KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 429 - 434

A COMPARATIVE EVALUATION OF MEDIUM- AND LARGE-SCALE FEATURE SELECTORS FOR PATTERN CLASSIFIERS

MINEICHI KUDO[1] AND JACK SKLANSKY

Needs of feature selection in medium and large problems increases in manv fields including medical and image processing fields. Previous comparative studies of feature selection algorithms are not satisfactory in problem size and in criterion function. In addition, no way has not shown to compare algorithms with different objectives. In this study, we propose a unified way to compare a large variety of algorithms. Our results show that the sequential floating algorithms promises for up to medium problems and genetic algorithms for medium and large problems.

1. INTRODUCTION

Feature selection aims mainly two goals: (1) reduction of the cost of extracting features and (2) improvement of the classification accuracy of a practical classifier. Especially, the second goal has received a great deal of attention in recent years according to the increase of the problem size. We compare many algorithms on medium (20–40 in feature number) and large problems (40– in feature number) (medium- and large-scale feature selection). In medium- and large-scale feature selection, there exist many garbage features which can degrade the performance of a practical classifier. In such a case, removing garbage features is useful to improve the classification accuracy of the classifier.

So far, a large number of algorithms have been proposed for feature selection and many comparative studies have been done [1, 3, 5, 10]. However, these studies do not treat large problems or use only *monotonic criterion* in which an addition of a feature improves or keeps the criterion value before. Since the latter limitation is not practical, we use the error rate of a classifier as our non-monotonic criterion. The error rate is estimated directly from the training data using the leave-one-out technique or the cross validation technique.

Another problem of previous comparative studies is that they compare algorithms in the entire range of the number of features. However, usually our main concern is in only a part where the classification accuracy is not degraded so much. We propose

[1] Supported by NSF and JSPS in Japan-U.S. Cooperative Science Program and by the Telecommunications Advancement Foundation.

a methodology to compare algorithms with diffrent objectives in a range where a loss of the classification accuracy is small.

2. ALGORITHMS

Let the original feature set be Y, $|Y| = n$, and a criterion function be $J(X)$ to evaluate a feature subset X. Algorithms for feature selection are divided into three categories in objective: (A) algorithms aim to find the best X of a given size $m(< n)$, (B) algorithms aim to find the smallest X with $J(X) \geq \theta$, and (C) algorithms aim to find the optimal X in an optimization function $O(X)$.

Algorithms compared are shown in Table 1 along with their objective types and search types. For the detail of the first six algorithms, see Kittler [3].

Table 1. Feature selection algorithms (the search types is (S) Sequential or (P) Parallel).

Obj.	Search	Algorithms
A	S	SFS, SBS, GSFS(g), GSBS(g), PTA(l,r), GPTA(l,r), SFFS SBFS, BAB$^+$
B	S	RBAB
C	P	GA, PARA

SFS, SBS, GSFS(g), GSBS(g): (Generalized) sequential forward (backward) serach method. SFS selects the best significant feature and then the best pair including the first one, and so on. SBS is the backward version. These algorithms are generalized to GSFS(g) and GSBS(g) in such a way that the best g-feature subsets is chosen for addition or deletion in the algorithms.

PTA(l,r), GPTA(l,r): (Generalized) Plus-l take-away-r algorithm. Go l stages forward (by adding l features) by SFS and go r stages backward (by deleting r features by SBS) and repeat this process. In the generalized algorithm (GPTA(l,r)), GSFS(l) and GSBS(r) are used instead of SFS and SBS.

SFFS, SBFS: The floating version of PTA(l,r). Unlike PTA(l,r), SFFS can backtrack unlimitedly as long as the backtrack finds a better feature subset than the feature subset obtained so far at the same size [5].

BAB$^+$: The improved branch and bound method [9]. This method gives the optimal solution when the criterion function J is monotonic.

RBAB: The relaxed branch and bound method [2]. RBAB aims to find the smallest subset for which the criterion value is not under a given threshold θ and the search is carried out for a larger set of subsets for which the criterion values are over $\theta - \delta(\delta > 0)$, where δ is called a *margin*.

GA: The genetic algorithm [6, 8]. In GA, a feature subset is represented by a binary string with length n, called a *chromosome*, with a zero or one in position i denoting the absence or presence of feature i. Each chromosome is evaluated in its fitness through an optimization function in order to survive to the next generation. A population of chromosomes is maintained and evolved by two operators of crossover and mutation. We use the following two optimization function for GA in accord with Type-A and B algorithms,

$$O_A(X) = \begin{cases} J(X) - \epsilon|X| & (|X| \leq m) \\ J_{\min} - \epsilon|X| & (|X| > m) \end{cases}$$

and

$$O_B(X) = \begin{cases} -|X| + (J(X) - J_{\min})/(J_{\max} - J_{\min} + \epsilon) & (J(X) \geq \theta) \\ -n + (J(X) - J_{\min})/(J_{\max} - J_{\min} + \epsilon) & (J(X) < \theta) \end{cases},$$

where ϵ is an arbitrary small positive constant and $J_{\max}$ and $J_{\min}$ are upper and lower bounds of J which are estimated in a preliminary feature selection described later. In addition, we use one more criterion O_C as $O_C(X) = J(X)$.

GA has arbitrariness in the population size N, the maximum number of generations T, the probability of crossover p_c, and the probability of mutation p_m. In this study, $N = 2n$ and $T = 50$. We use mainly two pairs of $(0.8, 0.1)$ and $(0.6, 0.4)$ for (p_c, p_m). We use the following two types of initial populations of chromosomes: (P1) $2n$ extreme feature subsets consisting of n distinct 1-feature subsets and n distinct $(n-1)$-feature subsets and (P2) $2n$ random feature subsets in which the number of features is in $[m-2, m+2]$ and all features appear as evenly as possible.

PARA: A parallel algorithm devised to compare with GA. This algorithm maintains a population of N feature subsets as the same as GA has but it updates the population only by local hill-climbing, that is, the population of the next generation is made from N best feature subsets from all unvisited supersets and subsets of the present N features subsets. In PARA, $N = 2n$ and $T = 50$.

3. METHODOLOGY OF COMPARISON

Once an algorithm is carried out over the entire range of the number of features, we can get a curve, called a *criterion curve*, of the algorithm. This is possible only when the algorithms are sequential (Type-A algorithms). Many previous studies compared algorithms in criterion curves. However, our main concern is in only a range of the number of features in which the criterion value does not decrease so much. To cope with this difficulty, we take the following approach. First, to capture the problem, we have one or two criterion curves using algorithms with a lower time complexity (in this paper, SFS and SBS). This approach is inspired by Siedlecki and Sklansky [7]. Then, based on this criterion curve(s), we classify the problem into one of three cases: (1) monotonic case (m.), (2) approximate monotonic case (a.m.) and (3) non-monotonic case (n.m.). Second, we choose either of two different settings

according to these cases (Figure 1). When the problem is monotonic or approximate monotonic, we determine a parameter $\alpha(= 1\%, 5\%)$ and find the point that is degraded with α as compared to the maximum criterion value $J_{\max}$. From this α-degradation point, we determine a criterion value J_α as a threshold θ and the corresponding number of features m_α.

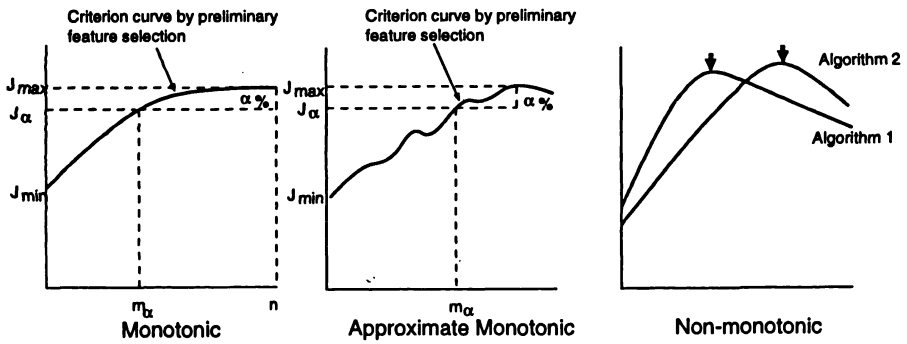

Fig. 1. Categories of problems.

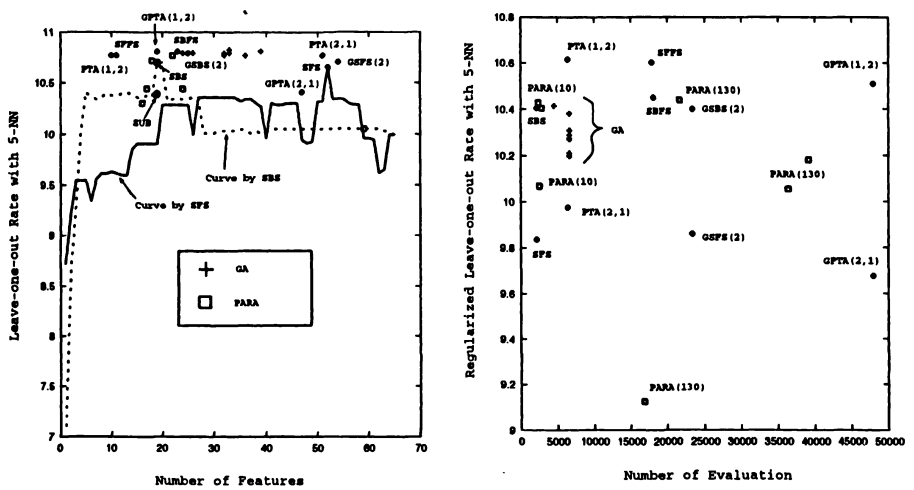

Fig. 2. Result graphs for mammogram (large) data. (a) m vs. $J(X)$. (b) the number of evaluation vs. $J(X)/(11-10) - |X|/(65-1)$.

Then, m_α is passed to Type-A algorithms and J_α to Type-B algorithms. For Type-C algorithms, both values are used in their optimization functions O_A and O_B. In addition, an upper bound $J_{\max}$ and a lower bound $J_{\min}$ in J are read from the criterion curve and used in O_A and O_B. If the problem is non-monotonic, we use Type-A and Type-C algorithms only. A Type-A algorithm is carried out so as to get the maximum point of the criterion curve (Figure 1). For Type-C algorithms, we use O_C as the optimization function, that is, J itself.

4. EXPERIMENTS

Used datasets are summarized in Table 2 with problem type. As an artificial data, we used a well-known Kittler's data [3, 5, 10] . The mammogram data are gathered from University of California, San Francisco (UCSF), the Mammographic Image Analysis Society (MIAS), and the University of California, Los Angels (UCLA). Other data are taken from UCI Repository of machine learning databases [4]. Some data are used again after some features are pre-selected. The results are shown in Table 3. The results of mammogram (large) data are shown in Figure 2. In two large problems, the results show that SFFS, SBFS and GA succeeded to find near-optimal solutions in the criterion function. In evaluation number, GA is superior to others.

Table 2. Experimental data (n: # of features, M: # of classes, K: # of training samples per class).

Database	n	M	K	Criterion J	Type
Vehicle [4]	18	4	199-218	(9-CV) linear classifier	a. m.
Mammogram (small)	19	2	57 and 29	(L) weighted 5-NN	a. m.
Kittler's data	20	2	1000 each	Mahalanobis	m.
Sonar (small)	20	2	111 and 97	(L) 1-NN	a. m.
Sonar (large)	60	2	111 and 97	(L) 1-NN	n. m.
Mammogram (large)	65	2	57 and 29	(L) weighted 5-NN	n. m.

Table 3. Summary of results.

Database	n	Top three algorithms in J (m if tie)	Best algorithms in J and Time
Vehicle	18	GA,SBS,BAB$^+$	SBS, BAB$^+$
Mammogram (small)	19	SBS,SBFS,PTA(1,2)	SBS, SBFS
Kittler's data	20	BAB$^+$, SBS,SBFS	BAB$^+$, SBS
Sonar (small)	20	PARA,GA,GPTA(1,2)	GSBS(2)
Sonar (large)	60	GA,SFFS,SBFS	GA, PARA
Mammogram (large)	65	GA,SBFS,GPTA(1,2)	GA

5. CONCLUDING REMARKS

Through many experiments, we obtained some conclusions as follows:

1. Preliminary feature selection using algorithms with a low time complexity is effective to capture the problem and to determine some parameters needed for algorithms in the further fine feature selection.

2. Among sequential search algorithms, the floating search algorithms (SFFS and SBFS) are effective in small and medium problems, but are time-consuming

in large problems. We recommend to use both forward and backward search algorithms together because of their dependance on problems.

3. GA is suitable for finding optimal solutions and is efficient in large problems. A repetition of a few runs with different sets of parameters are recommended. GA is also effective for finding the smallest feature subset with sufficient discriminatory information in approximate monotonic problems.

4. BAB$^+$ can be very efficient in monotonic problems and RBAB has a high possibility to find the best solution in approximate monotonic problems.

(Received December 18, 1997.)

REFERENCES

[1] F. J. Ferri, P. Pudil, M. Hatef and J. Kittler: Comparative study of techniques for large-scale feature selection. In: Pattern Recognition in Practice IV (E. S. Gelsema and L. N. Kanal, eds.), Elsevier Science B. V. 1994, pp. 403–413.

[2] I. Foroutan and J. Sklansky: Feature selection for automatic classification of non-gaussian data. IEEE. Trans. Systems Man Cybernet. *17* (1987), 187–198.

[3] J. Kittler: Feature set search algorithms. In: Pattern Recognition and Signal Processing (C. H. Chen, ed.), Sijthoff and Noordhoff, Alphen aan den Rijn 1978, pp. 41–60.

[4] P. M. Murphy and D. W. Aha: UCI Repository of machine learning databases [Machine-readable data repository]. Department of Information and Computation Science University of California, Irivne 1996.

[5] P. Pudil, J. Novovičová and J. Kittler: Floating search methods in feature selection. Pattern Recognition Lett. *15* (1994), 1119–1125.

[6] W. Siedlecki and J. Sklansky: A note on genetic algorithms for large-scale feature selection. Pattern Recognition Lett. *10* (1989), 335–347.

[7] J. Sklansky and W. Siedlecki: Large-scale feature selection. In: Handbook of Pattern Recognition and Computer Vision (L. F. Pau, C. H. Chen and P. S. P. Wang, eds.), Chapter 1.3, World Scientific 1993, pp. 61–123.

[8] M. R. Vriesenga: Genetic Selection and Neureal Modeling for Designing Pattern Classifier. Doctor Thesis, University of California, Irvine 1995.

[9] B. Yu and B. Yuan: A more efficient branch and bound algorithm for feature selection. Pattern Recognition *26* (1993), 6, 883–889.

[10] D. Zongker and A. Jain: Algorithms for feature selection: An evaluation. In: 13th International Conference on Pattern Recognition 1996, pp. 18–22.

Dr. Mineichi Kudo, Associate Professor, Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Sapporo 060-8628. Japan.
e-mail: mine@main.eng.hokudai.ac.jp

Prof. Dr. Jack Sklansky, Department of Electrical Engineering, University of California, Irvine, California 92697. U. S. A.
e-mail: sklansky@uci.edu