On the Greediness of Feature Selection Algorithms

Kan Deng and Andrew W. Moore
kdeng@ri.cmu.edu, awm@cs.cmu.edu

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract: Based on our analysis and experiments using real-world datasets, we find that the greediness of forward feature selection algorithms does not severely corrupt the accuracy of function approximation using the selected input features, but improves the efficiency significantly. Hence, we propose three greedier algorithms in order to further enhance the efficiency of the feature selection processing. We provide empirical results for linear regression, locally weighted regression and k-nearest-neighbor models. We also propose to use these algorithms to develop an off-line Chinese and Japanese handwriting recognition system with automatically configured, local models.

Keywords*: feature selection, cross-validation, function approximation, handwriting recognition, memory-based learning, instance-based learning.*

1. Introduction

A fundamental problem of machine learning is to approximate the function relationship *f()* between an input $X = \{x_1, x_2, \ldots, x_M\}$ and an output *Y*, based on a memory of data points, $\{X_i, Y_i\}$, $i = 1, \ldots, N$. Sometimes the output *Y* is not determined by the complete set of the input features $\{x_1, x_2, \ldots, x_M\}$, instead, it is decided only by a subset of them $\{x_{(1)}, x_{(2)}, \ldots, x_{(m)}\}$, where $m < M$. Theoretically, it is fine to use all the input features, including those irrelevant features, to approximate the underlying function between the input and the output. However, practically there are two problems which may be evoked by the irrelevant features involved in the learning process.

1. The irrelevant input features will induce greater computational cost. For example, using *k*d-trees [Moore et al, 97], locally weighted linear regression's computational expense is $O(m^3 + m^2 \log N)$ for doing a single prediction, where *N* is the number of data points in memory and *m* is the number of features used.

2. The irrelevant input features may lead to overfitting. For example, in the domain of medical diagnosis, our purpose is to figure out the relationship between the symptoms and their corresponding diagnosis. If by mistake we include the patient ID number as one input feature, an over-tuned machine learning process may come to the conclusion that the illness is determined by the ID number.

Another motivation for feature selection is that, since our goal is to approximate the underlying function between the input and the output, it is reasonable and important to ignore those input features with little effect on the output, so as to keep the size of the approximator model small. Based on information theory, [Akaike, 73] proposed several versions of model selection criteria, which basically are the trade-offs between high accuracy and small model size.

The feature selection problem has been studied by the statistics and machine learning communities for many years. It has received more attention recently because of enthusiastic research in data mining. According to [John, et al., 94]'s definition, [Kira, et al., 92; Almuallim, et al., 91; Moore et al., 94; Skalak, 94; Koller, et al., 96] can be labelled as "filter" models, while [Caruana et al., 94; Langley et al., 94]'s research is classified as "wrapped around" methods. In the statistics commu-

nity, feature selection is also known as “subset selection”, which is surveyed thoroughly in [Miller, 90].

The brute-force feature selection method is to exhaustively evaluate all possible combinations of the input features, and then find the best subset. Obviously, the exhaustive search’s computational cost is prohibitively high, with considerable danger of overfitting. Hence, people resort to greedy methods, such as forward selection. In this paper, we propose three greedier selection algorithms in order to further enhance the efficiency. We apply these algorithms to Chinese and Japanese handwriting recognition to test whether or not the features selected by the greedy algorithms are satisfactory to distinguish a set of characters. We also use real-world data sets from over ten different domains to compare the accuracy and efficiency of the variant algorithms in depth.

2. Cross Validation vs. Overfitting

The goal of feature selection is to choose a subset X_s of the complete set of input features $X = \{x_1, x_2, ... ,, x_M\}$ so that the subset X_s can predict the output Y with accuracy comparable to the performance of the complete input set X, and with great reduction of the computational cost.

First, let us clarify how to evaluate the performance of a set of input features. In this paper we use a very conservative form of feature set evaluation in order to avoid overfitting. This is important. Even if feature sets are evaluated by test-set cross-validation or leave-one-out cross validation, an exhaustive search of possible feature-sets is likely to find a misleadingly well-scoring feature-set by chance. To prevent this, we use the cascaded cross-validation procedure in Figure 1, which selects from increasingly large sets of features (and thus from increasingly large model classes). The score for the best feature set of a given size is computed by an independent cross-validation from the score for the best size of feature set.

Two notes about the procedure in Figure 1: First, the choice of 70/30 split for training and testing is somewhat arbitrary, but is empirically a good practical ratio according to more detailed experiments [Deng and Moore, 98]. Second, note that Figure 1 does not describe how we search for the best feature set of size j in Step 2a. This is the subject of Section 3.

1. Shuffle the data set and split into a training set of 70% of the data and a testset of the remaining 30%.
2. Let j vary among feature-set sizes: $j = (0, 1, 2, \ldots, m)$
   a. Let fs_j = best feature set of size j, where "best" is measured as the minimizer of the leave-one-out cross-validation error over the training set.
   b. Let $Testscore_j$ = the RMS prediction error of feature set fs_j on the test set.

   End of loop of (j).
3. Select the feature set fs_j for which the test-set score is minimized.

Figure. 1 Cascaded Cross-validation procedure for finding the best set of up to m features.

Evaluating the Feature Selection algorithms.
In our experiments, to evaluate the performance of a feature selection algorithm on a dataset we perform an additional outer layer of cross-validation (20-fold CV). This gives us a completely fair estimate of how well the feature selection algorithm performs. The full procedure, including the outer and inner parts, is given in Figure 2.

3. Feature selection algorithms

Forward feature selection

The forward feature selection procedure begins by evaluating all feature subsets which consist of only one input attribute. In other words, we start by measuring the Leave-One-Out Cross Validation (LOOCV) error of the one-component subsets, $\{X_1\}, \{X_2\}, \ldots, \{X_M\}$, where M is the input dimensionality; so that we can find the best individual feature, $X_{(1)}$.

Next, forward selection finds the best subset consisting of two components, $X_{(1)}$ and one other feature from the remaining $M - 1$ input attributes. Hence, there are a total of $M - 1$ pairs. Let's assume $X_{(2)}$ is the other attribute in the best pair besides $X_{(1)}$.

Afterwards, the input subsets with three, four, and more features are evaluated. According to forward selection, the best subset with m features is the m-tuple consisting of $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$, while overall the best feature set is the winner out of all the M steps. Assuming the cost of a LOOCV evaluation with i features is $C(i)$, then the computational cost of forward selection searching for a feature subset of size m out of M total input attributes will be

1. Collect a training data set from the specific domain.
2. Shuffle the data set.
3. Break it into *P* partitions, (say *P = 20*)
4. For each partition (*i = 0, 1, ..., P-1*)
   a. Let *OuterTrainset(i)* = all partitions except *i*.
   b. Let *OuterTestset(i)* = the *i*'th partition
   c. Let *InnerTrain(i)* = randomly chosen 70% of the *OuterTrainset(i)*.
   d. Let *InnerTest(i)* = the remaining 30% of the *OuterTrainset(i)*.
   e. For *j = 0, 1, ..., m*
      Search for the best feature set with *j* components, fs_{ij}.using leave-one-out on *InnerTrain(i)*
      Let $InnerTestScore_{ij}$ = RMS score of fs_{ij} on *InnerTest(i)*.
      End loop of *(j)*.
   f. Select the fs_{ij} with the best inner test score.
   g. Let $OuterScore_i$ = RMS score of the selected feature set on *OuterTrainset(i)*

   End of loop of *(i)*.
5. Return the mean Outer Score.

Figure. 2 Full procedure for evaluating feature selection of up to *m* attributes.

$$MC(1) + (M-1)C(2) + ... + (M-m+1)C(m)$$

For example, the cost of one prediction with one-nearest-neighbor as the function approximator, using a *k*d-tree with *j* inputs, is *O(j log N)* where *N* is the number of datapoints. Thus, the cost of computing the mean leave-one-out error, which involves *N* predictions, is *O(j N log N)*. And so the full cost of feature selection using the above formula is $O(m^2 M N \log N)$.

To find the overall best input feature set, we can also employ exhaustive search. Exhaustive search begins with searching the best one-component subset of the input features, which is the same in the forward selection algorithm; then it goes to find the best two-component feature subset which may consist of *any* pairs of the input features. Afterwards, it moves to find the best triple out of all the combinations of any three input features, etc. It is straightforward to see that the cost of exhaustive search is the following:

$$MC(1) + \binom{M}{2}C(2) + ... + \binom{M}{m}C(m)$$

Compared with the exhaustive search, forward selection is much cheaper.

However, forward selection may suffer because of its greediness. For example, if $X_{(1)}$ is the best individual feature, it doesn't guarantee that either $\{X_{(1)}, X_{(2)}\}$ or

$\{X_{(1)}, X_{(3)}\}$ must be better than $\{X_{(2)}, X_{(3)}\}$. Therefore, a forward selection algorithm may select a feature set different from that selected by exhaustive searching. With a bad selection of the input features, the prediction $\hat{Y}_q$ of a query $X_q = \{x_1, x_2, \ldots, x_M\}$, may be significantly different from the true Y_q.

Three Variants of Forward Selection

In this paper, we will investigate the following two questions based on empirical analysis using real world datasets mixed with artificially designed features.

1. How severely does the greediness of forward selection lead to a bad selection of the input features?

2. If the greediness of forward selection doesn't have a significantly negative effect on accuracy, how can we modify forward selection algorithm to be greedier in order to improve the efficiency even further?

We postpone the first question until the next section. In this paper, we propose three greedier feature selection algorithms whose goal is to select no more than *m* features from a total of *M* input attributes, and with tolerable loss of prediction accuracy.

Super Greedy Algorithm

Do all the 1-attribute LOOCV calculations, sort the individual features according to their LOOCV mean error, then take the *m* best features as the selected subset. We thus do *M* computations involving one feature and one computation involving *m* features. If nearest neighbor is the function approximator, the cost of super greedy algorithm is $O((M + m) N \log N)$.

Greedy Algorithm

Do all the 1-attribute LOOCVs and sort them, take the best two individual features and evaluate their LOOCV error, then take the best three individual features, and so on, until *m* features have been evaluated. Compared with the super greedy algorithm, this algorithm may conclude at a subset whose size is smaller than *m* but whose inner testset error is smaller than that of the *m*-component feature set. Hence, the greedy algorithm may end up with a better feature set than the super-greedy one does. The cost of the greedy algorithm for nearest neighbor is $O((M + m^2) N \log N)$.

Restricted Forward Selection (RFS)

1. Calculate all the 1-feature set LOOCV errors, and sort the features according to the corresponding LOOCV errors. Suppose the features ranking from the most important to the most independent are $x_{(1)}, x_{(2)}, \ldots, x_{(M)}$.

2. Do the LOOCVs of 2-feature subsets which consist of the winner of the first round, $X_{(1)}$, along with another feature, either $X_{(2)}$, or $X_{(3)}$, or any other one until $X_{(M/2)}$. There are $M/2$ of these pairs. The winner of this round will be the best 2-component feature subset chosen by RFS.

3. Calculate the LOOCV errors of $M/3$ subsets which consist of the winner of the second round, along with the other $M/3$ features at the top of the remaining rank. In this way, RFS will select its best feature triple.

4. Continue this procedure, until RFS has found the best *m*-component feature set.

5. From Step 1 to Step 4, RFS has found *m* feature sets whose sizes range from *1* to *m*. By comparing their LOOCV errors, RFS can find the best overall feature set.

The difference between RFS and conventional Forward Selection (FS) is that at each step to insert an additional feature into the subset, FS considers all the remaining features, while RFS only tries a part of them which seem more promising. The cost of RFS for nearest neighbor is $O(M\ m\ N \log N)$.

For all these varieties of forward selection, we want to know how cheap and how accurate they are compared with the conventional forward selection method. To answer these questions, we resort to experiments using real world datasets.

4. Experiment 1: Kanji Handwriting Recognition

Our goals in applying feature selection to Chinese and Japanese handwriting recognition are: (1) To demonstrate feature selection is important because it is a crucial part for an important application. (2) To compare the feature set found by the feature selection algorithms with a human expert's selection.

Feature selection for Kanji recognition

Although most of the research in handwriting recognition is for on-line systems [Singer et al., 94], there is no doubt that off-line systems are also very important especially in domains such as automatic tax form processing.

To date, the research for Chinese and Japanese character recognition is still preliminary[1]. Because the number of Kanji, i.e. Chinese characters, is over fifty thousand, it is hard to rely on any general-purpose global model to recognize all Chinese characters. Alternatively, a promising approach is to separate the Chinese characters into several groups. For each group, a local model is developed to distinguish the different characters.

Although it may be possible to build the local models off-line, manually, it is better if we have an on-line automatic configuration mechanism. Not only does this automatic system save software developers from tedious and time-consuming work, but also it is adaptive and can learn different personal handwriting styles.

In this section, we propose an idea to recognize Chinese and Japanese handwriting off-line, with automatically configured adaptive local models. We also give a prototype of this system.

Chinese characters are constructed by ten fundamental strokes.

The different combinations with different relative positioning determine different characters. For example, there are eight different Chinese characters plus “F” and the Japanese character “ka” containing two horizontal lines and one vertical line, illustrated in Figure. 3.

In this prototype system, some features are useful for recognition, while others may not be so significant, or, can be substituted, referring to Figure. 4. Notice: (1) The human expert’s selection, as shown in Figure. 4(a), is not the only functional

1. There are some Chinese and Japanese recognition products on the market. The product introductions claim that their accuracy is over 90%. However, we don’t know what kind of principles they apply. And we notice some of those products can only recognize rigidly written characters.

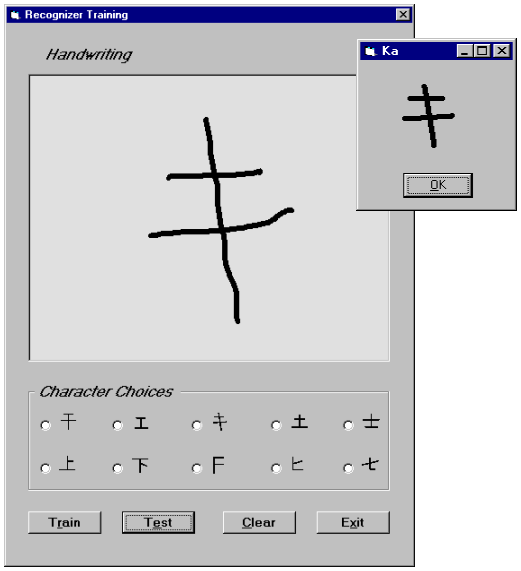
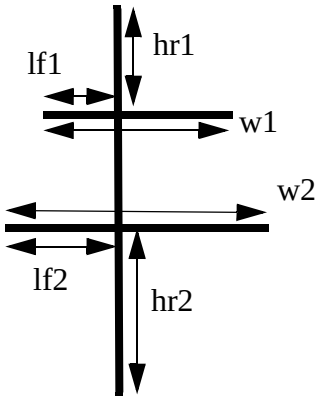
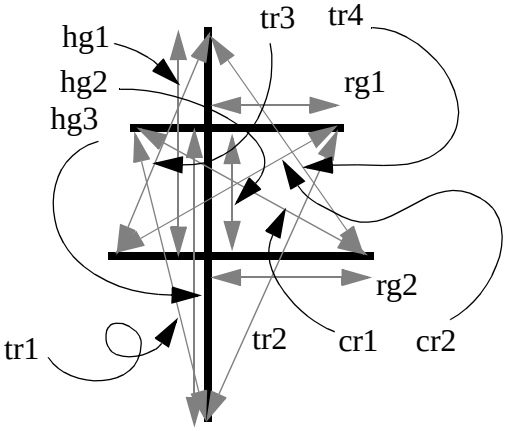


Figure. 3 A prototype of Kanji handwriting recognition system.



(a) Features selected by human expert



(b) Other candidate features, including those not-important

Figure. 4 The features used for the Kanji recognition prototype system.

set, there exist multiple options. (2) Among the multiple feature sets, some of them may lead to more accurate recognition than the others.

To find the features including those not-so-significant, we can follow these three steps:

- Figure out the horizontal lines, vertical lines, and other strokes, respectively.
- Sort the lines from top to bottom, or from left to right.
- Calculate all the possible features according to prior knowledge.

After we have found the candidate features, we can apply the various feature selection algorithms to select the proper features for the recognition job. In the experiment, we try four feature selection algorithms: Super-greedy (Super), Greedy (Greedy), Restricted Forward Selection (RFS) and conventional Forward Selection (FS). We request that any selected feature sets contain no more than eight components. To evaluate the goodness of the selected feature sets, we calculate their 20-fold scores. Since our procedure is carefully designed to avoid overfitting, the smaller a feature set's score is, the more accurately this feature set is able to recognize any one out of the ten characters. We also count the numbers of seconds consumed by the four algorithms so as to compare their computational costs.

Table 1. Kanji feature selection

Selection Methods	**m = 8 = Max Number of features**			**m = 12**	
	Selected feature set	**20fold score**	**Cost**	**20fold score**	**Cost**
Super	w2, lf1, lf2, hg1, tr1, tr2, tr3, tr4	0.038	532	0.018	529
Greedy	w2, lf1, lf2, hg1, tr2, tr3, tr4	0.041	767	0.022	916
RFS	hr1, w1, lf1, lf2, hg1, hg3, cr2, tr1	0.018	1414	0.016	1570
FS	hr1, hr2, w1, lf1, lf2, hg1, tr3	0.016	3586	0.018	4829
Human	hr1, hr2, w1, w2, lf1, lf2	0.016	--	0.016	--

In Table1, we observe that different selection algorithms may find different sets of features. When we carefully study these various sets with respect to Fig. 3, we find all of them are functional. Second, we find that the feature sets selected by RFS and FS are very similar to the human expert's preference, but different from the sets found by Super and Greedy. Third, although all of these feature sets have satisfactory accuracy, those found by the greedier algorithms lead to less accurate recognition performance. However, if we allow more components to enter the feature

sets, even the greedier algorithms' selections become more powerful. Finally, the greedier algorithms are cheaper than the others.

Future work

The prototype system is sufficient to demonstrate the importance and capability of the feature selection algorithms. But to pursue a good Kanji handwriting recognition system, some further work has to be done. Since this topic is a digression from the discussion of feature selection, we only give a brief introduction.

For more complicated kanji, for example 藏 which means "Tibet", the number of possible features will explode. Fortunately, every Chinese character can be split into some standard particles, and the number of these standard particles is no more than one hundred. Indexed by these particles and their relative positioning, any Chinese character can be represented by no more than five digits. One example is illustrated in Figure. 5. This technique is called Wang-coding or Five-stroke coding, which has become one of the national standard typing methods in China.

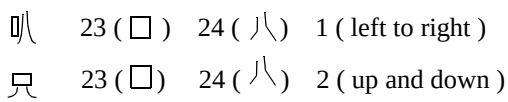


Figure. 5 . An illustration of Wang-coding of a Chinese character.

Now the remaining difficulty is how to find those standard particles from any Chinese characters. One promising approach is A^* search.

5. Experiment 2: Multiple Domains

In the last section, we demonstrated the importance and capability of feature selection. In this section, we compare the greedy algorithms with the conventional methods empirically. We run ten experiments; for each experiment, we try two datasets with different input dimensionalities; and for each dataset, we use three different function approximators.

Experimental design and results

To evaluate the influence of the greediness on the accuracy and efficiency of the feature selection process, we use twelve real world datasets from StatLib/CMU and UCI's machine learning data repository. These datasets come from different domains, such as biology, sociology, robotics, etc. The datasets each contain 62 to

1601 points, and each point consists of an input vector and a scalar output. The dimensionality of the input varies from 3 to 13. In all of these examples we set *m* (the maximum feature set size) to be 10.

Our first experiment demonstrates that Exhaustive Search (ES) is prohibitively time-consuming. We choose four domains with not-too-large datasets and limited input dimensionality for this test. Even for these easy cases, ES is far more expensive than the Forward Selection algorithm (FS), while it is not significantly more accurate than FS. However, the features selected by FS may differ from the result of ES. That is because some of the input features are not mutually independent.

Our second experiment investigates the influence of greediness. We compare the three greedier algorithms, Super Greedy, Greedy and Restricted Forward Selection (RFS), with the conventional FS in three aspects:(1) The probabilities for these algorithms to select any useless features, (2) The prediction errors using the feature set selected by these algorithms, and (3) The time cost for these algorithms to find their feature sets.

Table 2. Preliminary comparison of ES vs. FS

Domain (dim)	**20Fold Mean Errors**			**Time Cost**			**Selected Features**	
	ES	**FS**	**ES / FS**	**ES**	**FS**	**ES / FS**	**ES**	**FS**
Crab (7)	0.415	0.469	0.885	35644	522	68.28	A,F,G	A,E
Halibut (7)	57.972	52.267	1.109	61759	713	86.62	B,C,G	A,D,E,G
Irish (5)	0.863	0.905	0.954	138088	1142	120.91	A,C,E	A,D
Litter (3)	0.780	0.868	0.899	4982	117	42.58	A,B,C	A,B,C

For example, if a raw data file consists of three input attributes, *U*, *V*, *W* and an output *Y*. We generate a new dataset consisting of more input features, *U*, *V*, *W*, *cU*, *cV*, *cW*, R_1, R_2,..., R_{10}, and the output *Y*, in which *cU*, *cV* and *cW* are copies of *U*, *V* and *W* but corrupted with 20% noise, while R_1 to R_{10} are independent random numbers. The chance that any of these useless features is selected can be treated as an estimation of the probability for the certain feature selection algorithm to make a mistake.

Table 3. Greediness comparison

Domain (dim)	Funct. Apprx.	# Corrupt / Total Corrupts				# Noise / Total Noise			
		Super	Greedy	RFS	FS	Super	Greedy	RFS	FS
Bodyfat (13)	Nearest	0.23	0.12	0.10	0.12	0.10	0.05	0.05	0.06
	LocLin	0.31	0.08	0.17	0.18	0.00	0.00	0.05	0.20
	GlbLin	0.31	0.23	0.15	0.00	0.00	0.00	0.00	0.40
Boston (13)	Nearest	0.23	0.19	0.21	0.17	0.20	0.20	0.23	0.35
	LocLin	0.15	0.15	0.12	0.15	0.30	0.30	0.30	0.33
	GlbLin	0.15	0.12	0.15	0.23	0.40	0.30	0.30	0.40
:	(10 other results omitted for reasons of space)							:	
Mean over all twelve datasets	Nearest	0.37	0.27	0.17	0.21	0.23	0.10	0.11	0.13
	LocLin	0.38	0.18	0.17	0.20	0.24	0.13	0.13	0.18
	GlbLin	0.30	0.26	0.19	0.23	0.29	0.18	0.21	0.28
TOTAL	-	0.35	0.24	0.18	0.21	0.25	0.14	0.15	0.20

As we observe in Table 3, FS doesn't eliminate more useless features than the greedier competitors except the Super Greedy one. However, the greedier an algorithm is, the more easily it is confused by the relevant but corrupted features.

Since the input features may be mutually dependent, the different algorithms may find different feature sets. To measure the goodness of these selected feature sets, we calculate the mean 20fold score. As described in Section 2, our scoring is carefully designed to avoid overfitting, so that the smaller the score, the better the corresponding feature set is. To confirm the consistency, we test the four algorithms in all the twelve domains from StatLib and UCI. For each domain, we apply the algorithms to two datasets. Both of the datasets are generated based on the same raw data file, but with different numbers of corrupted features and independent noise. And for each dataset, we try three function approximators, nearest neighborhood (Nearest), locally weighted linear regression (LocLin) and global linear regression (GlbLin). For the sake of conciseness, we only list the ratios. If a ratio is close to 1.0, the corresponding algorithm's performance is not significant from that of FS. The experimental results are shown in Table 4. Besides, we also list the ratios of the number of seconds consumed by the greedier algorithms to that of FS.

Table 4. Greediness comparison

Domain (dim)	Funct. Apprx.	20Fold() / 20Fold(FS)			Cost() / Cost(FS)		
		Super	Greedy	RFS	Super	Greedy	RFS
Bodyfat (13)	Nearest	0.975	0.969	0.915	0.095	0.126	0.330
	LocLin	1.080	1.015	0.973	0.062	0.092	0.287
	GlbLin	0.984	0.981	0.966	0.084	0.109	0.247
Boston (13)	Nearest	0.876	0.872	0.881	0.105	0.145	0.389
	LocLin	1.091	1.091	0.969	0.058	0.080	0.270
	GlbLin	1.059	1.052	1.068	0.084	0.127	0.287
:	(10 other results omitted for reasons of space)					:	
Mean over all twelve datasets	Nearest	1.142	1.001	0.978	0.122	0.163	0.365
	LocLin	1.196	1.064	1.011	0.077	0.115	0.296
	GlbLin	1.029	1.025	0.995	0.091	0.138	0.301
TOTAL	-	1.122	1.030	0.995	0.097	0.138	0.321

First, we observe in Table 4 that the three greedier feature selection algorithms don't suffer great loss in accuracy, since the average ratios of the 20fold scores to those of FS are very close to 1.0. In fact, RFS performs almost as well as FS. Second, as we expected, the greedier algorithms improve the efficiency. Super greedy algorithm (Super) is ten times faster than forward selection (FS), while greedy algorithm (Greedy) seven times, and the restricted forward selection (RFS) three times. Finally, restricted forward selection (RFS) performs better than the conventional FS in all aspects.

To further confirm our conclusion, we do the third experiment. This time, we insert more independent random noise and corrupted features to the datasets. For example, if the original data set consists of three input features, $\{U,V,W\}$, the new artificial data file contains $\{U, cU, V, cV, cU * cV, W, cW, cV * cW, R_1,..., R_{40}\}$. The results are listed in Table 5 and Table 6.

Table 5. Greediness comparison with more inputs

	Funct. Apprx.	# Corrupt / Total Corrupts				# Noise / Total Noise			
		Super	Greedy	RFS	FS	Super	Greedy	RFS	FS
Mean Values	Nearest	0.29	0.33	0.30	0.38	0.04	0.04	0.03	0.04
	LocLin	0.38	0.38	0.25	0.41	0.05	0.03	0.02	0.03
	GlbLin	0.38	0.25	0.29	0.16	0.05	0.05	0.08	0.07
TOTAL	-	0.35	0.32	0.28	0.32	0.05	0.04	0.04	0.05

Table 6. Greediness comparison with more inputs

	Funct. Apprx.	20Fold( ) / 20Fold(FS)			Cost( ) / Cost(FS)		
		Super	Greedy	RFS	Super	Greedy	RFS
Mean Values	Nearest	1.197	1.056	1.001	0.080	0.080	0.282
	LocLin	1.202	1.059	1.040	0.071	0.084	0.281
	GlbLin	1.032	1.026	0.998	0.079	0.104	0.294
TOTAL	-	1.144	1.047	1.013	0.077	0.088	0.286

Comparing Table 3 with Table 5, we notice that with more input features, the probability for any corrupted feature to be selected remains almost the same, while that of independent noise reduces greatly. Comparing Table 4 with Table 6, with more input features, (1) the prediction accuracies of the feature sets selected by the variety of the algorithms are roughly consistent, because the 20fold scores in the two tables are almost the same; (2) the efficiency ratio of the greedier alternatives to FS is a little higher.

In summary, in theory the greediness of feature selection algorithms may lead to great reduction in the accuracy of function approximating, but in practice it doesn't happen quite often. The three greedier algorithms we propose in this paper improve the efficiency of the forward selection algorithm, especially for larger datasets with high input dimensionalities, without significant loss in accuracy. Even in the case the accuracy is more crucial than the efficiency, restricted forward selection is more competitive than the conventional forward selection.

6. Conclusion

In this paper, we propose three greedier variants of the forward selection method. Our investigation shows that the greediness of the feature selection algorithms greatly improves the efficiency, while does not corrupt the correctness of the selected feature set so that the prediction accuracy using the selected features remains satisfactory. As an application, we apply feature selection to a prototype system of Chinese and Japanese handwriting recognition.

7. Acknowledgments

Thanks to Garth Zeglin, Jeff Schneider and Scott Davies for their constructive comments and suggestions.

References

[Akaike, 73] Akaike, 1973. Information theory and an extension of the maximum likelihood principle. 2'nd International Symposium on Information Theory, B.N.Petrov and F. Csaki (eds.), Akademiai Kiado, Budapest, pp 267-281.

[Almuallim, et al., 91] Almuallim, H.; and Dietterich, T.G. 1991. Learning with many irrelevant features. In Proc. AAAI-91, pp 547-552. MIT Press.

[Caruana et al., 94] Caruana, R.; and Freitag, D. 1994. Greedy attribute selection. In Proc. ML-94. Morgan Kaufmann.

[Deng and Moore, 98] Deng, K. and Moore, A. W., 1998, Intense Cross Validation for Large Datasets, *In preparation, 1998*

[Heckerman, et al., 96] Heckerman, D.; and Chickering D.M., 1996. A comparison of scientific and engineering criteria for Bayesian model selection. Technical Report, MSR-TR-96-12, http://research.microsoft.com/research/dtg/heckerma/TR-96-12.htm

[John, et al., 94] John, G.; Kohavi, R.; and Pfleger, K., 1994. Irrelevant features and the subset selection problem. In Proc. ML-94, pp 121-129. Morgan Kaufmann.

[Kira, et al., 92] Kira, K.; and Rendell, L.A., 1992. The feature selection problem: Traditional methods and a new algorithm. In Proc. AAAI-92, pp 129-134. MIT Press.

[Koller, et al., 92] Koller, D.; and Shami, M., 1996. Toward optimal feature selection. In Proc. ML-96. Morgan Kaufmann.

[Langley et al., 94] Langley, P.; and Sage, S., 1994. Induction of selective bayesian classifiers. In Proc. UAI-94, pp 399-406. Seattle, WA: Morgan Kaufmann.

[Miller, 90] Miller, A.J., 1990. Subset selection in regression. Chapman an Hall, 1990.

[Moore et al., 94] Moore A.W; and Lee, M.S., 1994. Efficient algorithms for minimizing cross validation error. In Proc. ML-94. Morgan Kaufmann.

[Moore et al., 97] Moore, A.W., Schneider, J., and Deng, K. Efficient Locally Weighted Polynomial Regression. In Proc ICML-97, Morgan Kaufmann.

[Skalak, 94] Skalak, D.B., 1994. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In Proc. ML-94. Morgan Kaufmann.

[Singer et al., 94] Singer, Y.; Tishby, N., 1994. Dynamic encoding of cursive handwriting. Biological Cybernetics, 71(3), 1994. Springer-Verlag.