

# Recognition of Exon/Intron Boundaries Using Dynamic Ensembles

Xuanfu Wu and Zhengxin Chen  
Department of Computer Science  
University of Nebraska at Omaha  
Omaha, NE 68182-0500  
{xuanfuwu, zchen}@mail.unomaha.edu

## 1. Introduction

Many studies have been carried out in recognition of exon/intron boundaries. For example, PROCUSTES uses similarity-based approach to gene recognition [4]. Other examples include GRAIL (Gene Recognition and Assembly Internet Link) <http://compbio.ornl.gov/Graill-1.3/help/> (1996) and Glimmer (Gene Locator and Interpolated Markov Modeler) [2].

Since the problem of recognition of exon/intron boundaries can be cast as a classification task (e.g., [1]), ensemble learning [3,5] can be applied. An ensemble consists of a set of organized individual trained classifiers whose individual decisions are combined in a certain way for classification purpose. However, existing studies typically take static approaches which hampered flexibility for improved accuracy. To overcome this problem we have proposed the concept of dynamic ensemble and developed a new algorithm, BAGA, which combines bagging and genetic algorithm techniques.

## 2. Dynamic ensembles and BAGA algorithm

The concept of dynamic ensemble is originated and extended from the *overproduce and choose* paradigm proposed by Roli and Giacinto [6] for generating ensembles. As shown in Figure 1, the basic idea of the *overproduce and choose* paradigm is to produce an initial large set of “candidate” classifier ensembles, and then to select the sub-ensemble of classifiers that can be combined to achieve optimal accuracy.

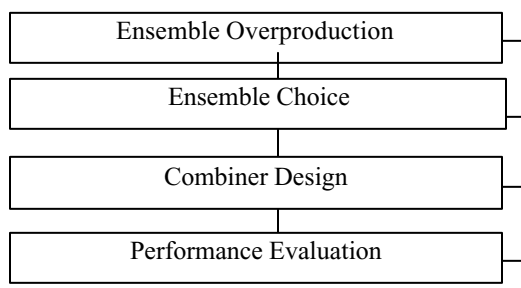


Figure 1. Overproduce and choose paradigm

Extending from the *overproduce and choose* paradigm, in a *dynamic ensemble*, several important factors of such an ensemble are determined at execution time rather than specified in advance (i.e., static, as used in traditional approaches), such as the number of classifiers should be determined at the execution time, the combination function should be adaptive, and, as a more advanced requirement, the structure (i.e., the overall configuration) of ensemble should be dynamic too. BAGA (which stands for BAGging + Genetic Algorithm (GA) is our answer for the first step of achieving dynamic ensembles. Figure 2 presents the skeleton of the basic BAGA Algorithm.

---

**Input:** training set S, Learner L, integer T (number of bootstrap samples), evaluation data set with size m Sv, population size P, maximum generation M, crossover probability Cp, mutation probability Mp

1. for  $i = 1$  to T {
2.      $S' =$  bootstrap sample from S (i.i.d. sample with replacement).
3.      $C_i = L(S')$
4. }
5. generate an initial population with size P of chromosome(CH), where the length of CH is T, and locus of CH is bit string(0/1),  
   for  $j = 1$  to P {
6. evolve the chromosome j, where the fitness function is measured  
    $f(CH_j) =$   
   fitness value of CHj on validate dataset Sv  
   using designed vote strategy  
   }
7. for  $i = 1$  to M {
8. for  $j = 1$  to P{
9.     selection (genetic algorithm operator)
10.    crossover (genetic algorithm operator based on parameter Cp)
11.    mutation (genetic algorithm operator based on parameter Mp)
12.    evolve the chromosome j at ith generation, where the fitness function is measured

$f(\text{CH}_j)(\text{chromosome}) =$   
 fitness value of  $\text{CH}_j$  on validate dataset  $S_v$  using  
 designed vote strategy

13. }
14. update the best fitness chromosome
15. }

Output: classifier  $C^*$  correspond to best fitness chromosome.

**Figure 2. The BAGA Algorithm**

Table 1 indicates several different variants used in the empirical study. In addition to accuracy, other evaluation function such as Compound Diversity (CD) is also used.

**Table 1. Algorithm variants**

Name	Combination Strategies	Evaluation Function
BAG	Simple majority voting	No Genetic Algorithm involved
BAGA_1	Simple majority voting	Based on Accuracy
BAGA_2	Simple majority voting	Based on Diversity (CD)
BAGA_3	Simple majority voting	Based on Accuracy + CD
PBAG (probability + bag)	Probabilistic voting	No Genetic Algorithm involved
BAGA_4	Probabilistic voting	Based on Accuracy
BAGA_5	Probabilistic voting	Based on Diversity (CD)
BAGA_6	Probabilistic voting	Based on Accuracy + CD

#### 4. Empirical studies

We have conducted multiple experiments on different datasets available at UCI Machine Learning Repository and on the Statlog data set at <http://www.liacc.up.pt/ML/statlog/datasets/dna/dna.doc.html>. The DNA dataset was taken from Statlog. The data is concerned with DNA-Primate splice-junction gene sequences, with associated imperfect domain theory. Recall that splice junctions are points on a DNA sequence at which superfluous DNA is removed during the process of protein creation in higher organisms. The problem posed in this dataset is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). The DNA classification problem consists of two subtasks: recognizing exon/intron boundaries (referred to as EI sites), and recognizing intron/exon boundaries (IE sites). Note that IE borders are usually referred to as “acceptors,” while EI borders are referred to as

“donors”. We have conducted various experiments, but due to space limitation, only a comparison of the accuracy rate is shown below (Table 2). Empirical study also indicates that an ensemble with dynamically selected 7 to 8 classifiers can achieve similar or even better results than using all of them (as in BAG).

**Table 2. Accuracy rates grouped by the three classes on DNA dataset**

Algorithm	Accuracy Rate (%)		
	EI class	IE class	neither
See5	89.37	83.96	92.49
BAG	91.55	90.93	93.47
GA1	92.57	91.36	93.43
GA2	95.02	80.11	87.20
GA3	92.64	90.96	93.37
PBAG	90.23	90.86	93.23
GA4	91.35	91.39	93.50
GA5	84.22	87.46	92.95
GA6	92.11	90.43	93.60

More studies on exon/intron boundary will be carried out on larger data sets such as the Exint Database [7].

#### References

- [1] W.-H. Au, K. C. C. Chan and X. Yao, A novel evolutionary data mining algorithm with application to churn prediction, *IEEE Trans. Evolutionary Computation*, 7(6), 532-545, 2003.
- [2] A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER, *Nucleic Acids Research*, 27, 23, 4636-4641, 1999.
- [3] T. G. Dietterich, Ensemble methods in machine learning, *Proc. 1<sup>st</sup> international workshop on multiple classifier systems*, pp.1-15, 2000.
- [4] M. S. Gelfand, A. A. Mironov, P. A. Pevzner (1996), Gene recognition via spliced sequence alignment, *Proc. Natl. Acad. Sci. USA*, 93, pp. 9061-9066
- [5] D. Opitz and R. Maclin, Popular Ensemble Methods: An Empirical Study, *Journal of Artificial Intelligence Research*, Vo. 11, 169-198, 1999.
- [6] F. Roli and G. Giacinto, Design of Multiple Classifier Systems, Chap. 8 in H. Bunke and A. Kandel (Eds.), *Hybrid Methods in Pattern Recognition*, World Scientific Publishing, pp. 199-226, 2002.
- [7] M. Sakharkar, M. Long, T. W. Tan, and S. J. de Souza: ExInt: an Exon/Intron database, *Nucleic Acids Res.* 28 (1), 191–192, 2000.