Modality Selection for Classification on Time-series Data

Murchana Baruah and Bonny Banerjee
Institute for Intelligent Systems, and Department of Electrical and Computer Engineering
The University of Memphis, Memphis, TN 38152, USA
{mbaruah,bbnerjee}@memphis.edu

ABSTRACT
In this paper, we investigate the optimal modality selection problem for time-series data in the context of late fusion. Multimodal emotion or action recognition is used as a testbed. Widely-used features and classifier are used for each modality drawn from five benchmark datasets. We experimented with four widely-used late fusion methods. From the classification accuracies obtained for all possible combinations of modalities in each dataset, we observe that the accuracy does not always improve with increase in number of modalities. We further show that expected information gain increases monotonically with classification accuracy in an useful interval and hence, can be used for selecting a subset of modalities for late fusion to achieve a high classification accuracy.

CCS CONCEPTS
• Computing methodologies → Machine learning.

KEYWORDS
Emotion, action, recognition, multimodal, fusion, information gain

ACM Reference Format:

1 INTRODUCTION

Real-world time-series data is often multimodal. Learning from multiple modalities facilitates learning a richer representation which helps to make more accurate inference [7, 33]. Each modality is expected to provide unique information, otherwise it would be redundant. A challenge associated with multimodal data classification is to select a subset of modalities to maximize accuracy. This optimal modality selection problem is investigated in areas such as multimedia [2, 20, 45] and wireless sensor networks [18, 27, 36]. This problem has been explored for early fusion (see [1] for review) but rarely for decision or late fusion.

In this paper, we investigate the optimal modality selection problem for time-series data in the context of late fusion. We experimentally evaluate the effect of combining different modalities on emotion recognition accuracy using two benchmark datasets (DEAP [25], HCI Tagging [41]) containing physiological signals, and on action classification accuracy using three benchmark datasets (PAMAP2 [40], UTD MHAD [9], Berkeley MHAD [35]) containing inertial, motion capture and depth data. Several methods have been reported for action or emotion recognition using these benchmark datasets (see for example [6, 10, 25, 35, 43, 44]). However, they did not report on how to select a subset of modalities, and their experiments are limited to a few datasets. Modality selection criteria, such as independence of classifiers [26] and classifier correlation [15], have been used in late fusion models, but were not evaluated on a number of benchmark datasets.

We compare emotion or action recognition accuracy from four late fusion methods and all possible subsets of modalities from each of the five benchmark datasets. We selected candidates for the different components of a late fusion model from the literature based on their wide usage: (1) features extracted from the signal in each modality, (2) classifier, (3) late fusion methods, and (4) modality selection criterion. These candidates allow us to experiment with multiple late fusion models and draw general conclusions.

The contributions of this paper are as follows:

(1) Our experimental results reveal that, contrary to expectation, emotion or action classification accuracy does not always increase with increase in number of modalities for different late fusion methods. More data might confuse a model as data is not always informative.

(2) We empirically show that information gain increases monotonically with classification accuracy in the accuracy interval $[a, 1] \times 100\%$ where $a \in [0, 0.5]$ for two or more classes.

(3) Our experimental results show that information gain is a useful metric for selecting the optimal subset of modalities for late fusion. For the five benchmark datasets and both action and emotion recognition, subsets selected using information gain yield results comparable to the highest classification accuracy.

2 MODELS AND METHODS

Our recognition model consists of four functions: feature extraction, classification, modality selection, and late fusion (see Fig. 1). We use widely-used feature extraction, classification and fusion methods, and modality selection metrics.

Let $\{M_1, \ldots, M_n\}$ be n modalities (or signals), and $x_i$ be the feature vector and $\lambda_i$ be the classifier for modality $M_i$ ($i = 1, \ldots, n$). Let there be m classes, $\{\omega_1, \ldots, \omega_m\}$, such that $P(\omega_k | x_i)$ represent the posterior probability for class $\omega_k$ from classifier $\lambda_i$.

1Typically late fusion models, including the one in this paper, ignore temporal dependencies across modalities as they fuse modalities only after the classifiers have made their decisions. As a result, late fusion models are to some extent indifferent to the heterogeneity of and synchronicity between the different modalities. They allow modality-specific parameters (e.g., sampling rate, window length, feature vector dimension) and representation (e.g., space-time vs. frequency-time) to be chosen for each
Theorem 2.1. Expected information gain is a convex function of classification accuracy. Given the number of classes \( c \), there exists a unique positive real number \( a \) such that the minimum of expected information gain occurs at classification accuracy \( a \). As \( c \to \infty \), \( a \to 0 \).

For the trivial case \( c = 1, a = 1 \). For the nontrivial case \( c = 2, a = 0.5 \) (computed from Eq. 2). Given a particular \( c \), for each classification accuracy \( \{0, 0.1, 0.2, \ldots, 1\} \), we randomly generated at least \( 10^6 \) confusion matrices and computed the corresponding mean (or expected) information gain. The expected information gain as a function of classification accuracy is shown in Fig. 2 for \( c = 2, 3, 7, 25 \). In each case, the expected information gain is a convex function of classification accuracy. As \( c \) increases, the unique location of the minimum of this function decreases. Since classification accuracy cannot be negative, the location is lower bounded by zero.

Figure 1: Block diagram for multimodal time-series classification using late fusion. ‘FE’ refers to feature extraction.

Figure 2: Expected information gain as a function of classification accuracy is shown for 2, 3, 7 and 25 classes from randomly generated confusion matrices.

This theorem entails that, for any given \( c \), in the accuracy interval \([a, 1]\), expected information gain increases monotonically with classification accuracy. Hence, information gain is probabilistically a sound measure of classification accuracy in \([a, 1]\) where \( a \) decreases as \( c \) increases.

We experiment with four methods to fuse the posterior probabilities obtained from each classifier at the decision level: product, average, Bayesian, and majority voting. Let \( Z \) be the pattern to be assigned to one of the \( m \) classes after fusion.

Definition 2.3. Product [6, 16, 24]. Assign \( Z \to \omega_j \) if \( \prod_{i=1}^{n} P(\omega_j|x_i) = \max_{k=1}^{m} \prod_{i=1}^{n} P(\omega_k|x_i) \) [24]. The product rule for fusion assumes that the joint probability distribution of the measurements obtained from the classifiers is conditionally independent.

Definition 2.4. Average [6, 16, 24, 38]. Assign \( Z \to \omega_j \) if \( \frac{1}{n} \sum_{i=1}^{n} P(\omega_j|x_i) = \max_{k=1}^{m} \frac{1}{n} \sum_{i=1}^{n} P(\omega_k|x_i) \) [24]. The average rule for fusion assumes that the prior is equal.

Definition 2.5. Bayesian [19]. Assign \( Z \to \omega_j \) if \( P(\omega_j|x_1, \ldots, x_n) \approx \max_{k=1}^{m} \sum_{i=1}^{n} P(\omega_k|\omega_j, x_i)P(\omega_j|x_i)P(\omega_i|x_i) \) [19], where \( \omega_j \) denotes the predicted class. The probabilities, \( P(\omega_k|\omega_j, \lambda_i) \) and \( P(\lambda_i|x_i) \), can be approximated from the confusion matrix of \( \lambda_i \).

Definition 2.6. Majority voting [24, 28]. Assign \( Z \to \omega_j \) if \( \Sigma_{i=1}^{n} \Delta_{ji} = \max_{k=1}^{m} \Sigma_{i=1}^{n} \Delta_{ki} \) [24]. The term \( \Sigma_{i=1}^{n} \Delta_{ji} \) adds the votes for the class \( \omega_k \) from the individual classifiers. In case of equal votes for multiple classes, the class with the highest posterior probability is selected.

3 EXPERIMENTAL SETUP

3.1 Experiments

3.1.1 Classification

To evaluate classification accuracy with increase in number of modalities through an exhaustive comparison, we construct a tree with \( n \) leaf nodes, each represents a set containing one modality, and the root node represents the set of \( n \) modalities. The tree has \( n \) levels. The \( i^{th} \) level has \( \binom{n}{i} \) nodes, \( i = 1, \ldots, n \). A non-leaf node representing a set \( Q \) containing \( q \) modalities has \( q \) child nodes,
Algorithm 1: Selecting modalities using information gain
1: Inputs: $P_1, \ldots, P_n$.
2: Output: Set $s$.
3: Initialize: $S \leftarrow \{M_{\text{best}}\}$, $S' \leftarrow \{M_1, M_2, \ldots, M_n\} - S$, $M_{\text{best}}$ is the modality yielding highest recognition accuracy.
4: Compute $G_S(1) \leftarrow G(T', T)$ using Eq. 2, where $T'$ are the predicted class labels from $M_{\text{best}}$.
5: for $i = 2$ to $n$ do
6:     for $j = 1$ to $|S'|$ do
7:         Compute predicted class labels, $T' \leftarrow f(\{P_{M_j} : \lambda_{M_j} \in S \cup S'_j\})$ such that $f$ represents any late fusion operation and $S'_j$ denotes the $j^{th}$ element in set $S'$.
8:     Compute information gain $G_S(j) \leftarrow G(T', T)$ using Eq. 2.
9: end for
10: $G_S(i) \leftarrow \max G_S$
11: $k \leftarrow \arg \max G_S$
12: $s \leftarrow S \cup S'_k$
13: $S' \leftarrow S' - S'_k$
14: end for
15: $k \leftarrow \arg \max G_S$
16: Return $s \leftarrow S_{1:k}$.

That data from 31 subjects constitute the train set and data from the remaining subject constitute the test set. The mean classification accuracy from 32-fold cross-validation is reported. The preprocessed dataset available in MATLAB format in the downloaded dataset is used in our experiments. We extract six statistical features for all the 7 modalities considered in our experiments: means and standard deviations of the raw signals, means of absolute values of the first and second differences of the raw signals, and means of absolute values of the first and second differences of the normalized signals [37], from a 6-second window without overlap [42] such that each window constitutes a datapoint.

HCI Tagging [41]: The individual rated scales (1-9) are mapped to three levels of each valence and arousal states such that 1-3 is mapped to low, 4-6 to medium and 7-9 to high, as in [44]. The valence and arousal classification is performed separately. Two-third of the dataset is used for training, as in [43]. The experiments are repeated 10 times. The mean classification accuracy is reported.

We extract 11 statistical features for all the 5 modalities considered in our experiments: the six features as in the DEAP dataset, skewness, kurtosis, min, max, and median [37, 42] from 6-second windows without overlap [42] such that each window constitutes a datapoint.

PAMAP2 [40]: We consider 12 actions for recognition, as in [40]. The train/test set split is obtained by leaving one subject out cross-validation; data from 8 subjects constitute the train set, as in [40]. The mean classification accuracy from 9-fold cross-validation is reported. We extract features from a 2-second window without overlap. Six statistical features as in the DEAP dataset are extracted from sensors 1 and 2. Three features are extracted from sensor 3 as in [9]: mean, variance, and standard deviation. The features are extracted for each dimension of the inertial signals and then concatenated.

UTD MHAD [9]: There are 27 action classes. Data from subjects $\{1, 3, 5, 7\}$ is used for training and subjects $\{2, 4, 6, 8\}$ for testing, as in [9]. The experiments are repeated 10 times and the mean of the classification accuracy over all the experiments is reported. We extract 8 statistical features from each 6-second window, as in [9], for inertial data across all the dimensions: mean, median, max, min, standard deviation, variance, skewness, kurtosis. Variance is extracted from 8 windows, as in [9], for each video for each dimension of the motion capture data. The features are computed over all dimensions, then concatenated over all the dimensions and windows. Depth motion map (DMM) features, as in [9] are extracted from the depth videos.

Berkeley MHAD [35]: There are 11 action classes. Data from the first 7 subjects are used for training and the last 5 for testing, as in [11, 35]. The experiments are repeated 10 times and mean classification accuracy is reported. Each inertial sequence is divided into 30 windows and each accelerometer sequence into 60 windows, as in [8]. Variance is extracted for inertial and accelerometer data from each window and across all the dimensions which is then concatenated over all the dimensions and windows. As in [8], DMM features constitute our feature vector for the depth videos.

4 EXPERIMENTAL RESULTS

Classification accuracy with respect to number of modalities. Our results show that an increase in the number of modalities
Table 1: Classification accuracy (%) reported for the best subset of modalities (i.e. the combination of modalities that yields highest accuracy from all fusion methods) for each of the three modality selection methods ($G, \rho, G_s$). The "Exhaustive" column reports the highest classification accuracy obtained by considering all possible combinations of modalities ($\sum_{i=1}^{n} (\frac{n}{i})$). All modalities present in a dataset are mentioned below the dataset name. The fusion method and the subset of modalities yielding the highest accuracy are reported. The baselines for similar experimental conditions as ours are shown. The highest accuracy for each dataset is highlighted.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exhaustive</th>
<th>$G$</th>
<th>$\rho$</th>
<th>$G_s$</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAP (valence)</td>
<td>75.72 - Product eeg, eog, emg, resp</td>
<td>75.72 - Product eeg, eog, emg, resp</td>
<td>75.41 - Bayesian eeg, eog, emg, plet</td>
<td>75.72 - Product eeg, eog, emg, resp</td>
<td>57.6 [25]</td>
</tr>
<tr>
<td>DEAP (arousal)</td>
<td>66.41 - Bayesian eeg, eog, emg, gsr, resp</td>
<td>66.40 - Bayesian eeg, eog, gsr</td>
<td>65.30 - Bayesian eeg, eog, emg</td>
<td>66.31 - Bayesian eeg, eog</td>
<td>62 [25]</td>
</tr>
<tr>
<td>HCI Tagging (valence)</td>
<td>68.63 - Product eeg, eeg, gsr, resp</td>
<td>68.63 - Product eeg, eeg, gsr, resp</td>
<td>68.63 - Product eeg, eeg, gsr, resp</td>
<td>68.63 - Product eeg, eeg, gsr, resp</td>
<td>56.83 [44]</td>
</tr>
<tr>
<td>HCI Tagging (arousal)</td>
<td>66.36 - Average eeg, eeg, gsr, resp</td>
<td>66.36 - Average eeg, eeg, gsr, resp</td>
<td>66.19 - Average eeg, eeg, gsr, temp</td>
<td>66.36 - Average eeg, eeg, gsr, resp</td>
<td>54.73 [44]</td>
</tr>
<tr>
<td>PAMAP2</td>
<td>90.98 - Product s1, s2, s3</td>
<td>83.57 - Product s1, s2, s3</td>
<td>90.98 - Product s1, s2, s3</td>
<td>90.98 - Product s1, s2, s3</td>
<td>89.24 [39]</td>
</tr>
<tr>
<td>UTD-MHAD depth, skel, iner</td>
<td>92.40 - Product depth, skel, iner</td>
<td>92.40 - Product depth, skel, iner</td>
<td>78.39 - Product skel, iner</td>
<td>92.40 - Product depth, skel, iner</td>
<td>79.1 [9]</td>
</tr>
<tr>
<td>Berkeley MHAD depth, skel, iner</td>
<td>97.05 - Average depth, skel, iner</td>
<td>97.05 - Average depth, skel, iner</td>
<td>94.22 - Average depth, skel, iner</td>
<td>97.05 - Average depth, skel, iner</td>
<td>98.23 [8]</td>
</tr>
</tbody>
</table>

Table 2: Percentage of #Cases (ref. Section 3.1.1) where classification accuracy decreases as a modality is added. The range of accuracy (%), stated within parentheses, is obtained from all possible combination of modalities ($\sum_{i=1}^{n} (\frac{n}{i})$).

<table>
<thead>
<tr>
<th>Method</th>
<th>DEAP (valence)</th>
<th>DEAP (arousal)</th>
<th>HCI Tagging (valence)</th>
<th>HCI Tagging (arousal)</th>
<th>PAMAP2</th>
<th>UTD-MHAD</th>
<th>Berkeley MHAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>47.85 (15.68)</td>
<td>64.17 (8.35)</td>
<td>18.67 (25.83)</td>
<td>16 (23.56)</td>
<td>0 (20.26)</td>
<td>0 (29.19)</td>
<td>0 (16.66)</td>
</tr>
<tr>
<td>Average</td>
<td>47.85 (15.60)</td>
<td>64.40 (8.35)</td>
<td>18.67 (25.44)</td>
<td>14.67 (23.72)</td>
<td>33.33 (13.32)</td>
<td>0 (24.08)</td>
<td>33.33 (18.91)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>34.01 (15.66)</td>
<td>48.30 (8.66)</td>
<td>6.67 (22.62)</td>
<td>6.67 (20.62)</td>
<td>22.22 (11.69)</td>
<td>0 (23.87)</td>
<td>22.22 (17.57)</td>
</tr>
<tr>
<td>Voting</td>
<td>47.39 (15.15)</td>
<td>49.21 (8.41)</td>
<td>46.67 (19.9)</td>
<td>41.33 (20.15)</td>
<td>0 (9.16)</td>
<td>0 (22.50)</td>
<td>0 (14.19)</td>
</tr>
</tbody>
</table>

may not increase the classification accuracy, especially for emotion recognition from physiological signals (ref. Table 1 and Table 2). Adding more modalities for fusion might add noise and create confusion leading to misclassification. The decrease in classification accuracy with increasing modalities is observed for all the fusion methods and all datasets except UTD-MHAD dataset, as shown in Table 2. Of the four fusion methods, Bayesian fusion yields the smallest decrease in accuracy with increase in number of modalities compared to other fusion methods for all the datasets. This is because Bayesian fusion takes into account the uncertainty of the classifiers for each class by combining the classifiers as a weighted combination of the error distribution over the classes.

**Evaluation of different metrics for selecting modalities.** Our results (ref. Table 1) show that, for all the datasets, the classification accuracy obtained from the subset of modalities selected using information gain is closer to the highest accuracy than that selected using correlation degree. Hence, information gain outperforms correlation degree as a criterion for modality selection. The lowest absolute difference between the true highest accuracy ("Exhaustive") and the accuracy obtained using selected modalities, added over all datasets, is 1.79 for product fusion, 1.7 for average fusion, 0.03 for Bayesian fusion, and 0.01 for majority voting, obtained using metrics $G_s, G, \rho$ and $G$ respectively. Since information gain increases monotonically with classification accuracy (ref. Theorem 1), it is a useful metric for selecting a subset of modalities that will yield high accuracy. As shown in Table 1, the subset of modalities selected using different metrics always yields a classification accuracy comparable, if not equal, to the highest accuracy.

The correlation degree criteria initially selects the modality highly correlated with the true class labels, then it selects modalities least correlated with the selected modalities. This helps in reducing redundancy. However, it can reduce relevant information as well which can lower the classification accuracy, as observed from our results. On the other hand, selecting modalities based on their individual score using information gain and filtering them using a threshold, allows selection of modalities highly correlated with the true class labels. This preserves relevant information but might have high redundancy. However, it outperforms correlation degree as a selection criteria, as seen from our experiments.

Information gain modality selection method (ref. Algorithm 1) selects the combination of modalities after fusion that has the highest correlation with the true class label. This yields the highest classification accuracy, comparable with the true best combination in our experiments. Algorithm 1 requires fusing the modalities
before computing the information gain. Hence, it depends on the fusion method while the other two, correlation degree and filtering using information gain, are independent of the fusion method.

5 CONCLUSIONS

In this paper, we investigated the optimal modality selection problem for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.

REFERENCES


In this paper, we investigated the optimal modality selection problem for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.

REFERENCES


In this paper, we investigated the optimal modality selection problem for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.

REFERENCES


[3] B. Banerjee and J. K. Dutta. 2014. SELP: A general-purpose framework for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.

REFERENCES


[3] B. Banerjee and J. K. Dutta. 2014. SELP: A general-purpose framework for time-series data in the context of late fusion. We analyzed multimodal emotion or action classification using four late fusion methods and five benchmark datasets. Our experimental analysis on product, average, Bayesian and majority voting late fusion methods show that the fusion methods perform differently based on the posterior distribution estimated by each modality. Our results show that for different fusion methods, increasing the number of modalities might not necessarily increase the classification accuracy. We analyze multiple methods for selecting a subset of modalities for late fusion and observe that information gain is an useful measure for selecting modalities which is consistent for all the datasets. The classification accuracy obtained from the selected subset of modalities is comparable to the highest accuracy in all cases.