Ensemble Selection for Evolutionary Learning using Information Theory and Price's Theorem

Stuart W. Card
Syracuse University
7417 S. Main St. P.O. Box 61
Newport NY 13416 USA
01-315-845-6249
cards@ntcnet.com

Chilukuri K. Mohan
EECS Department
Syracuse University
Syracuse NY 13244-4100 USA
01-315-443-2322
ckmohan@syr.edu

ABSTRACT
This paper presents an information theoretic perspective on design and analysis of evolutionary algorithms. Indicators of solution quality are developed and applied not only to individuals but also to ensembles, thereby ensuring information diversity. Price's Theorem is extended to show how joint indicators can drive reproductive sampling rate of potential parental pairings. Heritability of mutual information is identified as a key issue.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *genetic programming;* H.1.1 [**Models and Principles**]: Systems and Information Theory.

General Terms
Theory, Measurement.

Keywords
evolutionary computation, machine learning, ensemble models, group selection, mate selection, Price's Equation.

1. INTRODUCTION
Information theory is, by definition, the appropriate tool for quantifying entropy flows between organisms and their environment -- the vital essence of evolutionary learning. Rigorous mathematical definitions can be developed for intuitive notions that are traditionally addressed by attempting to find easily computable measures that roughly correspond to intuition. These include notions such as epistasis, crowding and diversity. The great advantage of such a mathematical development is the ability to design operators and evolutionary mechanisms, and evaluate existing ones from an information theoretic perspective. While information theory often has been applied to machine learning [1], it rarely has been applied to evolutionary algorithm analysis. We have developed information theoretic indicators of solution quality that can be applied not only to individuals but also to pairs, ensembles and entire populations, thereby ensuring information diversity.

A well-known result in evolutionary biology is *Price's Theorem*. Typical usage is of a formulation that can be derived from Slatkin's transmission-selection recursion as applied to Holland's canonical model of genetic algorithms; this assumes individual selection for reproduction and random mating [2]. If complementary individuals can be identified, they can be selected, not only for membership in ensemble models, but also for reproduction.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

2. INFORMATION THEORY
Akaike developed An Information Criterion (AIC) based on Kullback and Liebler's definition of information in relation to Fisher's sufficient statistics. Shannon's seminal paper [3] states

> "... we consider the source with the maximum entropy subject to the statistical conditions we wish to retain. The entropy of this source determines the channel capacity which is necessary and sufficient."

Shannon's half-century old paper still has much to offer researchers in evolutionary learning. We consider the following channels: the hidden process, between the observable inputs and outputs; the selection and genetic operators, between parent and offspring generations of the population; and the evolutionary learning process itself, between the environment and the genome.

The *sufficiency* of a model (genetic programming function) $\mathbf{f}_j$ is the extent to which its output data set $\mathbf{Z}_j$ captures the mutual information between the input data set $\mathbf{X}$ and the target output data set $\mathbf{Y}$. An ensemble of m models is *epsilon sufficient iff*

$$\mathrm{I}(\mathbf{Y};\mathbf{Z}^m)/\mathrm{I}(\mathbf{Y};\mathbf{X}) \geq 1-\varepsilon \quad \textbf{(1)}$$

The *necessity* of a model is the fraction of its entropy that contributes to its explanation of the target. As above, we define an ensemble model as *epsilon necessary iff*

$$\mathrm{I}(\mathbf{Y};\mathbf{Z}^m)/\mathrm{H}(\mathbf{Z}^m) \geq 1-\varepsilon \quad \textbf{(2)}$$

The residual entropy in the target that is not explained by an individual model is $\mathrm{H}(\mathbf{Y}|\mathbf{Z}_j)$, the excess entropy in an individual model that does not contribute to its explanation of the target is $\mathrm{H}(\mathbf{Z}_j|\mathbf{Y})$ and their sum is the information theoretic measure of the total error of that individual model. This can be generalized to ensembles *without requiring that we know how to compose the constituent individual models into a single higher order model*. It can be made relative to the total ensemble model and target entropy, and inverted to yield an overall ensemble solution quality index that incorporates both sufficiency and necessity:

$$ {}_{N}\mathrm{I}(\mathbf{X};\mathbf{Y};\mathbf{Z}^{m}) = \mathrm{I}(\mathbf{Y};\mathbf{Z}^{m}) / (\mathrm{I}(\mathbf{Y};\mathbf{X}) + \mathrm{H}(\mathbf{Z}^{m}) - \mathrm{I}(\mathbf{Y};\mathbf{Z}^{m})) \quad (3) $$

3. PRICE'S THEOREM

We adopt the formulation of [2], where: F(j) is the measurement function for the property of interest as exhibited by genotype j; T($j \leftarrow k_1,k_2$) is the transmission function giving the probability that genotype j is produced by parental genotypes k_1 and k_2; p(j) is the frequency of genotype j in the population at the current generation; p(j)' is that frequency at the next generation; w(j) is the reproductive sampling rate of genotype j; and $\varphi(k_1,k_2)$ is the expectation of F() in the offspring of parents k_1 and k_2.

Slatkin's transmission-selection recursion is:

$$ \mathrm{p}(j)' = \sum_{k_1,k_2} (\mathrm{T}(j \leftarrow k_1,k_2) \times \frac{\mathrm{w}(k_1)}{\overline{\mathrm{w}}} \mathrm{p}(k_1) \frac{\mathrm{w}(k_2)}{\overline{\mathrm{w}}} \mathrm{p}(k_2)) \quad (4) $$

This leads to Price's Equation:

$$ \Delta\overline{\mathrm{F}} = Cov(\varphi(k_1,k_2), \frac{\mathrm{w}(k_1)\,\mathrm{w}(k_2)}{\overline{\mathrm{w}}^{2}}) \quad (5) $$

If we change our assumptions, from individual reproductive selection and random mating, to pair selection for reproduction, we must slightly revise Slatkin's recursion:

$$ \mathrm{p}(j)' = \sum_{k_1,k_2} (\mathrm{T}(j \leftarrow k_1,k_2) \times \frac{\mathrm{w}_2(k_1,k_2)}{\overline{\mathrm{w}}_2} \mathrm{p}(k_1,k_2)) \quad (6) $$

The pair frequency (joint density) factors into the individual parental frequencies. Our change to Slatkin's recursion then falls through the proof of Price's Theorem to the conclusion:

$$ \Delta\overline{\mathrm{F}_2} = Cov(\varphi(k_1,k_2), \frac{\mathrm{w}_2(k_1,k_2)}{\overline{\mathrm{w}_2}}) \quad (7) $$

We set the pair sampling rate equal to the measurement function, equal to our joint solution quality index (from Equation 3):

$$ \mathrm{w}_2(k_1,k_2) = \mathrm{F}_2(k_1,k_2) = {}_{N}\mathrm{I}(\mathbf{X};\mathbf{Y};\mathbf{Z}_{k1},\mathbf{Z}_{k2}) \quad (8) $$

Recalling the definition of $\varphi(k_1,k_2)$, we set

$$ \hat{\varphi}(k_1,k_2) = \frac{\mathrm{F}_2(k_1,k_2)}{\overline{\mathrm{F}_2}} \quad (9) $$

yielding

$$ \Delta\overline{\mathrm{F}_2} = Cov(\varphi(k_1,k_2), \hat{\varphi}(k_1,k_2)) \quad (10) $$

which is about the best for which one could hope. The question then becomes "How strong is the covariance?", or equivalently, "How accurate is our estimator of φ?" We are addressing this question of the *heritability*[1] of ${}_{N}\mathrm{I}()$ in ongoing work. The joint mutual information based fitness function of the parents not only covaries with the expected fitness of the offspring; it also incorporates an estimate of the opportunity for improvement due to crossover (the positive variance tail or 'evolvability').

4. CONCLUSIONS

Information theory has numerous important applications to evolutionary learning. It enables explicit treatment of information flows between the environment and the genome, and the gain or loss of information due to evolutionary steps. It is useful in all phases of evolutionary computation, from terminal and non-terminal set selection, through survival and reproductive selection, to objective evaluation of ensemble models as such at end-of-run. Information theoretic indices are easily defined and provide justifiable, heritable, general, computable, commensurate indicators of fitness and diversity, which are undeceived by many transformations. They can measure quality of an ensemble without requiring knowledge of how to compose its constituent simple models into a single complex model; this can be used to guide group selection and non-random mating. Price's Equation may be revised to predict evolutionary dynamics under ensemble selection for reproduction and/or survival.

Much remains to be done in the broad area of applying information theory to evolutionary learning. While we have attempted to apply the basics of entropy, mutual information, etc. to the analysis of evolutionary algorithms, there are many results from the machine learning field [1] and the feature set selection [4] area that might be applied profitably. Our immediate concern is general proof of heritability of ${}_{N}\mathrm{I}()$ across recombination and mutation. This may lead into a re-interpretation of schema theory in terms of information theory. Longer term fundamental work is required to understand generalization in evolutionary learning in terms of information theory; this likely will involve Kullback-Liebler divergence of the joint density of the training inputs and outputs versus that of the testing inputs and outputs, as well as "No Free Lunch" considerations. Information theory may support building blocks in genetic programming and partly explain emergent speciation.

5. REFERENCES

[1] Principe J., Fisher III, and Xu D. Information Theoretic Learning. *Unsupervised Adaptive Filtering*. NY, 2000.

[2] Altenberg, L. The Schema Theorem and Price's Theorem. *Foundations of Genetic Algorithms 3*. Morgan Kaufman, San Francisco, 1995.

[3] Shannon, C. A Mathematical Theory of Communication. *The Bell System Technical Journal, Vol. 27* (1948)

[4] Torkkola, K. On Feature Extraction by Mutual Information Maximization. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[1] Narrow-sense: in evolutionary biology, the fraction of the phenotypic variance that can be used to predict changes in population mean; see http://en.wikipedia.org/wiki/Heritability