An information theoretic perspective on multimodal signal processing
-
ITS Technical Report
TR_ITS_2005.38

Mihai Gurban and Jean-Philippe Thiran
Signal Processing Institute,
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email: {mihai.gurban, jp.thiran}@epfl.ch
web: itswww.epfl.ch

Abstract

Multimodal signals can be defined in general as signals originating from the same physical source, but acquired through different devices, techniques or protocols. This applies for example to audio-visual signals, medical or satellite images. Understanding the joint dependencies of such signals is the first step toward intelligent means for their analysis. Information theory offers a rich theoretical framework in which such dependencies can be emphasized and from this, new methods of signal analysis can be developed.

Measures derived from information theory have been used in classification, in a preprocessing step aimed at finding the relevant features. These features can either be simply chosen from the available data (feature selection) or be obtained as the result of some linear or nonlinear transform (feature extraction). The common point of such techniques is the optimization of information theoretic measures of the obtained features, measures that should relate to the content of relevant information present in them.

Extending these methods to multimodal signals would offer powerful means to understand in which way such signals are related and thus, how to extract the relevant information from them.

1 Information theoretic background

Information theory is a vast domain that has grown from communication theory, with contributions spanning from computer science or physics to statistics. Two of its theorems that are important in the context of signal processing are the data processing inequality and Fano's inequality.

1.1 The data processing inequality

Most of the time, data needs to be preprocessed before it can be analyzed. The data processing inequality proves that no processing of the data can increase the amount of information in it.

The theorem is formally stated as follows [1]: if X, Y and Z are three random variables forming a Markov chain $X \to Y \to Z$ (that is, X and Z are conditionally independent given Y, $p(x,z|y) = p(x|y)p(z|y)$), and $I(X;Y)$, $I(X;Z)$ are the Shannon mutual information values measured between X and Y, respectively X and Z, then $I(X;Y) \geq I(X;Z)$.

Assuming that the relevant information that is sought in the signal is represented by the random variable X and the data by Y, there is a dependency between them. If the data passes through some transformation $g(Y) = Z$, these three random variables form a Markov chain $X \to Y \to g(Y)$. This means that $I(X;Y) \geq I(X;g(Y))$. The inequality shows that no function applied on the data can possibly increase the amount of relevant information, but only decrease it (or keep it unchanged).

Note that this does not mean that preprocessing data is wrong or futile. The inequality just states what can or can not be obtained from the data. Indeed, processing the data can reduce its dimensionality, remove noise or redundancy, but it can never add information that was not there initially. However, knowing how much relevant information was contained in the original data is usually impossible to compute in practice, because of the difficulties of computing information theoretical measures.

1.2 Fano's inequality

When estimating one random variable from the value of another dependent random variable, some errors will be made. Fano's inequality quantifies the probability of making such errors, tying it to the conditional entropy of the two. Formally, if X and Y are two random variables and $H(X|Y)$ is the Shannon conditional entropy between them, then the probability of error when estimating X from Y will verify this inequality [1]:

$$p_e \geq \frac{H(X|Y) - 1}{logN} \tag{1}$$

where N is the number of elements in the range of X.

Assuming again that the information that we are seeking is represented by the random variable X, and the data by Y, we have a quantitative lower bound on the error that is made when trying to infer something from the data.

As an example, X could represent the class labels in the case of supervised classification. Then the bound gives us a measure of the best performance achievable by a hypothetical ideal classifier. However, there are no guarantees that a particular classifier will reach such a performance. In practice, it would be our goal to ensure that Fano's bound is as low as possible, so that any classifier, if properly trained, has a chance of performing well. The bound only gives the lowest theoretical error, and minimizing it ensures that classifiers can potentially reach such a low error.

So minimizing the lower bound of the error does not guarantee a low error. However, if the bound is high, the error will certainly be high regardless of the classifier.

1.3 Mutual information in classification

In the case of supervised classification, several methods of feature selection or extraction use mutual information as a measure of the quality of the features.

In feature selection, the aim is to choose from the data only the features that are relevant to the problem. In the case of extraction, features are obtained as the result of linear or nonlinear transforms applied on the data. The purpose of both selection and extraction is to reduce the dimensionality of the data, while at the same time retaining only the information that is necessary for classification.

Dimensionality reduction is desirable because it makes it easier to train a good classifier. The training process is faster when the data has a lower dimensionality, and the same accuracy may be attained with less samples. If noise is removed from the data through this operation, accuracy may even improve, although we know from the data processing inequality that such transforms on the data can never add information that was not present there before.

Let C be a random variable representing the class labels and X the data. The following relations describe the connections between the entropies of the two variables and their mutual information [1]:

$$\begin{aligned} I(C;X) &= H(C) + H(X) - H(C,X) && (2) \\ &= H(C) - H(C|X) && (3) \\ &= H(X) - H(X|C) && (4) \end{aligned}$$

The same relations can be illustrated on a Venn diagram, as in figure 1. Intuitively, mutual information measures the information shared between C

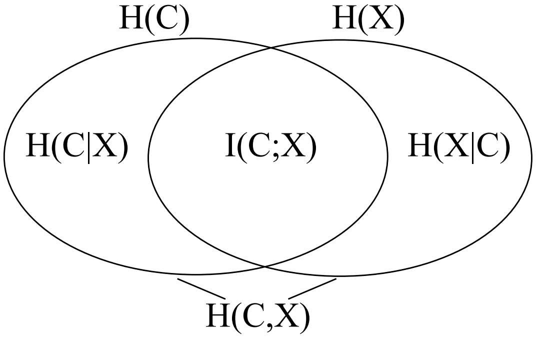

Figure 1: Venn diagram for the mutual information and conditional entropies for class labels (C) and data (X).

and X. If they were independent, their mutual information would be zero. Conversely, if C and X were identical, their mutual information would be equal to the entropy of C (or X) alone.

Applying some transform on the data to obtain features results in a new random variable F=f(X) (feature selection can be considered a particular case of a data transform). The three random variables C, X and F form a Markov chain, $C \rightarrow X \rightarrow F$.

Intuitively, the extracted features should contain as much information about the class labels as possible. So, one criterion for the extraction of features should be to maximize the mutual information between C and F, as this measure shows how much the features can say about the class labels. Indeed, many algorithms for feature selection or extraction take such an approach.

The same reasoning can be justified through Fano's inequality, as:

$$p_e \geq \frac{H(C|F) - 1}{logN} = \frac{H(C) - I(C;F) - 1}{logN} \tag{5}$$

p_e expresses here the probability of making an error when estimating the class labels from the features, that is, when doing the classification. Since the class entropy $H(C)$ is a constant, trying to minimize the lower bound of this error probability is equivalent to maximizing the mutual information. It is evident from the data processing inequality that this maximization can not add information, since $I(C;F) \leq I(C;X)$, but we can try to extract as much information as possible from the data, approaching $I(C;X)$ while keeping the dimensionality low, as illustrated in figure 2.

The conditional entropy $H(C|F)$ plays an important role in Fano's inequality. Indeed, it shows how much information about the class labels can not be inferred from the features, or how hard it is to separate the classes, given the features. When $H(C|F)$ is high, the classes are difficult to discern. Intuitively, this means that the class probability distributions overlap,

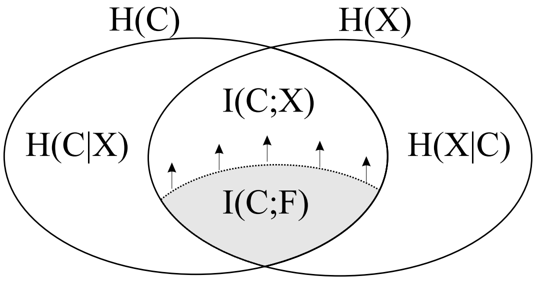

Figure 2: Venn diagram showing how the mutual information between the class labels and the extracted features relates to the MI between the class labels and the original data.

leading to a high error probability. At the other end, a low $H(C|F)$ means the class probability distributions are clearly separated and the probability of error is consequently low.

Considering now the conditional entropy of the features given the classes, $H(F|C)$, we have a measure of how much information in the features is irrelevant to the task. This can be considered noise and should ideally be kept as small as possible, ensuring that the classifier is not supplied with irrelevant data.

As mutual information is hard to compute in a high dimensional space, practically all feature selection or extraction algorithms use approximations to deal with this problem. The MIFS algorithm [2] ("Mutual Information-based Feature Selection") proposed by Batitti in 1994 selects features based on their individual performance, that is, how large is the mutual information between one feature individually and the class label. This avoids the difficulty of computing MI in high dimensions, but may miss features which are only relevant when taken together. The typical example for this is a XOR classification problem (where the class labels are the result of a XOR operation between the binary features), because here, one individual feature is irrelevant to the classification task, which can only be performed when the features are taken together.

To avoid picking redundant features, MIFS penalizes features that are too similar, that is, they have high mutual information between them. Many related algorithms exist, such as those proposed in [3] or [4].

Bollacker [5] prefers the joint mutual information as a criterion, searching pairs of features that together have a high mutual information with the class label. His SMIFE (Separated Mutual Information Feature Extractor) algorithm is similar to PCA, in the sense that a joint mutual information matrix is built and its eigenvectors are found. Analogous to PCA, these eigenvectors should be directions of high mutual information with the class label in the feature space.

A different approach is taken by Principe [6]. In his methodology for

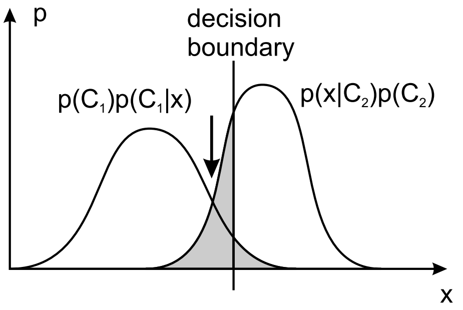

Figure 3: A depiction of the joint probability for a two-class example. In this case, if the vertical line represents the decision boundary, the samples from the grayed area will be misclassified. It can be easily seen that the error is minimal when the boundary is placed on the point where the two probability functions intersect [9].

information theoretic learning, he derives new measures for entropy and mutual information, measures that include the nonparametric probability density estimator implicitly. They are easier to compute, as they are based on the interaction between pairs of samples. His methodology describes in general how to adapt the free parameters of learning machine according to information theoretic criteria. This methodology is applied by Torkkola [7, 8] for feature extraction. He aims to find dimensionality reducing functions, either linear or nonlinear, that maximize the mutual information of the new features with the outputs.

1.4 Bayes error and information theory

If X is the data and C_i are the classes, the posterior probability is defined as $p(C_i|x)$. This is the probability that a sample belongs to the class C_i after having observed the feature vector x. Ideally, a classifier should select the class having the largest posterior probability, that is, choose the class C_k such that $p(C_k|x) > p(C_i|x)$ for any $i \neq k$. It can be proven that this leads to a minimal probability of misclassification [9]. However, the true probability distributions are hard to estimate, so in practice the probability of error will be higher.

From Bayes' rule, the posterior probability is equal to

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)}$$

Since $p(x)$ is independent of the class, the classification rule can also be written as $p(x|C_k)p(C_k) > p(x|C_i)p(C_i)$ for any $i \neq k$. Figure 3 shows how in the simple two-class case this rule leads to a minimum classification error.

In the general case, if the class of a sample is chosen as

$$C_k = \arg\max_i p(C_i|x),$$

the probability of error in the case of the sample x is given by $1-\max p(C_i|x)$. The total probability of error, called *Bayes error*, is the expectation of this measure over all the samples:

$$\begin{aligned} p_e &= E_x\left[1-\max_i p(C_i|x)\right] \\ &= \int p(x)\left[1-\max_i p(C_i|x)\right]dx \\ &= 1-\int \max_i p(x|C_i)p(C_i)dx \end{aligned}$$

Since we used the classification rule that gives the minimal classification error, Bayes error is the smallest error that can be theoretically obtained. No classifier can perform better. However, computing the bound numerically depends on the estimation of the probability density function.

The link to information theory is shown by Vasconcelos [10] through the following bound on Bayes error:

$$p_e \geq \frac{1}{logN}H(C|F) - \frac{log(2N-1)}{logN} + 1, \tag{6}$$

where N is the number of classes.

This bound is very similar to Fano's bound, the only difference between them being an additive constant $\frac{1}{logN}\log\frac{2N}{2N-1}$. The extremes of the two bounds are equal. Furthermore, since the constant is always positive, it follows that the bound of Vasconcelos is higher than Fano's bound. When the number of classes grows, the difference between the two bounds decreases rapidly. A justification of this difference may be the fact that Fano's bound as presented here is a weakened form of a higher error bound. The initial form of the inequality, as found in [1]:

$$H(p_e) + p_e logN \geq H(C|F) \tag{7}$$

does not allow the bound on the probability of error to be expressed explicitly.

In conclusion, minimizing the conditional class entropy $H(C|F)$ or correspondingly maximizing the mutual information $I(C;F)$ leads to minimizing a bound on Bayes error.

2 The gains of multimodal signal processing

2.1 The meanings of "multimodality"

The word *multimodal* is used by researchers in different fields and often with different meanings. There is sometimes a confusion between *multimodal* and *multimedia*. We will present a number of different situations where the word *multimodal* appears, and its associated meaning. We will also emphasize the difference between the meaning given to *multimodal* in the context of *multimodal interfaces* (or systems), compared to that of *multimodal signals*.

Most commonly, the word *multimodal* is used in the field of human-computer (or man-machine) interaction. Here, a *modality* is a natural way of interaction: speech, vision, face expressions, handwriting, hand gestures or even head and body movements. Using several such modalities can lead to *multimodal speaker tracking systems*, *multimodal speech recognizers*, or, more generally, *multimodal interfaces*. Such interfaces aim to facilitate the human-computer interaction, augmenting or even replacing the traditional keyboard and mouse. An example is given by audio-visual speech recognition [11], which aims to improve the performance of audio-only speech recognition through the use of the visual modality.

As will be shown in the following, apparently not all multimodal interfaces or systems use or analyze multimodal signals. The difference will become apparent after the overview of such systems and the modalities employed.

For psychologists, *sensory modalities* represent the human senses (sight, hearing, touch and so on). This is not very different from the previous interpretation. However, other researchers can use the word *modality* in a completely different way. For example, linked to the concept of *medium* as a physical device, *modalities* are the ways to use such *media* [12]. The pen device is a medium, while the actions associated to it, like drawing, pointing, writing or gestures are all modalities.

Generally, the word *multimodal* is associated to the input of information. However, there are cases where the output of the computer is considered multimodal, as is the case of *multimodal speech synthesis*, the augmentation of synthesized speech with animated talking heads.

In the context of medical image registration, a *modality* can be any of a large array of imaging techniques, ranging from anatomical, such as X-ray, MRI or ultrasound, to functional, like fMRI or PET [13]. *Multimodal registration* is the process of bringing images from several such modalities into spatial alignment. The same term is used in remote sensing, where the modalities are images with different spectrums [14] (visible, infrared or microwave for example).

A *multimodal biometrics system* [15] can rely on fingerprints, face, iris, retina, signature, voice or other types of information, all with the purpose of establishing the identity of the user. The modalities named here are not as closely related as, for example, medical images.

With all these different situations, defining *multimodal systems* in general is difficult. All *multimodal systems* extract meaning from multiple sources of information. However this is quite vague and insufficient, as not all systems using multiple input sources can be called multimodal.

Defining *multimodal signals* may be just as difficult. They are defined as signals originating from the same physical source [16] or phenomenon, and thus manifesting some dependency. This dependency is present even if they might have been distorted, affected by noise or changed in other ways which would make it difficult to emphasize their common origin. Note however that this definition excludes some combinations of signals previously mentioned as modalities. Gestures and speech do not have a dependency at the signal level, and the same is true for fingerprints and signatures.

Take for example a pen and voice "put-that-there" interface (which is definitely a multimodal interface). In this case, the signals, voice and pen gestures, are purely complementary, and we would not expect any correlation between the two other than the moment in time when they occur. The two signals do not originate from the same physical source, and they do not convey the same information.

In multimodal biometrics, multimodal databases can be built by joining separate sets of single modality data collected from different persons. This can be done as there is no correlation at the signal level between fingerprints and handwritten signatures (or faces), so the experiments on such databases are still valid.

We shall limit the scope of our discussion to signals that do have a common source, that we expect to have a dependency or to be correlated in some way. From the information theoretic point of view, that means there should be features in the multimodal signals that have a high mutual information between them. This mutual information represents a measure of the information that is shared between the multimodal signals.

2.2 Different goals

But is the common information the only useful information in the multimodal signals? Not necessarily. Most of the time, the information that we seek may be more than just the shared information between the signals. Take for example multimodal medical image registration. There may be details visible in one modality and invisible in another. This complementarity is exactly the reason why multiple modalities are used in the first place. The shared information is the part used for registration, but the complementary information remains in the registered images.

When the goal is feature extraction, the features should contain all (or most of) the information that is relevant to the given task. When samples

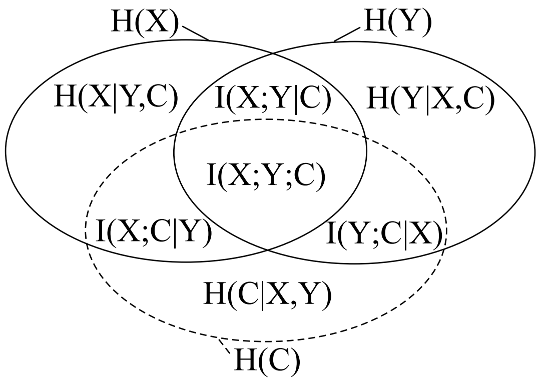

Figure 4: Venn diagram for the mutual information and conditional entropies for class labels (C) and features from two modalities (X and Y).

representing this relevant information are available, they could be used to guide the feature extraction process. An example for this case is audio-visual speech recognition. Here, the relevant information is the phoneme class to which a sample belongs, so the features extracted from both the audio and visual modalities should exhibit a dependency with these phoneme classes. Extracting features from only the part that is common may miss the information that is complementary.

On the other hand, when the goal is the synchronization or alignment of the two signals, the extracted features should be associated to the information that is common to the two signals. This is the case for example for audio-visual speaker tracking. Having a video sequence of people speaking, the goal is to find the region in the image which shows the highest dependency with the sound. This region should correspond to the mouth of the active speaker at a given time, as the movements of the mouth should correlate to the audio [17]. Here, the relevant information is unavailable, as we do not know for sure which features best express the synchrony between the two signals. But it can be assumed that the mutual information is a good indicator of this synchrony.

Making an analogy with machine learning, we could call the first case, the one where the relevant information is known, supervised, and the second unsupervised.

2.3 The supervised case

Assume that, besides the multimodal signals, we also have class labels available. Let X, Y and C be three random variables, the first two representing signals, and the third the class labels. Figure 4 represents the information theoretic measures showing the way information can be distributed between the three. We will present next the significance of each component and how it can be used.

$I(X;Y;C)$ is a measure of how much information is present in both modalities, and is at the same time relevant for our problem. However, being present in both means it can also be considered redundant in one of them. Once identified, features that seem to be tied to this type of information could be picked from only one modality, the one that is more reliable at a given time.

$I(X;C|Y)$ and $I(Y;C|X)$ measure the amount of relevant information is specific to one of the signals X and Y. This is the complementary information that is most of the time the reason why several signals are used. If all the relevant information was common to the signals, then it could be extracted only from one, and the second would be unnecessary.

Take for example the case of audio-visual speech recognition. The class information is in this case represented by phoneme labels. Features from the two modalities that are correlated with these labels are relevant, and they will have a high mutual information with the class label. If they also have a dependency between them, that means that their information is common to the two modalities, and they have some redundancy. If not, then they are complementary, and each feature has specific information about the class label.

$I(X;Y|C)$ is the information present in both signals, but irrelevant to our problem. This could be noise that is accidentally correlated in both signals, but could also be real information that is not needed. Depending on the problem, the class labels may not justify all the dependencies between the modalities. The class labels do not model the underlying physical source of the signal, but rather the information that we seek about that source.

We return to the example of audio-visual speech. There may be some correlation between the audio and video signals which is specific to each individual. This information is not correlated with the phoneme labels, and increases the value of $I(X;Y|C)$. Imagine now that, instead of trying to recognize the spoken words, we want to identify who is speaking. In this case, the audio-visual correlation that is specific to each person becomes relevant information, while the one related to the phonemes becomes irrelevant.

The entropies are also significant. $H(C|X,Y)$ shows how difficult it is to discern the classes, given the features. Its value is linked to the probability of misclassification, as shown by Fano's bound.

$H(X|C,Y)$ and $H(Y|C,X)$ quantify how much irrelevant information is specific to each modality. When extracting features, the corresponding conditional entropies of the features should be kept low.

Two things can be gained from exploiting the multimodality: complementarity and redundancy. Complementarity implies that each modality brings its own contribution, and the result is more useful than individual modalities. Redundancy means that some information is duplicated and, depending on the application, this can be used. The redundancy can lead to better noise tolerance, or even allow a system to function in the extreme case when one of the modalities has been lost completely.

Let us consider now the extraction of features in this context. Let F_X be the features extracted from modality X, and F_Y from Y. In all supervised applications, the extracted features should have a maximum mutual information with the class labels, so $I(C; F_X; F_Y)$, $I(C; F_X|F_Y)$ and $I(C; F_Y|F_X)$ should be simultaneously maximized.

Since the information represented by $I(C; F_X; F_Y)$ appears in both modalities, including it in both F_X and F_Y leads to some redundancy. Depending on the application, this redundancy may either be considered undesirable and minimized, or it may contribute to the system's reliability and tolerance to noise.

An extraction scheme where one modality is favored may also be taken into consideration. If one of the modalities, for example X, is more reliable than another, features could be extracted based mainly on X, representing the information $I(C; F_X)$, while additional features would augment them with the complementary information $I(C; F_Y|F_X)$. A possible example for this is building an augmented audio vector for speech recognition, by adding information from the video stream.

To model this process of feature extraction, two parallel Markov chains can be built:

$$C \rightarrow X \rightarrow F_X \rightarrow \hat{C}$$
$$C \rightarrow Y \rightarrow F_Y \rightarrow \hat{C}$$

Assuming that the modalities are used individually to estimate the class label $\hat{C}$, two bounds can be set on the two error probabilities:

$$p_{eX} \geq \frac{H(C|F_X) - 1}{logN} = B_X$$
$$p_{eY} \geq \frac{H(C|F_Y) - 1}{logN} = B_Y$$

But if the modalities are used together, the error probability will be bounded by:

$$p_{eXY} \geq \frac{H(C|F_X, F_Y) - 1}{logN} = B_{XY}$$

Since conditioning reduces entropy [1],

$$H(C|F_X, F_Y) \leq H(C|F_X) \text{ and } H(C|F_X, F_Y) \leq H(C|F_Y).$$

It follows that the bound obtained when using both modalities, B_{XY}, is lower than both bounds for individual modalities, $B_{XY} \leq B_X$ and $B_{XY} \leq B_Y$.

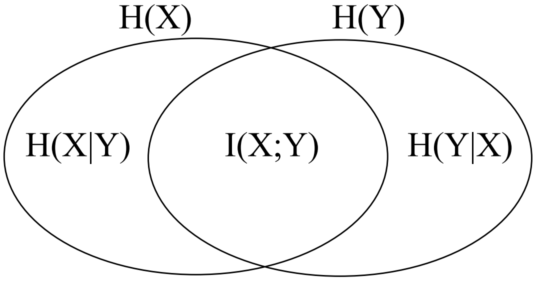

Figure 5: Venn diagram for the mutual information and conditional entropies for features from two modalities (X and Y), when class labels are unavailable.

This shows that the multimodal approach has indeed the potential to lead to a lower error probability.

Since:

$$H(C|F_X, F_Y) = H(C) - I(C; F_X|F_Y) - I(C; F_Y|F_X) - I(C; F_X; F_Y),$$

minimizing the bound B_{XY} means that $I(C; F_X; F_Y)$, $I(C; F_X|F_Y)$ and $I(C; F_Y|F_X)$ should be maximized simultaneously, confirming our previous statement.

Maximizing the mutual information between the features and the class information is a promising concept, but estimating this mutual information is difficult. The reasons for this will be detailed in section 3.

2.4 The unsupervised case

Let us consider the case when class labels are unavailable, in applications such as image registration or speaker tracking. Here, the goal may be to synchronize or align several signals, without knowing the type of information that is relevant to this task. This alignment is achieved when samples considered similar by some criteria originate from the same sampling positions in the different modalities (image registration), or regions of high dependency are identified (speaker tracking). Taking a similar approach as in the previous section, features could be extracted in such a way that the relation between the signals is emphasized.

Figure 5 shows how the information could be distributed between two modalities if we do not know which part of it is relevant. An analogy could be drawn with the supervised case, considering that there is a part of the mutual information that is relevant for our task and another part that is irrelevant, but this time the border between them is hidden.

The mutual information $I(X; Y)$ is a measure of the information shared between the two signals, X and Y. Making a parallel to the previous case, the fact that it is common to both signals may not necessarily mean it is also relevant in its entirety. However, if the two have a common dependency, that will be reflected as a higher value of their mutual information. Determining however how much of the common dependency is due to information that is relevant to the particular problem is difficult without knowing more. For some specific problems, we already know that the common information is important, as is the case for medical image registration. Extracting features with high mutual information between modalities can be further justified with Fano's inequality, as will be shown shortly.

The entropies $H(X|Y)$ and $H(Y|X)$ measure the amount of information characteristic to each of the signals. Again referring to the previous case, this information might not be entirely irrelevant. Some prior knowledge could be added, specific to each modality. Let us take for example a multimodal speaker tracking application. If face detection is used before identifying the speaker, then this means that information that is specific to the visual modality has been used. In this way, information specific to only one of the modalities is used before searching for a dependency between them.

For multimodal medical image registration, typically the mutual information is used to find a transform that maximizes the dependency between the images. However, the registered images still contain all the information, including that which is specific or complementary. For example, in [16], the gradients are used as features for the registration, which is done by maximizing the mutual information, but then the images themselves are used for the diagnostic. This means that although the information that was common to the modalities was used for registration, the complementarity is still present, and, in the end, it is this complementarity of the modalities that justifies the multimodal approach.

As can be seen from these examples, not only the mutual information between the modalities is important. The information characteristic to each modality should be taken into account and used whenever possible.

The signals representing each modality are all sampled signals. When trying to achieve a synchronization of these signals, the goal is to find the sampling positions in one modality corresponding to positions in the other. This is the case in image registration for example. Let us consider O as a random variable representing the sampling positions in the original signals, and $\hat{O}$ as the estimated sampling positions. If X and Y are two different modalities, and F_X, F_Y are the features extracted from these modalities, then, following the approach of Butz [16], two parallel Markov chains can be built:

$$O \rightarrow X \rightarrow F_X \rightarrow \hat{F_Y} \rightarrow \hat{Y} \rightarrow \hat{O}$$
$$O \rightarrow Y \rightarrow F_Y \rightarrow \hat{F_X} \rightarrow \hat{X} \rightarrow \hat{O}$$

Here, $\hat{F_X}$ and $\hat{F_Y}$ are feature values estimated from the other modality.

From these values, the corresponding sampling positions $\hat{O}$ can also be estimated. The probability of error when estimating $\hat{O}$ can be bounded by [16]:

$$p_{e1} = p(\hat{O} \neq O) \geq 1 - \frac{I(F_X; \hat{F_Y}) + 1}{\log N} = B_1$$
$$p_{e2} = p(\hat{O} \neq O) \geq 1 - \frac{I(F_Y; \hat{F_X}) + 1}{\log N} = B_2$$

where N is the number of sampling positions in the range O.

When estimated correctly, the variables $\hat{F_X}$ and $\hat{F_Y}$ will have the same joint probability density functions, $p(\hat{F_X}, F_Y) = p(F_X, F_Y)$ and $p(F_X, \hat{F_Y}) = p(F_X, F_Y)$. This in turn leads to $I(\hat{F_X}; F_Y) = I(F_X; \hat{F_Y}) = I(F_X; F_Y)$, showing that the lower bounds for the probability of error are equal [16]:

$$B_1 = B_2 = 1 - \frac{I(F_X; F_Y) + 1}{\log N}$$

This justifies the use of mutual information when searching for the features that best represent the synchrony of the signals. Such features should have a high mutual information between modalities. However, as will be shown in the next section, it is also desirable that the entropy of these features be small.

2.5 Feature efficiency

When looking for features having maximum mutual information, there is always the danger of adding superfluous information to these features. This is reflected in the joint entropy of the features, $H(F_X, F_Y)$. A high joint entropy means that there is a lot of irrelevant information added in the features.

This is why *efficient* features [16] are desirable, features that have high mutual information, showing that they reflect the relation between the modalities, but at the same time have a low entropy, meaning that there is little superfluous information present. Of course, since $H(F_X, F_Y) \geq I(F_X; F_Y)$, the joint entropy is always higher than the mutual information, but ideally, for highly efficient features, their ratio should be close to one.

The *feature efficiency coefficient* measures precisely this ratio between the mutual information and the joint entropy of features:

$$e(F_X, F_Y) = \frac{I(F_X; F_Y)}{H(F_X, F_Y)} \tag{8}$$

As both values are positive, and $H(F_X, F_Y) \geq I(F_X; F_Y)$, the value of the efficiency coefficient is always between 0 and 1.

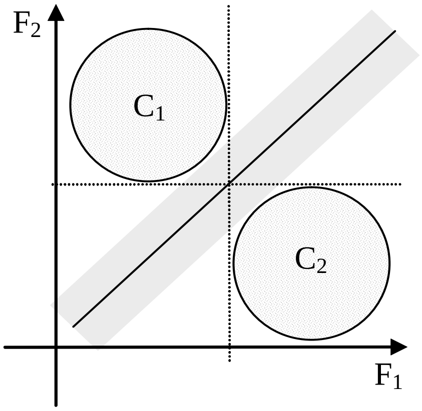

Figure 6: An example of two perfectly separated classes in two dimensions. As the two classes are linearly separable on each of the two directions, one of F_1, F_2 is redundant.

Maximizing the feature efficiency coefficient tries to strike a balance between minimizing the error bound through the maximization of the mutual information, and at the same time keeping the joint entropy low.

2.6 The role of redundancy

Redundancy means that the same information can be present in several different locations. Whether this is desirable or not entirely depends on the context.

In the case of supervised classification, where a low dimensionality is sought, redundancy in the extracted features should be kept low. Indeed, if a feature does not add new information to that already present in the other features, then it is useless. A high mutual information between features is an indicator of redundancy in them.

However, some redundancy can still be useful. Take for example the case in figure 6. Each of the features F1 or F2 can perfectly separate the two classes, so one of them is redundant. But when using only one of them, the margin of separation is small. When taken together, the margin is greatly increased, showing that redundancy in the features can sometimes have a positive effect. In this particular case, the redundancy can be eliminated by transforming the coordinates through a rotation, where one axis would be parallel to the maximum margin separating line, and the other perpendicular. This would lead to one feature that has all the separating power required, and another which is superfluous.

For both the supervised and unsupervised cases, redundancy can be present either between the features of a single modality or between the modalities themselves. Keeping some redundancy between modalities can improve the reliability of the application. If one of the modalities is degraded, the information in the other one can still be used. This redundancy between the modalities is given by the information that is common to them. Such common information uncovers the relation between the modalities, and it is this relation that we seek to interpret.

3 Pitfalls of estimation

3.1 The curse of dimensionality

The term "curse of dimensionality" refers to the fact that, with the increase of dimensionality, the same amount of data becomes less and less informative. For example, 100 observations in a one-dimensional space can be used to build a histogram and draw some conclusions, while in a 10-dimensional space, 100 observations become only isolated points.

Computing the values of entropy and mutual information requires the estimation of (joint) probability density functions, and this estimation is affected by the curse of dimensionality. The number of examples that is required grows exponentially with the dimensionality of the space.

Since computing the mutual information between multidimensional variables is a difficult problem, usually approximations are sought. Most of the time the value of mutual information is either approximated by a sum of low-dimensional terms, or computed directly in a low-dimensional transformed space.

3.2 A worst-case scenario

The values obtained for entropy and mutual information are only estimates based on the limited number of observations available. It may be possible that these estimates are not accurate at all. The following is a worst-case example of how misleading estimates can be.

Assume we want to estimate the values of a random variable X from a related random variable Y, with the help of some transform $g(Y)$, such that $\hat{X} = g(Y)$. Maximizing the mutual information between $g(Y)$ and X should in this case minimize the bound given by Fano's inequality on the probability of error. So a possible approach is to find the transform $g(Y)$ such that $I(X; g(Y))$ is maximized. If g is a very flexible transform however, it could overfit, learning in the worst-case perfectly the values of X from the training set. If $g(Y) = X$ for all the available examples, then the mutual information is indeed maximized, reaching $I(X; X) = H(X)$. However, this contradicts the data processing inequality, $I(X; g(Y)) \leq I(X; Y)$, as $I(X; Y) \leq H(X)$.

In fact, this is only an apparent contradiction. The computed value of the mutual information $I(X; g(Y))$ is only an estimation, given the points in the training set. The real value of $I(X; g(Y))$ should be much smaller, as $g(Y)$ has been overfitted.

This example shows the difference between computed and real mutual information. In this case, the real probability densities of X and $\hat{X}$ are different, while the estimated probabilities are identical. To avoid such problems, some method of regularization is necessary.

While extracting features based on information theoretic measures, Torkkola [8] observed that "flexible non-linear transforms" generalize poorly, especially when there are few data points. Similarly, his conclusion was that regularization is required.

4 Conclusion

Information theory is a very useful tool in the analysis of multimodal signals. Methods that are already used in supervised classification can be adapted to find relevant information in multimodal signals, and expose the relations between modalities.

The use of information theoretic measures can be justified by the minimization of an error bound. Choosing from one modality features that convey a maximum of information, either about the class or about the other modalities, minimizes the lower bound of the error, be it classification or registration error. There are no guarantees that in practice the lower bound can be reached, but its minimization potentially allows the error to decrease. Obviously, if this bound was high, the error would certainly be high.

Estimation also plays a very important role. Without accurate estimation, the information theoretic measures can be misleading, as was shown. Because of the curse of dimensionality, these measures can not be practically computed for high-dimensional data, as the number of required samples is very high.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

[2] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[3] N. Kwak and C. H. Choi, "Improved mutual information feature selector for neural networks in supervised learning," *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 1313–1318, 1999.

[4] M. Deriche and A. Al-Ani, "A new algorithm for EEG feature selection using mutual information," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1057–1060, 2001.

[5] K. D. Bollacker and J. Ghosh, "Linear feature extractors based on mutual information," *Proceedings of the 13th International Conference on Pattern Recognition*, pp. 720–724, 1996.

[6] J. C. Principe, D. Xu, Q. Zhao, and J. W. Fisher III, "Learning from examples with information theoretic criteria," *Journal of VLSI Systems*, vol. 26, pp. 61–77, 2000.

[7] K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. 17th International Conf. on Machine Learning*, pp. 1015–1022, Morgan Kaufmann, San Francisco, CA, 2000.

[8] K. Torkkola, "Nonlinear feature transforms using maximum mutual information," *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 4, pp. 2756–2761, 2001.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[10] N. Vasconcelos, "Feature selection by maximum marginal diversity," *Neural Information Processing Systems*, pp. 1351–1358, 2002.

[11] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in audio-visual speech processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

[12] C. Benoit, J. C. Martin, C. Pelachaud, L. Schomaker, and B. Suhm, "Audio-visual and multimodal speech systems," in *Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume* (D. Gibbon, ed.), 2000.

[13] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[14] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[15] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, 2004.

[16] T. Butz and J. P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, no. 85, pp. 875–902, 2005.

[17] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," *Proceedings of the International Conference on Image and Video Retrieval*, pp. 488–499, 2003.

Contents