA consists of the classifier accuracy and the size of the features. So it belongs to a wrapper method.

Coding: A feature subset is represented by a binary string with length of n (n is the number of features), with a zero or one denoting corresponding feature whether to be selected.

Initialization of GA population: We rank the features with SU value greater than threshold according to their weight. Then we set the selection probabilities of each feature: set the probability to be p_1 for the feature ranking first and $p_2$for the feature ranking last, and then generate probabilities for the other features according to arithmetic sequence[12]

Design of fitness function: the design of fitness is an important step. We take into account both the classifier accuracy and the size of the feature subset. The classifier accuracy is thought of as a more important factor. We design the following fitness function:

X denotes the subset, $|X|$ denotes the number of features in X, $error(X)$ denotes the classifier error rate, N is the number of all features, and control the relative importance of and $|X|$ ($\alpha = 2$, $\beta = 1$ in this paper).

Selection, Crossover and Mutation: We adopt tournament selection, uniform crossover, and standard mutation. Tournament selection operates by randomly selecting a set number of candidates, from which the two fittest chromosomes survive. The survivors, called the parent chromosomes, are then subjected to crossover and mutation[19].

4 Empirical Study

4.1 Experimental Setup and Results

In our experiments, we choose 5 datasets from the UCI Machine Learning Repository[20]. A summary of data sets is presented in Table 1.

Table 1: Summary of UCI data sets.

Title	Features	Instances	Classes
dermatology	33	366	6
lung-cancer	56	32	2
Breast Cancer	30	569	2
soybean-large	35	307	19
ionosphere	34	351	2

We choose the Naive Bayes Classifier to evaluate the goodness of the final feature subset by 10-cross-validation. We compare the classifier accuracy of SU-GA-W, GA-Wrapper, SU, and full data sets (Table 3). The size of the selected feature subset of SU-GA-W, GA-Wrapper and SU is also compared. (Table 2). SU only uses symmetrical uncertainty and GA-Wrapper uses genetic algorithms to search feature subset.

The SU threshold is 0.15 in our experiments. The features with SU value lesser than the threshold are removed; the others are ranked according to their SU value, and then guide the initialization of population in GA. Parameters of GA are set as follows: size of population (20), maximum number of generations (20), probability of crossover and mutation are set to be 0.6 and 0.033 respectively. The Naive Bayes Classifier accuracy and the size of subset are used to evaluate the fitness of individuals. The classifier accuracy is obtained by 5-cross-validation method.

Table 2: Number of features selected on UCI data sets.

Title	**All features**	**SU**	**GA-Wrapper**	**SU-GA-W**
dermatology	33	24	10	**9**
lung-cancer	56	12	25	**2**
Breast Cancer	30	18	8	**3**
soybean-large	35	21	17	**13**
ionosphere	34	32	**9**	10
AVERAGE	37.6	21.4	13.8	**7.4**

Table 3: The classifier accuracy on UCI data sets.

Title	**All features**	**SU**	**GA-Wrapper**	**SU-GA-W**
dermatology	97.26	97.27	**98.91**	**98.91**
lung-cancer	84.38	87.5	**90.63**	**90.63**
Breast Cancer	92.97	92.97	95.96	**96.84**
soybean-large	92.18	92.18	**93.81**	92.84
ionosphere	82.62	82.91	92.02	**92.59**
AVERAGE	89.88	90.57	94.27	**94.36**

4.2 Analysis and Discussion

From Table 3, we observe that SU-GA-W gets the highest classifier accuracy on the most data sets. From Table 2, it is clear that SU-GA-W achieves higher level of dimensionality reduction by selecting less number of features than other methods.

From experiments on various data sets above, we can see that SU-GA-W obtains better performance than GA-Wrapper and SU. It is an effective feature selection algorithm.

5 Conclusion

In this paper, we adopt two-phase algorithm, called SU-GA-W, to select optimal feature subset. This method is combination of filter and wrapper approaches. The filter phase removes features with lower SU and guide the initialization of GA population. The wrapper phase searches final feature subset. The experiments demonstrate the effectiveness of this method on various data sets.

References

[1] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(3):301–312.

[2] H. Liu, H. Motoda, and L. Yu, Feature Selection with Selective Sampling, *Proceedings of the 19th International Conference on Machine Learning*. July 8-12, 2002. Sydney, p. 395-402.

[3] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 2003, 53:23–6.

[4] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pages 365–369.

[5] D. Manoranjan, K. Choi, P. Scheuermann, H. Liu, Feature Selection for

[6] Clustering - A Filter Solution, *Proceedings of ICDM* 2002. Dec 9 - 12, 2002, p. 115-122.

[7] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, pages 284–292.

[8] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 1997, 97:245–271.

[9] H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 1994, 69(1-2):279–305.

[10] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 1997, 97(1-2): 273–324.

[11] M. Dash and H. Liu, Feature Selection for Classification. *Intelligent Data Analysis* - An International Journal, Elsevier, Vol. 1, No. 3, 1997, pages 131 – 156.

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

[13] Li-Xin Zhang, Jia-Xin Wang, Yan-Nan Zhao and Ze-Hong Yang. A novel hybrid feature selection algorithm: using ReliefF estimation for GA-Wrapper search. *Machine Learning and Cybernetic*s, International Conferenc**e,** 2003, page(s): 380- 384.

[14] Press, W. H., Flannery, B. P., Teukolsky, S. A., &Vetterling, W. T. (1988). Numerical recipes in C Cambridge University Press, Cambridge.

[15] L. Yu and H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03)*, Washington, D.C. 2003, pp. 856-863. August 21-24.

[16] H. Liu and R. Setiono, Feature Selection via Discretization of Numeric Attributes, IEEE *Trans. on Knowledge and Data Engineering*, VOL.9, NO. 4, July/August 1997. pp. 642-645.

[17] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, USA, 2nd edition, 2001.

[18] Holland, J. H., *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, MI., 1975.

[19] J.R. Koza, M.A. Keane and M. Streeter. Evolving Inventions. *Scientific American*, February 2003:52-59

[20] H. Vafaie, I.F. Imam. Feature selection methods: genetic algorithms vs. greedy-like search, *Proceedings of the International Conference on Fuzzy and Intelligent Control Systems*, 1994.

[21] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.