Fitness Approximation in Estimation of Distribution Algorithms for Feature Selection

Haixia Chen[1], Senmiao Yuan[1], and Kai Jiang[2]

[1] College of Computer Science and Technology, Jilin University,
Changchun 130025, China
hxchen2004@sohu.com
[2] The 45th Research Institute of CETC, Beijing 101601, China
kjiang2004@sohu.com

Abstract. Estimation of distribution algorithms (EDAs) are popular and robust algorithms that combine two technical disciplines of soft computing methodologies, probabilistic reasoning and evolutionary computing, for optimization problems. Several algorithms have already been proposed by different authors. However, these algorithms may require huge computation power, which is seldom considered in those applications. This paper introduces a "fast estimation of distribution algorithm" (FEDA) for feature selection that does not evaluate all new individuals by actual fitness function, thus reducing the computational cost and improve the performance. Bayesian networks are used to model the probabilistic distribution and generate new individuals in the optimization process. Moreover, fitness value is assigned to each new individual using the extended Bayesian network as an approximate model to fitness function. Implementation issues such as individual control strategy, model management are addressed. Promising results are achieved in experiments on 5 UCI datasets. The results indicate that, as population-sizing requirements for building appropriate models of promising solutions lead to good fitness estimates, more compact feature subsets that give more accurate result can be found.

1 Introduction

Feature selection (FS) is one of the most important issues in the community of data mining, machine learning, pattern recognition, etc[1]. There are two basic approaches to feature selection: wrapper approaches and filter approaches[2]. While wrappers give better results in terms of the accuracy of the final classifier, being a NP-hard problem, the selection process becomes more complex with the number of features and instances in the given task increasing.

Estimation of distribution algorithms (EDAs) are a quite recent topic in optimization techniques. Using different assumption of the joint probability distribution, different algorithms have been proposed and good results have been observed [1], [3], [4]. In the wrappers for feature selection optimization, the time to run EDAs is dominated by the 'slow-to-compute' fitness function evaluation. To compound the problem further, it is often necessary for EDAs to select a large population size for distribution estimation and use a large number of generations to obtain an acceptable solution and

S. Zhang and R. Jarvis (Eds.): AI 2005, LNAI 3809, pp. 904 – 909, 2005.

avoid premature convergence. For evolutionary algorithms models, there are two main ways to reduce the computational cost by integrating approximate models that exploit knowledge of past evaluation into the optimization: evolution control and surrogate approach [5]. As it is difficult to construct an approximate model that is globally correct due to the high dimensionality, ill distribution and limited number of training samples, it is found that the surrogate approach is likely to converge to a false optimum, which is an optimum of the approximate model, but not the one for the actual fitness function. Therefore, the evolution control approach is of more practical importance.

This paper introduces a "fast estimation of distribution algorithm" (FEDA) to deal with the computational overburden comes along with the wrapper approach for feature selection. It uses Bayesian Networks (BNs) to estimate the probability distribution of each generation. In addition, the BNs are extended as approximate models to assign fitness values. As those assigned fitness values are not the actual fitness values, the individual control strategy and model management strategy are proposed to find those informative individuals with high fitness values or in an unexplored region. The main aim of the approximate model is not only to assign fitness but also to find informative individuals for updating itself.

2 The FEDA for Feature Selection

The framework for FEDA can be summarized as follows. First, a group of individuals are initialized randomly. Then, individuals are selected according to the individual control strategy for actual evaluation and added into the population by steady-state strategy. In the first generation, as there is no model for fitness estimation, all individuals are controlled. Third, a Bayesian network are build from the population, fitness statistics of the selected population are collected according to the Bayesian network structure and used to extend the model. Last, new individuals are generated by sampling from the Bayesian network and sent to the approximate model for fitness assignment. The process terminates if the stopping criterion is meet. Otherwise, it goes to the second step for the next iteration.

Given dataset ***D*** with *d* features: $X = \{X_1, X_2, \cdots, X_d\}$. The purpose of the learning system is to induce a classifier *c* in the hypothesis space ***H*** that can best describe the dataset. Each individual in the search can be represented by a binary string of *d* bits, $s = \{s_i\}, s_i \in \{0,1\}, i = 1,...,d$, with each bit indicating whether a feature is present (1) or absent (0). And the aim of the wrapper is to select a feature subset s in the feature subset space $\boldsymbol{S}$ that satisfies $s = \arg\max_{c\in H, s'\in S}(P(c,s',\boldsymbol{D}))$. Where $P(c,s,\boldsymbol{D})$ measures the performance of the classifier *c* with a feature subset *s* on the dataset ***D***, and comprises the accuracy measure [1] and the 'parsimony' measure that indicates how many features and data acquirement and preparation cost we have managed to save.

Estimation of Distribution Algorithms (EDAs) replace the crossover and mutation operators in GAs by building a probabilistic model and sampling from the model. Bayesian networks are used in this paper as the probabilistic models. And the traditional Bayesian networks are extended to act as approximate models for fitness approximation so as to avoid additional cost in model building and to exploit the complicated expressive power of the Bayesian networks.

For every variable S_i of the population and each possible value s_i of the variable, let $f_c(s_i \mid \pi_i)$ denote the contribution of S_i at value s_i to the total fitness.

$$f_c(s_i \mid \pi_i) = f(s_i \mid \pi_i) - \sum_{s_i} p(s_i \mid \pi_i) f(s_i \mid \pi_i) . \tag{1}$$

Where $\boldsymbol{\Pi}_i$. is the set of parent variables (nodes) that S_i. has in the BN and π_i is its possible instantiation. $f(s_i \mid \pi_i)$ is average fitness of individuals with $S_i = s_i$ and $\boldsymbol{\Pi}_i = \pi_i$. The fitness of an individual can be estimated as the sum of the average fitness and the contribution of every bit:

$$f_{est}(s) = \overline{f} + \sum_{i=1}^{d} f_c(s_i \mid \pi_i) . \tag{2}$$

To accommodate this extension, the conditional probability table of BN is modified by adding two additional entries that represent $f_c(s_i \mid \pi_i)$ to each row.

2.1 Individual Control Strategy

The first problem comes along with FEDA is how should individuals be evaluated by the actual fitness function. There are two possibilities to combine the actual fitness function with the approximate fitness function. One is individual-based evolution control in which a certain number of individuals within a generation are evaluated with the actual fitness function. Another is generation-based evolution control, which means in every T generations, T' ($T' < T$) generations will be actually evaluated. We prefer individual control in FEDA because EDA evolves with a large population, and we want to find the near optimum as soon as possible. The problem now is to determine which and how many individuals should be evaluated/controlled by the actual fitness function. To this end, a strategy with four units is proposed.

1. Model Fidelity Test. The model fidelity factor is defined as:

$$\alpha = L_{\max} / L_t(BN) . \tag{3}$$

Where $L_{\max}$ is the maximum log likelihood of BN to the population. $L_t(BN)$ is the log likelihood for the current model. t denotes the number of iteration. $0 < \alpha \le 1$.

2. Exploration. For each new sampled individual s, the probability for sampling it $\hat{p}(s)$ is calculated according to the current BN. If $\hat{p}(s) < 1/|\boldsymbol{S}|$, then the individual is put into the exploration unit, otherwise it is sent to the exploitation unit that will be addressed below. For individual that is put into the exploration unit, its approximate fitness $\hat{f}_t(s)$ is calculated according to formula 2 and compared to the average fitness $\overline{f}_t$. If $\hat{f}_t(s) < \overline{f}_t$, then the individual is not controlled. Otherwise, the individual is evaluated by the actual fitness function.

3. Exploitation. For individuals in the exploitation unit, the best β percent is controlled. We call β the exploitation factor.

4. Random Control. For the rest individuals in the exploitation unit, random numbers between 0 and 1 are obtained. If the number is higher than $\min(\alpha, \eta)$, then the individual is truly evaluated. η is the random factor that is less than 1.

The strategy is designed to find the unexplored region and the promising region. Individuals with low sampling probability imply two cases. For one thing, the individual is really not so good and should be eliminated from the evolutionary process. For another, the individual is in fact a good one but not in the current search region. The exploration unit tries to find the latter one. The approximate model gives a consistent answer indicating a really bad individual when it assigns a low approximate fitness to an individual with low probability. However, when an individual with low probability is assigned a high fitness, we may conclude that the approximate model itself doesn't confirm its decision. So the individual should be controlled. For individuals with high probability, as the current population gives enough information for model building, the approximate fitness is sound to a great degree and only a small percentage of individuals with highest approximate fitness should be actually evaluated. However, for small percentage value, the fitness values of most individuals in the population will be estimated. Such an approach might affect the selection process and cause the population converges to a sub-optimum point. To make up for this, the random control unit demands at least some additional true evaluations will occur to every generation according on the fidelity of the model.

2.2 Model Management Strategy

Another problem comes along with FEDA is whether all or just the fraction of actually evaluated individuals should be used for model update. Due to the high dimensionality, ill distribution and limited number of training samples, it is very difficult to construct an approximate model that is globally correct. In fact, it is of more practical importance to build an approximate model that represents the promising individuals step by step. The individual control strategy described above can be viewed as a good filter for actively selecting the most informative individuals for model update.

With no prior knowledge, the evolutionary process often starts with a population generated according to uniform distribution. So the approximate model is often of poor quality at the beginning of the search. With more promising individuals are found by exploration and exploitation, the model quality also improves as the evolutionary search proceeds. Thus, another general rule for model management is that the model should be updated with more controlled individuals at the beginning. After a certain number of generations, the model becomes much more reliable and the control frequency can reduce. The individuals that will be controlled in each generation are determined by three factors, α, β and η. α is determined online by the performance of the model. η is used jointly with α to ensure at least some individuals are controlled. Only β can be controlled easily. We let it decreases with the evolutionary process. Those controlled individuals replace the ones with the lowest fitness values in the father population. In this way, only individuals that are thought as informative, promising are introduced into the individual pool.

3 Experiments

Experiments were carried out on the German, soybean, chess, anneal and mushroom datasets from UCI repository [7]. We used a population with 2000 individuals. The search finished when no improvement was found in two consecutive generations or when a maximum of 20 generations was reached. Exploitation factor β decreased from 0.1 to 0.05 with step length 0.01, and kept constant in the following generations. Random factor $\eta = 0.95$. For each individual, a naive Bayes (NB) classifier was constructed. We compared FEDA with FSS-EBNA[1].

Fig. 1 shows the best fitness value as a function of the number of true evaluations for German dataset. Similar results were observed for the other datasets. It is noted that FEDA can achieve better results with less actual evaluation than FSS-EBNA. About 60% actual evaluations are saved by introduction of the approximate model. Furthermore, FEDA can find better results than FSS-EBNA. It seems contradictory as FSS-EBNA takes more actual evaluations. We think the individual control strategy and model management strategy can answer for this. In FSS-EBNA all the new generated individuals are evaluated by the actual fitness function and enter into the next generation. However, in FEDA, only individuals that meet the individual control strategy are controlled. The model fidelity test, exploration, exploitation, and random control units in the individual control strategy help to find informative individuals. That is, the strategy actively finds those individuals that help most. It is like a filter that rejects those useless individuals from model building. As we can only build a local model in practice, the active learning implied by the strategy helps to find better models. It is also noted that we take a pessimistic termination criterion in the compare. As FEDA takes less actual evaluations, we hope it can evolve more and find better results given the same time as FSS-EBNA.

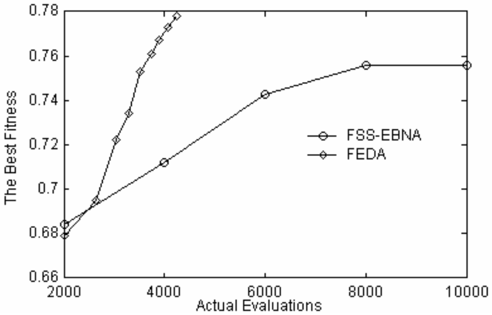

Fig. 1. The best fitness value as a function of the number of actual evaluations for FEDA and FSS-EBNA

Fig. 2 illustrates the accuracy and size of the final optimal feature subset for the five datasets averaged over 30 runs. In each run, the performance of the algorithm is determined by five-fold cross validation on the whole dataset. A more compact subset with higher accuracy is more desirable. So points in the right bottom are preferable to the ones in the upper left. The graph clearly shows that with the help of the approximate model, FEDA can find a more compact classifier with higher accuracy.

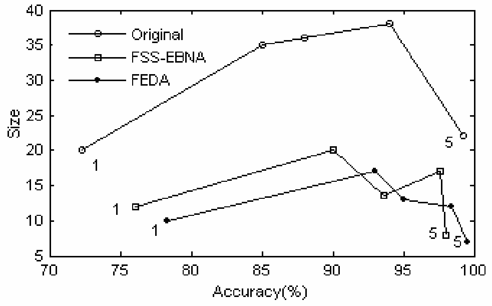

Fig. 2. Performance Comparison between FEDA and FSS-EBNA for the five dataset

4 Summary

This paper presents a fitness approximation strategy in EDA for FS. To exploit all of the cumulated knowledge about the search process, the conditional probability tables of the BNs are extended to incorporate contributions of each bit under different states. Individual control strategy and model management strategy are proposed to find those informative individuals and limit the number of actual fitness evaluations to a minimum. Experimental results show that the algorithm can get a more accurate, more compact subset with less computational cost.

References

1. Inza, I., Larranaga, P.., Etxeberria, R. and Sierra, B.. Feature Subset Selection by Bayesian Network-based Optimization. Artificial Intelligence 123 (2000) 157-184
2. Kohavi, R., John, G. H.. Wrappers for Feature Subset Selection. Artificial Intelligence 97 (1997) 273-324
3. Bengoetxea, E., Larranaga, P., Bloch, I., Perchant, A., Boeres, C.. Inexact Graph Matching by means of Estimation of Distribution Algorithms. Pattern Recognition 35 (2002) 2867-2880
4. Chen, H. X., Yuan, S. M., Jiang, K.. Bayesian Network Optimization Algorithm Based on Holding Strategy. Computer Engineering and Application 14 (2005) 61-65
5. Jin, Y. A Comprehensive Survey of Fitness Approximation in Evolutionary Computation. Soft Computing (2003)
6. Pelikan, M., Sastry, K.. Fitness Inheritance in the Bayesian Optimization Algorithm. IlliGAL Report No. 2004009 (2004)
7. Blake, C. L. and Merz, C. J.. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, Dept. of Information and Computer Science (1998)