An Information Theoretic Perspective on Multiple Classifier Systems

Gavin Brown

School of Computer Science, University of Manchester,
Kilburn Building, Oxford Road, Manchester, M13 9PL
gbrown@cs.man.ac.uk
http://www.cs.man.ac.uk/~gbrown/

Abstract. This paper examines the benefits that *information theory* can bring to the study of multiple classifier systems. We discuss relationships between the mutual information and the classification error of a predictor. We proceed to discuss how this concerns ensemble systems, by showing a natural *expansion* of the ensemble mutual information into "accuracy" and "diversity" components. This natural *derivation* of a diversity term is an alternative to previous attempts to *artificially define* a term. The main finding is that diversity in fact exists at multiple orders of correlation, and pairwise diversity can capture only the low order components.

1 Introduction

Information Theory sparked a revolution in the practice of electronic communications [1] and has since been successfully applied in countless fields, from anthropology to biology to cosmology. In the last decade or so, it has found significant uptake in Machine Learning. Suppose there is a message Y, encoded and sent to us by a friend through a communications channel, that we receive as a signal X. We would like to decode the received signal X, and recover the correct message Y; that is, we will perform a decoding operation, $\hat{Y} = g(X)$. In Machine Learning terms, we imagine that the friend transmitting the message has access to a particular object, for which Y is the correct class label. They 'encode' the object as a feature vector X. Our task is to decode that feature vector and recover the correct class label, using our predictor function $g(\cdot)$. Using this analogy, information theory provides us with a language and a set of mathematical tools to analyze the situation. One of the most interesting observations it can provide is a bound on the error of our predictor, dependent on the chosen features X. This bound, known as *Fano's inequality*, applies for *any* predictor: be it a simple decision stump, or a nonlinear support vector machine.

We can also use information theory to understand multiple classifier systems. To make the link, consider the received signal X not to be a set of features, but as a set of classifier outputs, which we will use to form an ensemble. In this case, the predictor $g(\cdot)$ corresponds to the ensemble combiner function. In this work

J.A. Benediktsson, J. Kittler, and F. Roli (Eds.): MCS 2009, LNCS 5519, pp. 344–353, 2009.

we investigate the link in detail, in particular addressing the notion of ensemble diversity.

This paper is structured as follows. Section 2 provides a tutorial introduction to the basics of information theory, including the lesser known concept of *multivariate* mutual information. Section 3 describes how an understanding for the concept of diversity can *naturally* emerge as an expansion of the ensemble mutual information. Section 4 uses this result to characterize and explain the behaviors of Adaboost versus Bagging, sections 5 and 6 present related work and conclude with a look ahead to what advantages this approach might bring to MCS.

2 Background

In this section we review the required elements of information theory, and their relation to Machine Learning. Due to space limitations this is necessarily brief; for an extended treatment the reader might consult reference [2] or [3].

2.1 Information Theory Basics

The fundamental unit of information theory is the *entropy* of a random variable [1]. The entropy, denoted $H(X)$, quantifies the uncertainty present in the distribution of X. It is defined[1] as,

$$H(X) = -\sum_{i=1}^{|X|} p(x_i) \log p(x_i). \tag{1}$$

The base of the logarithm is arbitrary, but decides the "units" of the entropy. When using base 2, the units are 'bits', when using base e, the units are 'nats'. To compute this, we need an estimate of the distribution $p(X)$. This is estimated by frequency counts from data, that is $p(x_i) = \frac{\#x_i}{N}$, the fraction of observations taking on value x_i from the total number of observations N.

If the distribution is highly biased toward one particular event $x \in X$, i.e. little uncertainty over the outcome, then the entropy is low. If all events are equally likely, i.e. maximum uncertainty over the outcome, then $H(X)$ is maximal[2]. Following the rules of standard probability theory, entropy can also be *conditioned* on other events. The *conditional entropy* of X given Y is denoted,

$$H(X|Y) = -\sum_{j=1}^{|Y|} p(y_j) \sum_{i=1}^{|X|} p(x_i|y_j) \log p(x_i|y_j). \tag{2}$$

This can be thought of as the amount of uncertainty remaining in X after we learn the outcome of Y.

[1] In this work we restrict ourselves to discrete RVs, and note $z \log(z) \rightarrow 0$ with $z \rightarrow 0$.

[2] In general, $0 \leq H(X) \leq \log(|X|)$.

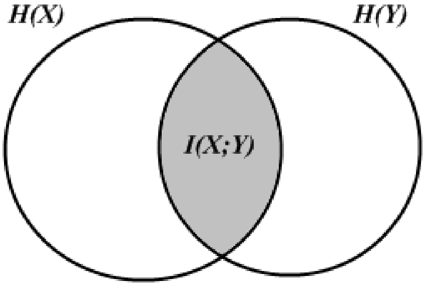

Fig. 1. Illustration of various information theoretic quantities

We can now define the *Mutual Information* between X and Y, i.e. the amount of information *shared* by X and Y, as follows.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_X \sum_Y p(xy) \log \frac{p(xy)}{p(x)p(y)}. \end{aligned} \tag{3}$$

It should be noted that Mutual Information is symmetric, i.e. $I(X;Y)=I(Y;X)$. The relation between all these quantities can be seen in figure 1. The Mutual Information can also be conditioned on other events—the *conditional mutual information* is,

$$\begin{aligned} I(X_1;X_2|Y) &= H(X_1|Y) - H(X_1|X_2Y) \\ &= \sum_Y p(y) \sum_{X_1} \sum_{X_2} p(x_1x_2|y) \log \frac{p(x_1x_2|y)}{p(x_1|y)p(x_2|y)}. \end{aligned} \tag{4}$$

This can be thought of as the information still shared between X_1 and X_2 after the value of Y is revealed. The conditional mutual information will emerge as a particularly important property in understanding the message of this paper.

2.2 Relationship to Machine Learning

Suppose there is a message Y, that was sent through a communications channel, and we received the value X. We would like to decode the received value X, and recover the correct Y. That is, we will perform a decoding operation, $\hat{Y} = g(X)$. In ML terms: Y is the original (unknown) class label distribution, X is the particular set of features chosen to represent the problem, and g is our predictor. The set of features chosen may or may not be sufficient to perfectly recover Y; that is, there may be an error in prediction. Information theory can provide a bound on $p(\hat{Y} \neq Y)$, for *any* predictor g.

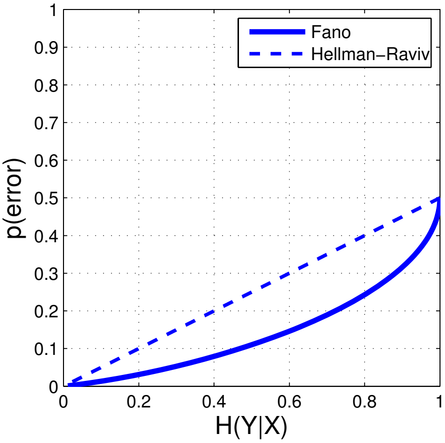

Fig. 2. Fano's inequality [4] provides a lower bound on the Bayes rate, while Hellman-Raviv [5] provides the upper bound. Picking features to reduce conditional entropy (equivalent to maximising mutual information) causes this bound to be minimized.

The error of predicting target variable Y from input X is tightly bounded by two inequalities [4,5]. The bounds state,

$$\frac{H(Y) - I(X;Y) - 1}{\log(|Y|)} \leq p(g(X) \neq Y) \leq \frac{1}{2}H(Y|X). \tag{5}$$

In order to maximise the chances of our predictor guessing the correct class label, we should have maximum $I(X;Y)$. Given the definition (3), this is equivalent to *minimizing* $H(Y|X)$, illustrated in figure 2. As the mutual information $I(X;Y)$ grows, the bound is minimized—whether or not the bound can be reached depends on the ability of our classifier, i.e. the function $g(X)$.

For example, if the conditional entropy is measured to be $H(Y|X) = 0.4$, then the minimum error rate by *any* classifier lies in the range $[0.079, 0.2]$. In other words, no classifier can possibly achieve better than error 0.079 with features X, and there exists a classifier that can achieve at least error 0.2. It should be noted that, in real ML problems, since we only ever have access to a *sample* of X (not the full distribution) this is in practice an estimated bound on the *training* error. We will investigate relations to the the generalization error in section 4.

We have now covered the basic properties of information theory. To complete the background necessary for this paper, we now briefly review the lesser-known topic of *multi-variate* mutual information.

2.3 Multi-variate Mutual Information

While Shannon's mutual information $I(X;Y)$ measures dependence between a *pair* of variables, the multivariate form, known as *Interaction Information* [6], can account for dependencies among *multiple* variables. For a set of size 2, the Interaction Information reduces to Shannon's definition. For *three* random variables, the Interaction Information is

$$I(\{X_1, X_2, X_3\}) = I(X_1; X_2|X_3) - I(X_1; X_2), \tag{6}$$

that is, a difference of the conditional mutual information and the simple mutual information. The case for n variables is defined recursively. A full treatment of this advanced topic is not possible given the limited space; for more information the reader is referred to reference [7]. The interaction information turns out to be useful in understanding the nature of ensemble diversity, which we will explore in the following section.

3 Mutual Information and Ensemble Classifiers

One of the long-standing problems in the MCS literature is to understand the nature of ensemble *diversity*. We know that ensemble members should exhibit some level of accuracy. We also know that ensemble members should not be identical, exhibiting some level of diversity. However, quantifying these statements has proved challenging [8]. In this section we take an information theoretic perspective.

3.1 Why Is Diversity So Elusive?

To answer this question [9] we return to one of the most well-known results in the MCS literature concerning the diversity issue. Tumer & Ghosh [10] related the ensemble classification error to the correlations between the individual predictor outputs. They showed that the error of a *linearly* combined ensemble could be decomposed neatly into accuracy and diversity components. This exemplary early work sparked much effort to find the corresponding accuracy-diversity terms for a majority voting ensemble. A fundamental message of this work is that *we should not expect the majority vote ensemble error to similarly decompose into additive accuracy-diversity terms.*

The neat situation in [10] is due to the linearity of the combination operator, and bias-variance properties of the squared loss function. When we have a *nonlinear* combination operator, and a *zero-one* loss function, the situation is more complicated. It is well appreciated that there exists no unique definition of bias and variance for zero-one loss. In the same fashion, there is no unique definition of *covariance* (diversity) with this loss function; instead, the literature has spawned a myriad of diversity definitions [8] with desirable and undesirable properties.

It is often the case in Machine Learning to use a *surrogate* loss function, and minimise that instead of the actual one of interest. Adaboost is the prime example of this—the distribution updates in the algorithm *do not directly minimise* classification error, but instead minimize a surrogate, an exponential loss which *bounds* the classification loss. In this way, when the exponential loss is small, we can be guaranteed the classification loss will also be at least as small. In the following section we take a similar approach, remembering that the classification error rate can be bounded by the *mutual information*, using Fano's inequality.

3.2 A 'Natural' Definition of Diversity

In this section we show a diversity term emerges naturally when we measure the ensemble mutual information. This draws on a recent result in the feature selection literature [7], described and adapted for the MCS community in the Appendix. For a set of classifiers $S = \{X_1, ..., X_M\}$, remembering that our objective is to maximise $I(X_{1:M}; Y)$, we have the expansion,

$$I(X_{1:M}; Y) = \sum_{i=1}^{M} I(X_i; Y) - \sum_{\substack{\boldsymbol{X} \subseteq S \\ |\boldsymbol{X}|=2..M}} I(\{\boldsymbol{X}\}) + \sum_{\substack{\boldsymbol{X} \subseteq S \\ |\boldsymbol{X}|=2..M}} I(\{\boldsymbol{X}\}|Y). \quad (7)$$

The expansion consists of three terms. The first, $\sum_{i=1}^{M} I(X_i; Y)$ is the sum of each individual classifier's mutual information with the target. Since the mutual information is actually only a *bound* on the accuracy, not the *actual* accuracy, it is misleading to say this is an 'accuracy' term. Instead, we refer to the first term as the *relevancy* of a classifier output to the target. The final combination function g will determine if this provides good accuracy in combination with the other classifiers.

The second contains terms of the form $I(\{\boldsymbol{X}\})$ and is independent of the class label Y, and so is the closest analogy to the (now almost mythical) concept of 'diversity'. It measures the interaction information among *all possible subsets of classifiers*, drawn from the ensemble. We refer to this as the ensemble *redundancy.* Notice this term is *subtractive* from the overall mutual information. A large value of $I(\{\boldsymbol{X}\})$ indicates strong correlations between the classifiers, and reduces the value of $I(X_{1:M}; Y)$, and hence the overall achievable accuracy.

The third contains terms of the form $I(\{\boldsymbol{X}\}|Y)$ and is a function of the class label Y. This therefore does not correspond to the folklore definition of 'diversity', that it should be a function solely of the classifier outputs. We call this the *conditional redundancy.* Notice that this term is *additive* to the ensemble mutual information. While it is commonly accepted that we should have low correlations between ensemble members, this term indicates that we in fact need *strong class-conditional correlations.* The balance between these conditional and unconditional terms is similar to aiming for a small within-class variance (maximizing the dependency $I(\{\boldsymbol{X}\}|Y)$) and a large between-class variance (minimizing the dependency $I(\{\boldsymbol{X}\})$).

3.3 Low-Order and High-Order Diversity

We have found that through an expansion of the ensemble mutual information, terms which we might call 'diversity' appear naturally. The *redundancy* is a traditional diversity term, and the *conditional redundancy* is the same form but conditioned on the class label. The sum of these two values is what we refer to as the "*diversity*" of the classifier set. It should be noted that the summations in eq(7) are *over all possible subsets of classifiers drawn from the ensemble.* We can expand this sum over subsets, to give us a breakdown of diversity,

$$\begin{aligned} I(X_{1:M};Y) = \sum_{i=1}^{M} I(X_i;Y) &- \sum_{|\boldsymbol{X}|=2} I(\{\boldsymbol{X}\}) + \sum_{|\boldsymbol{X}|=2} I(\{\boldsymbol{X}\}|Y) \\ &- \sum_{|\boldsymbol{X}|=3} I(\{\boldsymbol{X}\}) + \sum_{|\boldsymbol{X}|=3} I(\{\boldsymbol{X}\}|Y) \\ &- \ldots \qquad\qquad + \ldots \\ &- \sum_{|\boldsymbol{X}|=M} I(\{\boldsymbol{X}\}) + \sum_{|\boldsymbol{X}|=M} I(\{\boldsymbol{X}\}|Y). \end{aligned}$$

This breakdown has the form,

$$\begin{aligned} I(X_{1:M};Y) = \textit{Individual Mutual Info } &+ \textit{2-way diversity (pairwise)} \\ &+ \textit{3-way diversity} \\ &+ \textit{...-way diversity} \\ &+ \textit{M-way diversity} \end{aligned}$$

where the diversity measure is the multivariate mutual information. This expansion reflects the true complexity of the accuracy-diversity issue. Diversity is *not* simply a pairwise measure between classifiers, such as the Q-statistics or the Double-Fault measures. Diversity in fact exists on numerous *levels* of interaction between the classifiers.

4 Monitoring Low-Order Diversity Components

In the previous section we showed that diversity exists at multiple levels of correlation within an ensemble. If the classifiers were statistically independent, then all diversity terms would be zero, and we would have simply $I(X_{1:M};Y) = \sum_{i=1}^{M} I(X_i;Y)$. If the classifiers only exhibited pairwise interactions, the breakdown be as above but omitting the 3-way and above diversity terms. This assumption of pairwise interactions gives us,

$$I(X_{1:M};Y) \approx \sum_{i=1}^{M} I(X_i;Y) - \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_j,X_k) + \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_j,X_k|Y) \quad (8)$$

The ensemble information is thus approximated by a sum of the relevancy, the pairwise redundancy, and the pairwise conditional redundancy. In figures 3, 4, and 5 we monitor these three components to characterize the behavior of Adaboost and Bagging. All information measurements are made on *training* data, and used to explain the performance on test data. Examining the pairwise components we find Adaboost succeeds by decreasing redundancy, but has no effect on the conditional term. Bagging has no effect on either, reflected in the poor test error. Further comment is provided in the figure captions.

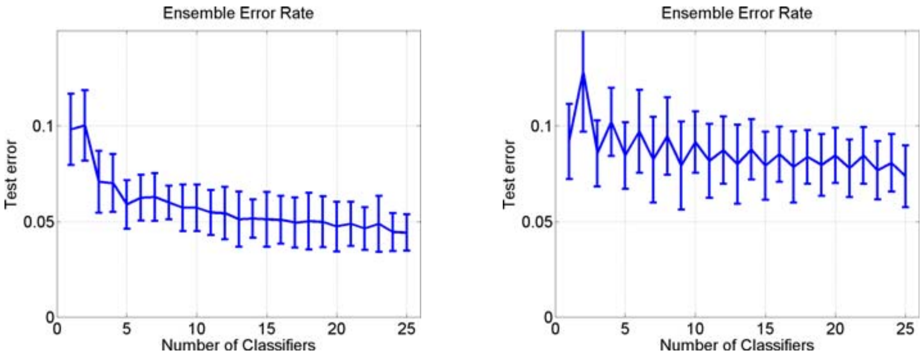

Fig. 3. Adaboost (left) and Bagging (right) errors using decision stumps on the Breast Cancer data. Graphs show standard deviation over ten trials of 2-fold cross validation.

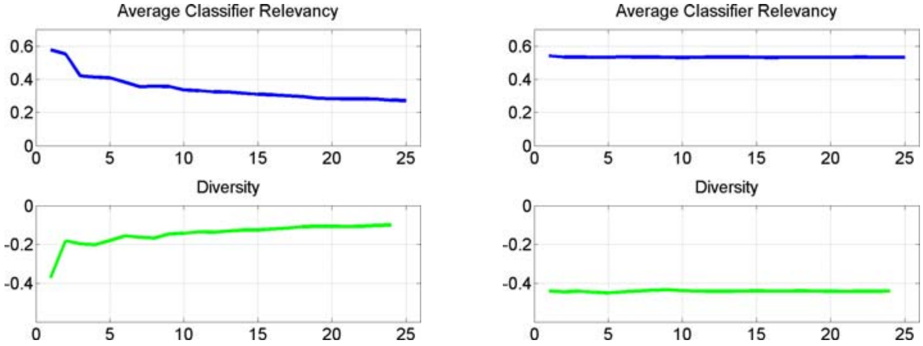

Fig. 4. The Relevancy-Diversity tradeoff. On the left we see the average relevancy of Adaboost classifiers decreases over time, but the diversity component compensates this by also rising. On the right, Bagging maintains almost constant classifier relevancy and very low diversity, explaining the poor test error in figure 3.

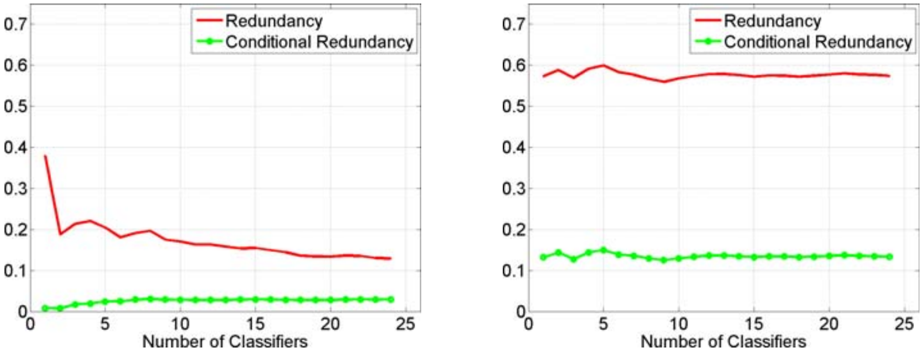

Fig. 5. Second order components of the ensemble mutual information. Adaboost (left) decreases the redundancy of its classifiers, though maintains constant conditional redundancy. Bagging (right) allows the redundancy to rise very slightly at small ensemble size, but has no significant effect on either component.

5 Related Work

Meynet and Thiran [11] suggest a heuristic cost function, designed to balance ensemble accuracy with diversity. The cost function consists of two information theoretic terms. The first is simply the average mutual information between each ensemble member and the class label, which they call the Information Theoretic Accuracy, $ITA = \frac{1}{M}\sum_{i=1}^{M} I(X_i;Y)$. The second is the reciprocal of the average pairwise mutual information between ensemble members, which they call the Information Theoretic Diversity,

$$ITD = \Big(\frac{1}{\binom{M}{2}}\sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_j;X_k)\Big)^{-1}. \tag{9}$$

Thus, the task is to simultaneously maximise ITA and ITD, though it is clear that a tradeoff will occur between the two. The authors represent the tradeoff by a second-order polynomial: the Information Theoretic Score is defined,

$$ITS = (1+ITA)^3.(1+ITD) \tag{10}$$

Comparing the form of ITA and ITD to the results in section 3.2, it is clear that ITS includes two of the necessary components to take account of pairwise interactions between ensemble members. The final term necessary is the class-conditional $I(X_i;X_j|Y)$, and the higher-order terms are assumed zero. The main difference between this heuristic and the current work is that ITS was *hand-designed*, whereas we have shown a natural *derivation* of a diversity term.

6 Conclusion

This paper examined the issue of ensemble diversity from an information theoretic perspective. A major advantage of information theoretic criteria is they capture higher order statistics of the data. In contrast, the squared error criterion can capture only second-order statistics. The main finding was an expansion of the ensemble mutual information which naturally involves “accuracy” and “diversity” component, although diversity is shown to exist at several levels, having low and high order elements.

The advantage of this approach is that $g(\cdot)$ can be *any* function, that is, any ensemble combiner function. In this paper we showed preliminary results with the majority vote combiner, as this has traditionally been of most interest regarding the ‘diversity’ question. Extensions to this work might assess how effective different combiner functions are at ‘decoding’ the information contained in the ensemble.

References

1. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. 27(3), 379–423 (1948)
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (1991)
3. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2003)
4. Fano, R.: Transmission of Information: Statistical Theory of Communications. Wiley, New York (1961)
5. Hellman, M., Raviv, J.: Probability of error, equivocation, and the Chernoff bound. IEEE Transactions on Information Theory 16(4), 368–372 (1970)
6. McGill, W.: Multivariate information transmission. IEEE Trans. Inf. Theory 4(4), 93–111 (1954)
7. Brown, G.: A New Perspective on Information Theoretic Feature Selection. In: Proceedings of Intl. Conf. on Artificial Intelligence and Statistics (2009)
8. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, Hoboken (2006)
9. Kuncheva, L.: That Elusive Diversity in Classifier Ensembles. In: Proc. 1st Iberian Conf. on Pattern Recognition and Image Analysis, pp. 1126–1138 (2003)
10. Tumer, K., Ghosh, J.: Error Correlation and Error Reduction in Ensemble Classifiers. Connection Science 8(3-4), 385–403 (1996)
11. Meynet, J., Thiran, J.: Information Theoretic Combination of Classifiers with Application to AdaBoost. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 171–179. Springer, Heidelberg (2007)

Appendix: Expansion of the Ensemble Mutual Information

Theorem 1
Given a set of classifiers $S = \{X_1, ..., X_M\}$, and a class label Y, their Shannon mutual information can be expanded as

$$I(X_{1:M}; Y) = \sum_{T \subseteq S} I(\{T \cup Y\}), \qquad |T| \geq 1. \tag{11}$$

That is, the Shannon Mutual Information between $X_{1:M}$ and Y expands into a sum of Interaction Information terms. Note that $\sum_{T \subseteq S}$ should be read, "sum over all possible subsets T drawn from S".

Proof: See ref [7].

Example: As an illustrative example for an ensemble of size $M = 3$, the Shannon information between the joint variable $X_{1:3}$ and a target Y can be re-written as

$$\begin{aligned} I(X_{1:3}; Y) = I(\{X_1, Y\}) &+ I(\{X_2, Y\}) + I(\{X_3, Y\}) \\ &+ I(\{X_1, X_2, Y\}) + I(\{X_1, X_3, Y\}) + I(\{X_2, X_3, Y\}) \\ &+ I(\{X_1, X_2, X_3, Y\}). \end{aligned} \tag{12}$$

Each term can then be separated into class unconditional $I(\{\boldsymbol{X}\})$ and conditional $I(\{\boldsymbol{X}\}|Y)$ according to the standard definition of interaction information. This gives us the expansion found in the main body of this paper.