Using Informational Confidence Values for Classifier Combination: An Experiment with Combined On-Line/Off-Line Japanese Character Recognition

Stefan Jaeger
Laboratory for Language and Media Processing
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
Jaeger@umiacs.umd.edu

Abstract

Classifier combination has turned out to be a powerful tool for achieving high recognition rates, especially in fields where the development of a powerful single classifier system requires considerable efforts. However, the intensive investigation of multiple classifier systems has not resulted in a convincing theoretical foundation yet. Lacking proper mathematical concepts, many systems still use empirical heuristics and ad hoc combination schemes. My paper presents an information-theoretical framework for combining confidence values generated by different classifiers. The main idea is to normalize each confidence value in such a way that it equals its informational content. Based on Shannon's notion of information, I measure information by means of a performance function that estimates the classification performance for each confidence value on an evaluation set. Having equalized each confidence value with the information actually conveyed, I can use the elementary sum-rule to combine confidence values of different classifiers. Experiments for combined on-line/off-line Japanese character recognition show clear improvements over the best single recognition rate.

1. Introduction

Research on multiple classifier systems has developed into two main branches: One major field of study concerns the generation of classifier ensembles producing high recognition results when their classification results are combined. The other major field deals with the actual integration of those ensembles. While steady progress has been made in the former [1, 3, 4], the latter still suffers from missing theoretical foundations. This lack of theoretical knowledge led researchers experiment with many different combination schemes [11]. For instance, simple voting techniques [2], elementary combination rules (e.g. sum-rule and product-rule [12]) as well as more complex methods (e.g. Dempster/Shafer's theory of evidence [14] and Behavior-Knowledge-Space method [6]) have all been utilized for combination purposes. However, up till now, researchers have not been able to show the general superiority of a particular combination scheme, neither theoretically nor empirically.

In this paper, which falls into the second major branch mentioned above and is an extension of my earlier work described in [7], I present an information-theoretical approach to classifier combination. My idea is to combine the information actually conveyed by each confidence value, and not the raw confidence values as they are provided by the classifiers in a multiple classifier system. In fact, I require that confidence and information are identical. In practice, this means replacing each confidence value by an estimate of its conveyed information. Naturally, I assume that the information conveyed by a confidence value depends somehow on its performance in the practical application domain. Accordingly, I measure the information of a confidence value using a performance function estimating its performance on a given evaluation set. Based on this performance function, I compute the informational content of the confidence value according to the well-known logarithmic notion of Shannon [16]. This implies that the elementary sum-rule is the most natural combination scheme for the newly computed informational confidence values.

My paper is structured as follows: Section 2 describes, in mathematical terms, my main idea of making confidence and information identical. It explains the crucial concept of "performance" and how I estimate the information of a confidence value. Section 3 presents practical recognition rates for handwritten Japanese characters. I will show that using the sum-rule in combination with my proposed method can considerably improve the recognition rate of a combined on-line/off-line Japanese character recognition system. Finally, a summary of my work concludes the paper.

2. Information and Confidence

Generally speaking, confidence values describe the trust a classifier has in its own recognition results. The higher the confidence in a class label, the higher the probability that this class label is indeed the correct label for an unknown test pattern. On the other hand, it is intuitively clear that confidence values convey some sort of information, which helps us determine the correct classification result. However, the exact amount of information conveyed is unknown.

My solution to this problem lies in two simple observations: First, in practical systems, confidence values are merely approximations of the, in the mathematical sense, correct confidence values. Even in extensively trained classifier systems, the confidence values will generally differ from their mathematically correct values. Second, if each classifier of a multiple classifier system connects to a central combination module via a linear transmission line, we can only transmit a single value at a time. This almost inevitably leads to the conclusion that confidence and information should essentially be the same, unless we take sequential transmission or parallel transmission via multiple channels into account. From these observations, I draw the conclusion that we need a learning process that matches confidence with information.

I will now begin formalizing these observations by introducing the notation for confidence and information, based on a general concept of performance. I do not further specify what "performance" actually means. In fact, we will see that the precise mathematical definition of performance follows almost immediately from the observations and assumptions made above:

Let K_C be the set of confidence values of a classifier C:

$$K_C = \{K_1, \ldots, K_i, \ldots, K_N\} \tag{1}$$

Furthermore, let $p(K_i)$ denote the performance of the i-th confidence value K_i. Then, according to the general observations made above, I set confidence and information equal by the following linear fixed point equation:

$$K_i \;=\; E * I\left(\overline{p}(K_i)\right) + C \tag{2}$$

In (2), the variable E is a multiplying factor just influencing the scale. However, we will later see that E represents also an expectation value in the statistical sense. The term C is merely a constant specifying an offset. It will play no further role in this paper. The expression $I\left(\overline{p}(K_i)\right)$ in (2) denotes the information conveyed by the complement $\overline{p}(K_i)$ of the performance $p(K_i)$. Note that the performance complement $\overline{p}(K_i)$, which we can imagine as the error rate of K_i for the time being, has the desired feature of providing more information for better performances of K_i. For my derivations in this paper, I use a special form of (2), which I obtain by setting C to zero and expressing $\overline{p}(K_i)$ as $1 - p(K_i)$. Also, I measure information according to Shannon, with the natural basis e as information bit. Under these assumptions, (2) can be restated as follows:

$$\begin{aligned} K_i &= E * I\left(1 - p(K_i)\right) \\ &= -E * \ln\left(1 - p(K_i)\right) \end{aligned} \tag{3}$$

This is a fixed point equation that assigns low confidence to values with poor performance. On the other hand, it assigns infinite confidence to values with perfect performance.

We can now also more precisely specify the performance function $p(K_i)$ by resolving (3) for $p(K_i)$. A straightforward transformation produces the following performance specification:

$$\begin{aligned} & K_i = E * I\left(1 - p(K_i)\right) \\ \Longleftrightarrow\quad & \frac{K_i}{E} = -\ln\left(1 - p(K_i)\right) \\ \Longleftrightarrow\quad & e^{-\frac{K_i}{E}} = 1 - p(K_i) \\ \Longleftrightarrow\quad & p(K_i) = 1 - e^{-\frac{K_i}{E}} \end{aligned} \tag{4}$$

This result shows that (3) is only met for a performance function $p(K_i)$ that equals the distribution of a random variable with exponential density and parameter $\lambda = \frac{1}{E}$.

Let me repeat some basic statistics in order to clarify this crucial result. The general definition of an exponential density function $e_\lambda(x)$ with parameter λ is:

$$e_\lambda(x) = \begin{cases} \lambda * e^{-\lambda x} & : \; x \geq 0 \\ 0 & : \; x < 0 \end{cases} \qquad \lambda > 0 \tag{5}$$

Its corresponding distribution $E_\lambda(k)$, which describes the probability that the exponentially distributed random variable assumes values lower than or equal to k, is therefore

$$\begin{aligned} E_\lambda(k) &= \int_{-\infty}^{k} e_\lambda(x)\, dx \\ &= \int_{0}^{k} \lambda * e^{-\lambda x}\, dx \\ &= \left[-e^{-\lambda x}\right]_0^k \\ &= 1 - e^{-\lambda k} \end{aligned} \tag{6}$$

The only difference between (4) and (6) lies in the parameter λ. Setting λ to $\frac{1}{E}$ makes both the performance function and the exponential distribution identical. This relationship between distribution and performance now also sheds light on the parameter E. Since the expectation value of an exponentially distributed random variable with parameter λ is $\frac{1}{\lambda}$, parameter E denotes the expected confidence.

As already mentioned in the beginning of this section, confidence values in practical systems will almost always violate the fixed point equation in (3). In other words, the equilibrium between information and confidence will usually be distorted in practice. Nevertheless, I will show that a simple training process is able to restore this equilibrium and adjust the confidence values so that they satisfy (3). Technically, there are two different ways of restoring the equilibrium in (3): We can either adjust the expectation E or the confidence K_i. In this paper, I concentrate on the latter, i.e. adjusting confidence, and assume E to be an invariable constant for each classifier.

Let me first derive the expectation E that I use in my experiments, before I then go into detail regarding the adjustment of confidence. Disregarding Parameter E for the time being, the following definite integral describes the average amount of information provided by all confidence values:

$$\begin{aligned}\int_0^1 -\ln(1-p(K))\ dp(K) &= \int_0^1 -\ln(K)\ dK \\ &= [K - \ln(K) * K]_0^1 \\ &= 1 \qquad (7)\end{aligned}$$

Accordingly, the average amount of information provided is exactly one bit (Euler-bit). Applying this result to (3) again confirms that Parameter E, which is merely a factor multiplying the integral in (7), equals the average expected information in the state of equilibrium. Motivated by these observations, I compute E as the expected information $E(R)$ conveyed by classifier C with recognition rate R for a recognition process that contains one bit overall information. Accordingly, the fixed point equation in (3) now formulates as:

$$K_i = -E(R) * \ln(1 - p(K_i)) \qquad (8)$$

In fact, the expectation value E has now become a function, which I derive via the information provided by C. Following the definition of information for confidence values, the information $I(C)$ conveyed by classifier C computes as

$$I(C) = I(1-R) = -\ln(1-R), \qquad (9)$$

where R again denotes the overall recognition rate of C. I can now compute the information $E(R)$, which we can expect from C for a one bit process, using the information $I(C)$ provided by C:

$$E(R) = \sqrt[I(C)]{R} = R^{\frac{1}{I(C)}} \qquad (10)$$

This concludes the derivation of E or rather $E(R)$ in (3) and (8), respectively.

Let us now turn to the second option of restoring the equilibrium in (3) and (8), namely adjusting confidence values. Adjusting a confidence value to its proper value requires the knowledge of its actual performance. Based on its true performance, we can compute its correct value satisfying the fixed point equation in (8). Therefore, I estimate the true performance on a training set, and replace the old confidence values by the new values computed according to (8). Motivated by the relationship between performance function and distribution function in (4), I compute the following estimate $\hat{p}(K_i)$ of $p(K_i)$:

$$\hat{p}(K_i) = E\left(\frac{R_i}{R}\right) \qquad (11)$$

R_i denotes the partial recognition rate on all patterns with confidence values $K \leq K_i$, measured on an evaluation set. E is the expectation function defined in (10). In general, R_i will be a monotonously increasing function over the set of confidence values K_C with $R_i \leq R$, so that the estimated performance $\hat{p}(K_i)$ will converge on 1 for increasing confidences. The estimate $\hat{p}(K_i)$ is therefore an estimate of the percentage of information that classifier C has realized when classifying a test pattern with confidence K_i.

I also experimented with the following performance estimate:

$$\hat{p}(K_i) = E\left(\frac{I(1-R_i)}{I(1-R)}\right) = E\left(\frac{\ln(1-R_i)}{\ln(1-R)}\right) \qquad (12)$$

This performance estimate is perhaps even closer to my idea of estimating the performance of informational confidence values by approximating an exponential distribution. The estimate in (12) takes the relative informational content of partial recognition rates into account. In my practical experiments in Section 3, both performance estimates will provide similar performance.

Inserting the performance estimate $\hat{p}(K_i)$ directly into (8) provides the new confidence values $\hat{K}_i$, which are the trained estimates for the optimal confidence values K_i:

$$\hat{K}_i = -E(R) * \ln(1 - \hat{p}(K_i)) \qquad (13)$$

These new confidence values will replace the old ones when combining classifier C with other classifiers.

3. Experimental Evaluation

I evaluated my proposed method for a multiple classifier system comprising two recognizers for handwritten Japanese characters. Both recognizers are on-line classifiers expecting a point sequence over time. However, one recognizer operates on an off-line pictorial representation, which it generates by connecting neighboring on-line points [19, 10]. We can therefore consider this classifier to be an off-line classifier. Multiple classifier systems have a long tradition in handwriting recognition [21, 5, 20]. In particular, the combination of on-line and off-line recognition copes with the problem of stroke-order and stroke-number variations inherent in on-line data [8, 9]. This is especially important for Japanese or Chinese character recognition since the number of strokes per character, and thus the number of variations, is generally higher than in the Latin alphabet [8, 13].

In my experiments, both recognizers were trained with more than 1 million handwritten Japanese characters [15, 10]. The test set contained $54,775$ handwritten characters. Experiments in [17, 18] show that the performance estimate $\hat{p}(K_i)$ does not depend on the underlying data for these large data sets, only on the classifier. I can therefore safely compute the performance estimates $\hat{p}(K_i)$ on the training set, given its large size.

Table 1 shows my recognition results for n-best candidate lists ranging from $n = 1$ to $n = 10$. The second and third column contain the single recognition rates for the off-line and on-line recognizer respectively. We see that the off-line recognition rates are much higher than the corresponding on-line rates. Clearly, stroke-order and stroke number variations are partly responsible for this performance difference.

Column 4 and Column 5 show the number of test patterns for which the correct class label occurs either twice (AND) or at least once (OR) in the n-best lists of both classifiers.

The next five columns show the combined recognition rates for the sum-rule and differently computed confidence values. The (+)-column contains the results for unmodified confidence values, i.e. without any adjustment. Column (11) shows the combined rates for informational confidence computed according to the performance estimate in (11) and the fixed point equation in (13). In Column (12), I used instead the estimate in (12). Both estimates show similar behavior. However, I achieved a slightly better performance by using the partial recognition rates directly, i.e. applying the performance estimate in (11).

Column (11)' and (12)' list the recognition rates for the performance estimates in (11) and (12), respectively, but without the function E applied to the fractions $\frac{R_i}{R}$ and $\frac{\ln(1-R_i)}{\ln(1-R)}$. We see that the recognition rates are a bit lower than in the corresponding columns (11) and (12). This result confirms practically the usefulness of the normalization in (10).

Table 1 also shows that the combined on-line and off-line recognition rates outperform the off-line classifier, which is the best individual classifier, by more than 4.5%.

The plain sum-rule, without any informational confidence, already accounts for more than 3% of this improvement. Informational confidence increases the performance still further by more than one percent, reaching the best overall recognition rate of 93.6%. Considering the state-of-the-art in handwritten Japanese character recognition and the quality of the classifiers and real-world test data, this is an exceptional result. Also, my method exploits the candidate alternatives in the n-best lists to a fairly high extent, as indicated by the small difference between the "OR"-column and the actual recognition rates. Note that the n-best recognition rates for the informational confidence values are actually slightly lower than the rates provided by the plain sum-rule for $n \geq 3$. This could be directly related to the fact that I consider only $n = 1$ when computing the performance estimates for informational confidence values. I will investigate this effect further in future experiments. Nevertheless, the recognition rate for $n = 1$ is one of the best ever measured for this test set [13].

Figure 1 shows the partial recognition rates for both off-line and on-line confidence values on the left-hand and right-hand side, respectively. In both cases, the partial recognition rates describe a monotonously increasing function over confidence. The off-line function is steeper and reaches a higher level though. Figure 2 shows the new informational confidence values computed according to (11) and (13). Due to the better performance of the off-line classifier, the off-line confidence values reach a higher level than the on-line confidence values.

4. Summary

Classifiers of a classifier ensemble generally provide confidence values with quite different mathematical properties, especially when they are based on different underlying architectures. These differences pose a considerable problem for classifier combination, making it even impossible sometimes. My aim is to overcome this problem by adjusting confidence values so that they become directly comparable. In order to do so, I presented an information-theoretical method that tries to establish an equilibrium between information and confidence. In the state of equilibrium, which I defined by a fixed point equation, each confidence value matches exactly its informational content. My idea of information is based on Shannon's definition of information in combination with a so-called performance function. I showed that the performance function is the distribution of an exponentially distributed random variable. This causal relationship motivated the use of partial recognition rates to compute performance estimates for each confidence value. Using these performance estimates, I computed adjusted confidence values that replace the old values. The newly computed informational confidence values allow me to apply the elementary sum-rule when combining confidence values from different classifiers.

Note that for sufficiently large training sets the performance estimate is a monotonously increasing function. My proposed method will therefore not affect the single recog-

n-	Single		Truth		Combined				
best	off-line	on-line	AND	OR	(+)	(12)’	(11)’	(12)	(11)
1	89.07	80.57	74.76	94.88	92.43	93.25	93.43	93.57	**93.60**
2	93.69	85.17	81.93	96.93	95.60	95.87	95.88	95.87	95.86
3	95.09	86.90	84.43	97.56	96.53	96.59	96.55	96.55	96.52
4	95.82	87.76	85.68	97.90	96.99	96.95	96.89	96.86	96.83
5	96.24	88.27	86.40	98.11	97.27	97.16	97.10	97.06	97.04
6	96.51	88.70	86.96	98.25	97.50	97.30	97.24	97.17	97.14
7	96.74	88.99	87.37	98.36	97.64	97.42	97.35	97.28	97.25
8	96.93	89.25	87.71	98.47	97.80	97.51	97.44	97.37	97.34
9	97.04	89.48	87.99	98.52	97.89	97.57	97.48	97.42	97.39
10	97.16	89.68	88.26	98.57	97.97	97.63	97.54	97.48	97.44

Table 1. Single and combined recognition results.

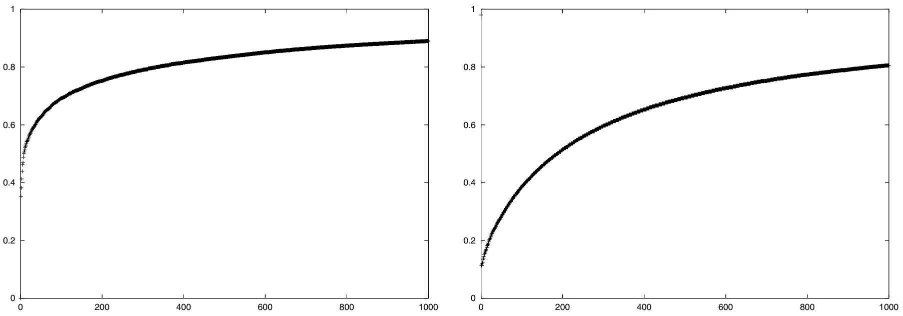

Figure 1. Partial off-line (left) and on-line (right) recognition rates.

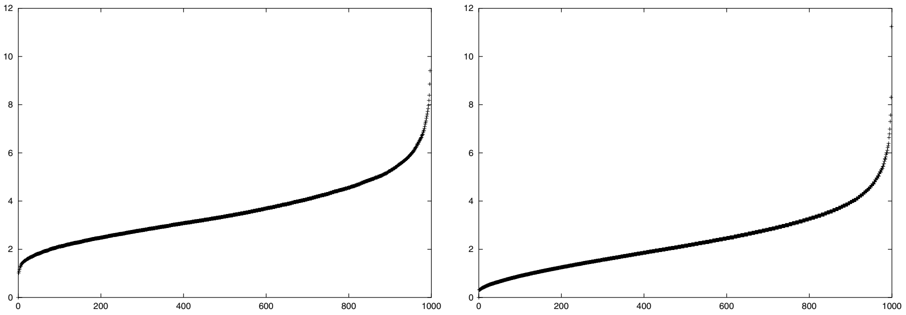

Figure 2. Informational off-line (left) and on-line (right) confidence values.

nition rates in a classifier ensemble, only the combined recognition rates of multiple classifiers.

Practical experiments for combined on-line/off-line Japanese handwriting recognition showed the good performance of my method.

In future experiments, I hope to be able to show that informational confidence values can be useful for application domains other than handwriting recognition, and for multiple classifier systems with more than two classifiers.

Acknowledgment:
I would like to thank Masaki Nakagawa, Ondrej Velek, and Akihito Kitadai from Tokyo University of Agriculture & Technology for their support in generating classifier logfiles for my practical experiments. I also gratefully acknowledge the support by the US Department of Defense (Contract MDA90402C0406) and the National Science Foundation (Grant EIA0130422).

References

[1] L. Breiman. Bagging Predictors. *Machine Learning*, 2:123–140, 1996.

[2] M. V. Erp, L. G. Vuurpijl, and L. Schomaker. An Overview and Comparison of Voting Methods for Pattern Recognition. In *Proc. of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)*, pages 195–200, Niagara-on-the-Lake, Canada, 2002.

[3] Y. Freund and R. Schapire. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[4] S. Guenter and H. Bunke. New Boosting Algorithms for Classification Problems with Large Number of Classes Applied to a Handwritten Word Recognition Task. In *4th International Workshop on Multiple Classifier Systems (MCS)*, pages 326–335, Guildford, UK, 2003. Lecture Notes in Computer Science, Springer-Verlag.

[5] T. Ho, J. Hull, and S. Srihari. Decision Combination in Multiple Classifier Systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 16(1):66–75, 1994.

[6] Y. Huang and C. Suen. A Method of Combining Multiple Experts for Recognition of Unconstrained Handwritten Numerals. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 17(1):90–94, 1995.

[7] S. Jaeger. Informational Classifier Fusion. In *Proc. of the 17th International Conference on Pattern Recognition (to be published)*, Cambridge, UK, 2004.

[8] S. Jaeger, C.-L. Liu, and M. Nakagawa. The State of the Art in Japanese Online Handwriting Recognition Compared to Techniques in Western Handwriting Recognition. *International Journal on Document Analysis and Recognition*, 6(2):75–88, 2003.

[9] S. Jaeger, S. Manke, J. Reichert, and A. Waibel. Online Handwriting Recognition: The Npen++ Recognizer. *International Journal on Document Analysis and Recognition*, 3(3):169–180, 2001.

[10] S. Jaeger and M. Nakagawa. Two On-Line Japanese Character Databases in Unipen Format. In *6th International Conference on Document Analysis and Recognition (ICDAR)*, pages 566–570, Seattle, 2001.

[11] A. Jain, P. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[13] C.-L. Liu, S. Jaeger, and M. Nakagawa. Online Recognition of Chinese Characters: The State-of-the-Art. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(2):198–213, 2004.

[14] E. Mandler and J. Schuermann. Combining the Classification Results of Independent Classifiers Based on the Dempster/Shafer Theory of Evidence. In L. K. E.S. Gelsema, editor, *Pattern Recognition and Artificial Intelligence*, pages 381–393, 1988.

[15] M. Nakagawa, K. Akiyama, L. Tu, A. Homma, and T. Higashiyama. Robust and Highly Customizable Recognition of On-Line Handwritten Japanese Characters. In *Proc. of the 13th International Conference on Pattern Recognition*, volume III, pages 269–273, Vienna, Austria, 1996.

[16] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Tech. J.*, 27(623-656):379–423, 1948.

[17] O. Velek, S. Jaeger, and M. Nakagawa. A New Warping Technique for Normalizing Likelihood of Multiple Classifiers and its Effectiveness in Combined On-Line/Off-Line Japanese Character Recognition. In *8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 177–182, Niagara-on-the-Lake, Canada, 2002.

[18] O. Velek, S. Jaeger, and M. Nakagawa. Accumulated-Recognition-Rate Normalization for Combining Multiple On/Off-line Japanese Character Classifiers Tested on a Large Database. In *4th International Workshop on Multiple Classifier Systems (MCS)*, pages 196–205, Guildford, UK, 2003. Lecture Notes in Computer Science, Springer-Verlag.

[19] O. Velek, C.-L. Liu, S. Jaeger, and M. Nakagawa. An Improved Approach to Generating Realistic Kanji Character Images from On-Line Characters and its Benefit to Off-Line Recognition Performance. In *16th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 588–591, Quebec, 2002.

[20] W. Wang, A. Brakensiek, and G. Rigoll. Combination of Multiple Classifiers for Handwritten Word Recognition. In *Proc. of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)*, pages 117–122, Niagara-on-the-Lake, Canada, 2002.

[21] L. Xu, A. Krzyzak, and C. Suen. Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE Trans. on Systems, Man, and Cybernetics*, 22(3):418–435, 1992.