Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion

Dahua Lin[1] and Xiaoou Tang[1,2]

[1] Dept. of Information Engineering,
The Chinese University of Hong Kong, Hong Kong, China
dhlin4@ie.cuhk.edu.hk
[2] Microsoft Research Asia, Beijing, China
xitang@microsoft.com

Abstract. The paper introduces a new framework for feature learning in classification motivated by information theory. We first systematically study the information structure and present a novel perspective revealing the two key factors in information utilization: class-relevance and redundancy. We derive a new information decomposition model where a novel concept called class-relevant redundancy is introduced. Subsequently a new algorithm called Conditional Informative Feature Extraction is formulated, which maximizes the joint class-relevant information by explicitly reducing the class-relevant redundancies among features. To address the computational difficulties in information-based optimization, we incorporate Parzen window estimation into the discrete approximation of the objective function and propose a Local Active Region method which substantially increases the optimization efficiency. To effectively utilize the extracted feature set, we propose a Bayesian MAP formulation for feature fusion, which unifies Laplacian Sparse Prior and Multivariate Logistic Regression to learn a fusion rule with good generalization capability. Realizing the inefficiency caused by separate treatment of the extraction stage and the fusion stage, we further develop an improved design of the framework to coordinate the two stages by introducing a feedback from the fusion stage to the extraction stage, which significantly enhances the learning efficiency. The results of the comparative experiments show remarkable improvements achieved by our framework.

1 Introduction

Pattern recognition in a high dimensional space, such as face recognition, is a challenging problem due to the difficulties brought by "the curse of dimensionality". Hence, it is crucial to extract a compact set of features to describe the samples so that the classification can be performed efficiently and robustly in a feature space of much lower dimension.

In the literatures of learning, feature extraction has been studied extensively. PCA[1] and LDA[2][3][4] are among the most popular algorithms. The former finds a subspace best preserving the sample variations, while the latter seeks a feature space where the ratio between the between-class scattering and the

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part I, LNCS 3951, pp. 68–82, 2006.

within-class scattering is maximized. Though some improved variants[5][6] are proposed, the fundamental limitation of PCA and LDA are yet to be solved: they are solely based on the second order statistical moments, thus may not work well in the practical cases where the distributions are nongaussian.

To break the limitation, we need a method which does not rely on parametric assumptions on the sample distribution. The intrinsic relationship between information theory and pattern recognition, established by the well known Fano's inequality[7], inspires a new way to the feature learning. In the past decade, many works have been done to apply information theory to the learning problems. Some [8][9][10][11] use infomax principle for sequential feature selection. However, they only concern the information conveyed by each individual feature without considering their relation, thus often produce feature sets with a large amount of redundancy. Some improved feature selection algorithms[12][13][14] try to tackle the problem by taking the diversity among the features into consideration. Nonetheless, the criteria of these methods are based on either heuristic rules without convincing justification or some very loose approximations. Hence, the improvement achieved is not significant.

So far the use of information theory in pattern recognition is basically restricted to the feature selection due to two difficulties: 1) No rigorous theory is available to study the inter-feature relation and how the relation affects the performance of the whole feature set; 2) The evaluation of entropy and mutual information incurs great computational difficulties in the optimization. Recently, Torkkola et al.[15][16] propose an infomax feature extraction method to learn a joint set of orthogonal features based on Renyi entropy. However, it suffers from the following drawbacks: 1) The Renyi approximation is not sufficiently justified and what effects it brings to the solution is unclear; 2) It is based on density estimation in a multi-dimensional space, which is computationally expensive and not robust; 3) It does not account for the inter-feature relations.

In this paper, to address the two difficulties, we first systematically investigate the structure of information conveyed by the feature set and present an information decomposition model. It shows that the effectiveness of the feature set is influenced by two key factors: the class relevance and the inter-feature redundancy. As a novel approach, our model also points out that *the redundancy can be factorized into class-relevant and irrelevant ingredients* and introduces the concept *class-relevant redundancy* with theoretically well-founded formulation. We then derive the *Conditional Informative Feature Extraction* algorithm which maximizes the information conveyed by the whole feature set by explicitly reducing the class-relevant redundancies. To attack the computational difficulty, we couple the discrete approximation with the 1D Parzen window technique and further propose a *Local Active Region method*, which substantially reduces the computational cost from $O(n^2)$ to $O(n)$ and thus enables large-scale application of the method.

We also develop the Bayesian Feature fusion algorithm to effectively utilize the feature set by incorporating Laplacian sparse prior and Multivariate logistic regression into the Bayesian MAP formulation, where the features are adap-

tively weighted. Considering that the separate treatment of feature extraction and fusion incurs inefficiency, we finally improve the framework architecture to coordinate the two stages by introducing a feedback from the fusion stage to the extraction stage. By the new design, both the learning efficiency and the effectiveness of the resultant feature set are greatly enhanced.

2 Conditional Informative Feature Extraction

2.1 Problem Formulation and Features

Consider a multiclass classification problem: the training set consists of n samples from C classes, which is denoted by $\{(\mathbf{x}_i, c_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X}$ is a d-dimensional vector representing the i-th sample, c_i is its class label. For discrimination, we extract a set of features, denoted by $F = \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$. Each feature is a functional: $y^{(t)} : \mathcal{X} \to R$, which maps a sample vector to a scalar. For each sample $\mathbf{x}$, all the m feature values constitute a feature vector, denoted by $\mathbf{y}(\mathbf{x}) = \left[y^{(1)}(\mathbf{x}), y^{(2)}(\mathbf{x}), \ldots, y^{(m)}(\mathbf{x})\right]^T$. For succinctness, we denote the features for the i-th training sample by $\mathbf{y}_i = [y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(m)}]^T$.

Linear features are the most widely used features in the literature owning to its simplicity and effectiveness. Each linear feature is parameterized by a projection vector $\mathbf{w}$ subject to $||\mathbf{w}|| = 1$, and the feature value for the sample $\mathbf{x}$ can be extracted by $y = \mathbf{w}^T\mathbf{x}$. In the cases where the sample distribution is highly nongaussian, linear features are insufficient to classify the samples well. To tackle the difficulty, we can extract nonlinear features by kernelization, where a nonlinear mapping ϕ is employed to map the original vector space to a Hilbert space of much higher dimension. Each feature can be regarded as a projection of such mapping. Assume that the projection vector in the Hilbert space can be expanded by $\mathbf{w}^\phi = \sum_{i=1}^n a_i\phi(\mathbf{x}_i)$, then with the kernel trick, the feature value can be computed by $y = \mathbf{a}^T[k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n)]^T$, where $\mathbf{a} = [a_1, \ldots, a_n]$ is the vector of expansion coefficients.

2.2 The Information Maximization Principle

In information theory, the *entropy* of a random feature $\mathbf{y}$, denoted by $H(\mathbf{y})$, contains two-fold meanings: 1) $H(\mathbf{y})$ measures the uncertainty on $\mathbf{y}$, 2)$H(\mathbf{y})$ represents the total information conveyed by $\mathbf{y}$. Based on the notion that information stems from uncertainty, the mutual information $I(\mathbf{x}; \mathbf{y})$ is defined by $I(\mathbf{x}; \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$, which indicates that the information delivered from $\mathbf{x}$ to $\mathbf{y}$ equals the reduction of uncertainty of $\mathbf{y}$ when $\mathbf{x}$ is known. [7] gives a comprehensive treatment to the concepts of information theory.

Intuitively, when we know more about the classes, we can classify the objects more accurately. This rationale leads to the *infomax principle* for feature learning, which advocates to learn features by maximizing the mutual information between the features and the classes. The principle is validated theoretically by Fano's inequality[7]

$$P(\hat{c} \neq c) \geq \frac{H(\mathbf{y}|c) - 1}{\log C} = \frac{H(c) - I(\mathbf{y}; c) - 1}{\log C}, \tag{1}$$

where $\hat{c}$ is the decision made based on the feature vector $\mathbf{y}$, c is the true underlying class. This inequality relates the lower bound of the Bayes error to the mutual information between the features and the classes. Vasconcelos[10] reinforces the relation by showing that: *The infomax solution is near optimal in the minimum Bayes error sense.*

2.3 The Information Decomposition and the Conditional Objective

Since each sample is usually described by multiple features, there may exist some relations between the features. How do the inter-feature relations affect the process of information utilization? To answer this question, we first study the structure of the joint information by examining the two-feature case.

$$H(y^{(1)}) = I(y^{(1)}; c) + H(y^{(1)}|c) \tag{2}$$

$$H(y^{(2)}) = I(y^{(2)}; c) + H(y^{(2)}|c) \tag{3}$$

$$\begin{aligned} H(y^{(1)}y^{(2)}) &= I(y^{(1)}y^{(2)}; c) \\ &\quad + H(y^{(1)}y^{(2)}|c) \end{aligned} \tag{4}$$

$$\begin{aligned} H(y^{(1)}y^{(2)}) &= H(y^{(1)}) + H(y^{(2)}) \\ &\quad - I(y^{(1)}; y^{(2)}) \end{aligned} \tag{5}$$

$$\begin{aligned} I(y^{(1)}y^{(2)}; c) &= I(y^{(1)}; c) + I(y^{(2)}; c) \\ &\quad - [I(y^{(1)}; y^{(2)}) - I(y^{(1)}; y^{(2)}|c)] \end{aligned} \tag{6}$$

Fig. 1. The important formulas characterizing the information structure

Suppose we have two features $y^{(1)}$ and $y^{(2)}$ to represent the samples. Then the information carried by $y^{(1)}$ and $y^{(2)}$ are $H(y^{(1)})$ and $H(y^{(2)})$ respectively. The information conveyed by the joint set of two features is $H(y^{(1)}y^{(2)})$. Based on information theory, we deduce the formulas given in fig.1, which characterize the relations between these quantities and those between information and classification. Though they are simple, however, careful analysis of them leads us to an insightful perspective on the information structure:

1) Eq.(2-4) indicate that the information conveyed by the features consists of two parts: the class-relevant part $I(y; c)$ and the irrelevant part $H(y|c)$. Only the former contributes to classification.
2) Eq.(5) gives another view: when two features are used, the joint information of the feature set would be less than the sum of information conveyed by individual features due to the redundancy, which is measured by the mutual information between the two features.

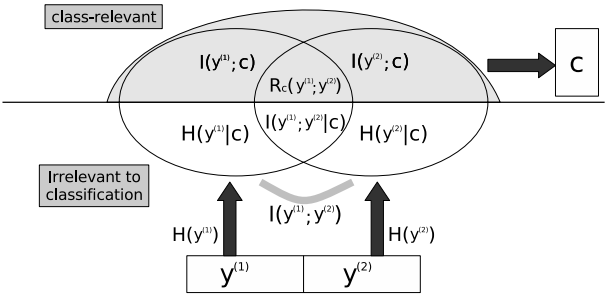

Fig. 2. Illustration of Joint Information Decomposition

3) Eq.(6) combines the class-relevance factor and the redundancy factor to depict the information structure: the class-relevant information conveyed by the joint set is equal to the sum of the individual class-relevant information delivered by $y^{(1)}$ and $y^{(2)}$ minus the *class-relevant redundancy*. For conciseness, we denote it by $R_c(y^{(1)}; y^{(2)}) = I(y^{(1)}; y^{(2)}) - I(y^{(1)}; y^{(2)}|c)$, then Eq.(6) can be rewritten as

$$I(y^{(1)}y^{(2)}; c) = I(y^{(1)}; c) + I(y^{(2)}; c) - R_c(y^{(1)}; y^{(2)}). \tag{7}$$

The fig.2 illustrates the two-feature information decomposition model and gives a clear picture to the information structure.

The Eq.(7) can be generalized to the case of multiple features with mathematical induction. It results in the following theorem:

Theorem 1. *Assume that* $\forall i \neq j, k_1, k_2, \ldots \notin \{i, j\}$ $I(y^{(i)}; y^{(j)}|y^{(k_1)}, y^{(k_2)}, \ldots) = I(y^{(i)}; y^{(j)})$ *and* $I(y^{(i)}; y^{(j)}|c, y^{(k_1)}, y^{(k_2)}, \ldots) = I(y^{(i)}; y^{(j)}|c)$, *then*

$$I(\mathbf{y}; c) = I(y^{(1)}y^{(2)} \cdots y^{(m)}; c) = \sum_{t=1}^{m} I(y^{(t)}; c) - \sum_{t=1}^{m-1} \sum_{u=t+1}^{m} R_c(y^{(t)}; y^{(u)}). \tag{8}$$

The theorem states that when the communication of any two features is not affected by other features, the joint class-relevant information equals the sum of the individual feature information minus the total pairwise redundancies. We can rewrite Eq.(8) by

$$I(\mathbf{y}; c) = \sum_{t=1}^{m} \left[I(y^{(t)}; c) - \sum_{u=1}^{t-1} R_c(y^{(u)}; y^{(t)}) \right]. \tag{9}$$

This form enables us to extract features sequentially, given that $t-1$ features are extracted, the t-th feature can be extracted by optimizing the *Conditional Informative Objective* as

$$\theta_t = \underset{\theta_t}{\operatorname{argmax}} \left\{ I(y^{(t)}; c) - \sum_{u=1}^{t-1} R_c(y^{(u)}; y^{(t)}) \right\}, \tag{10}$$

where θ is the parameter for the t-th feature. Accordingly, the feature extraction algorithm based on Eq.(10) is called *Conditional Informative Feature Extraction*.

Discussion
The significance of the information decomposition model lies in three aspects:

First, it is the first work to present an insightful view into the composition of information with a classification context, where two key factors: *class-relevance* and *inter-feature redundancy* are revealed and analyzed with solid theoretical foundation.
Second, a novel concept called *class-relevant redundancy* is introduced, which serves a key role in the information-oriented classification. This concept reflects the compound influence of class-relevance and redundancy, which has not been discussed in previous literatures.
Third, Eq.(8) integrates the two factors to form an approximation of joint information with the second-order interactions taken into account. The condition when the approximation is exact is also given. This formulation on one hand explicitly exploits the redundancies among features, which plays an important role in learning, on the other hand ignores the higher-order interactions which will lead to exponentially increasing complexity. In this sense, it achieves a good trade-off between the accuracy and the complexity.

3 The Efficient Optimization

According to the Asymptotic Equipartition Property[7], when a reasonably large set of samples are available, the entropy can be approximated by the sample mean as

$$H(y) = -\int_{\mathcal{R}} p(y)\log(p(y))dy = -E\left\{\log(p(y))\right\} \approx -\frac{1}{n}\sum_{i=1}^{n}\log(p(y)). \tag{11}$$

To evaluate $p(y)$, we apply the nonparametric Parzen window technique instead of relying on any parametric assumptions that are often violated in practical cases. Here, we use a Gaussian kernel, defined by $\phi(r) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp(-\frac{r^2}{2\sigma^2})$, and σ controls the width of the kernel. Then the approximation is given by

$$p(y) \approx \frac{1}{n}\sum_{i=1}^{n}\phi(y - y_i) \tag{12}$$

In the following text, we try to unveil the underlying working mechanism of conditional infomax learning by studying the terms in the objective function given in Eq.(10).

1) Class-relevant Information. From Eq.(11) and Eq.(12), we have

$$I(y^{(t)}; c) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log\sum_{j:c_j=c_i}\frac{1}{n_k}\phi(y_i^{(t)} - y_j^{(t)}) - \log\sum_{j=1}^{n}\frac{1}{n}\phi(y_i^{(t)} - y_j^{(t)})\right\}. \tag{13}$$

We observe two types of terms: the terms representing the interactions between the samples in the same class gathered together by log-sum, and the terms representing the interactions between any pair of samples accumulated by negative log-sum. Considering that $\phi(y_i^{(t)}, y_j^{(t)})$ increases when $y_i^{(t)}$ and $y_j^{(t)}$ become closer, maximizing such an objective will agglomerate the feature values from the same class and disperse those from different classes. In this sense, the optimization process pursues a feature space beneficial to discrimination.

2) Redundancy. We have discussed that $I(y^{(u)}; y^{(t)})$ represents the inter-feature redundancy between $y^{(u)}$ and $y^{(t)}$. In the evaluation of the joint distribution $p(y^{(u)}, y^{(t)})$, we employ the parzen window technique with an isotropic 2D gaussian kernel, which can be expressed as $\phi(y^{(u)}y^{(t)}) = \phi(y^{(u)})\phi(y^{(t)})$. Then we have

$$I(y^{(u)}; y^{(t)}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\frac{1}{n}\sum_{j=1}^{n} \phi(y_i^{(t)} - y_j^{(t)})\phi(y_i^{(u)} - y_j^{(u)})}{\left[\frac{1}{n}\sum_{j=1}^{n} \phi(y_i^{(t)} - y_j^{(t)})\right]\left[\frac{1}{n}\sum_{j=1}^{n} \phi(y_i^{(u)} - y_j^{(u)})\right]}. \quad (14)$$

We find that the unit of the formula is "normalized" correlation between the kernel values for feature $y^{(u)}$ and $y^{(t)}$. Considering that the inter-sample relationship are characterized by the kernel values, and the correlation is a typical measurement of similarity, the redundancy is actually represented by the similarity between the inter-sample relations induced by the two features.

To further clarify how it affects the optimization, we introduce the affinity coefficients $\lambda_{ij}^{(u)} = \frac{\phi(y_i^{(u)} - y_j^{(u)})}{\sum_{j=1}^{n} \phi(y_i^{(u)} - y_j^{(u)})}$, which reflects the affinity between the sample i and j in the u-th feature space. Then Eq.(14) can be simplified to be

$$I(y^{(u)}; y^{(t)}) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \log \sum_{j=1}^{n} \lambda_{ij}^{(u)} \phi(y_i^{(t)} - y_j^{(t)}) - \log \sum_{j=1}^{n} \frac{1}{n} \phi(y_i^{(t)} - y_j^{(t)}) \right\}. \quad (15)$$

We can see that the formula assigns heavy weights on the sample-pairs which are close in the u-th feature space. Therefore minimizing the redundancy will encourage these pairs of samples go farther from each other, thus to create an inter-sample relationship in the t-th feature space, which are distinct from that in the u-th feature space.

As discussed before, some part of the total redundancy is irrelevant to classification, we need to subtract the term $I(y^{(u)}; y^{(t)}|c)$ to compensate its effect. Similar analysis can be applied to this term.

3) Derivative. The analysis above shows that all the terms in the objective function can be written in the following form:

$$f(y^{(t)}) = \pm \sum_{i=1}^{n} \log \sum_{j=1}^{n} \omega_{ij} \phi(y_i^{(t)} - y_j^{(t)}), \quad (16)$$

where ω_{ij} are some coefficients dependent on the specific term. For the terms with $\sum_{j:c_j=c_i}$, they can be expressed by Eq.(16) by setting $\omega_{ij} = 0$ when $c_j \neq c_i$.

When $\phi(\cdot)$ is an even function, the derivative w.r.t the feature values is derived as follows

$$\frac{\partial f}{\partial y_i^{(t)}} = \pm \sum_{k=1}^{n} \left[\frac{\omega_{ik}}{\sum_{j=1}^{n} \omega_{ij} \phi(y_i^{(t)} - y_j^{(t)})} + \frac{\omega_{ki}}{\sum_{j=1}^{n} \omega_{kj} \phi(y_k^{(t)} - y_j^{(t)})} \right] \psi(y_i^{(t)} - y_k^{(t)}). \tag{17}$$

With the derivatives given, we can use stochastic gradient descent to optimize the objective function.

3.1 Local Active Region Method

As shown in fig.3, both the potential and the force attenuates drastically as the distance increases. This observation implies that the interactions within a certain region centered at each sample dominates the objective function, which we call "Local Active Region". As a consequence, we can approximate the objective function and its derivative by retaining only the terms reflecting the interactions with the local regions.

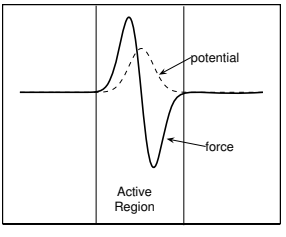

Fig. 3. The potential and the force

Retrieving the neighborhood of every sample is computationally expensive, especially when the sample number is large. Fortunately, we are handling the feature values in a 1-D space, therefore it is feasible to partition the whole value-range into small sections. Here we propose a simple scheme to establish the neighborhoods: suppose the minimum and maximum value of the current feature are y_{min} and y_{max} respectively. Then we divide the range $[y_{min}, y_{max}]$ into sub-sections. The feature values of all samples are categorized into one of the sub-sections. For each sample, the samples residing in the same sub-section constitute its neighborhood. To attain a satisfactory level of accuracy and robustness, the section length is determined so that the average number of samples in each section is about 5.

By employing the simplified way to build neighborhood and discarding the non-neighboring interactions, the time complexity is reduced from $O(n^2)$ to $O(n)$. Such a great improvement in computational efficiency makes the large scale application of infomax learning feasible. Moreover, our algorithm has two important advantages: 1) The Parzen window estimation is performed in 1D and 2D spaces instead of a multidimensional space such as in MMI[15][16], thus it is robust and accurate. 2) The system with only local interactions favors the preservation of local consistency and hence effectively reduces the risk of overfitting.

4 Bayesian Feature Fusion with Sparse Prior

After obtaining the set of features, a question arise naturally: how to combine the features to give the final decision? In many literatures, it is a typical approach to directly compute the Euclidean distance in the feature space, and classify a sample to the nearest class. Though simple, these methods neglect the different contributions of different features thus fails to optimally utilize the features.

In our framework, we assign different weights to different features and evaluate the dissimilarities between samples in the following weighted form:

$$d(\mathbf{y}_i, \mathbf{y}_j) = \sum_{t=1}^{m} b_t \left(y_i^{(t)} - y_j^{(t)} \right)^2. \tag{18}$$

It is known that to achieve a good generalization capability, it is crucial to control the model complexity in order to prevent over-fitting, thus it is desirable to reduce the redundant components by giving a sparse estimating on the coefficients. It has been shown[17] that the Laplacian prior is favorable to sparse estimation.

$$p(\mathbf{b}) \propto \exp\left(\alpha \sum_{t=1}^{m} |b_t| \right). \tag{19}$$

Considering the discriminant learning context, we employ the multivariate logistic regression model to give the conditional likelihood of $\mathbf{b} = [b_1, \ldots, b_m]^T$ as follows

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \mathbf{b}) \propto \prod_{i=1}^{n} p(c_i | \mathbf{y}_i; \mathbf{b}) = \prod_{i=1}^{n} \frac{\exp\left(-d(\mathbf{y}_i, \mathbf{m}_{c_i})\right)}{\sum_{k=1}^{C} \exp\left(-d(\mathbf{y}_i, \mathbf{m}_k)\right)}, \tag{20}$$

where $\mathbf{m}_k$ is the mean vector of the k-th class. By incorporating Laplacian prior and logistic likelihood into the Bayesian MAP learning formulation, we have

$$\mathbf{b} = \underset{\mathbf{b}}{\operatorname{argmax}}\, p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \mathbf{b}) p(\mathbf{b}). \tag{21}$$

A well balance can be achieved between the sparsity and the discriminative power in the learning process. The optimization can be accomplished by Sparse Regression[17][18] proposed by Figueiredo et al.

5 The Integrated Framework for Feature Learning

Traditionally, there are two typical paradigms for feature learning: one first generates a large pool of simple features and then selects a subset from it[8][11][12], while the other directly learns discriminant features from the raw representation and then combines them[1][2][19][15]. They both suffers from a limitation: due to the separate treatment of the two stages, the feature extracted or selected in the 1st stage may not be useful in the fusion or decision stage. Though we can

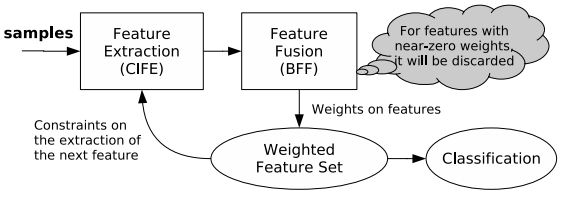

Fig. 4. The Architecture of the Integrated Framework

tackle the problem by extracting a sufficiently large set in the first step, it will inevitably incur considerable waste.

To achieve high efficiency while guaranteeing sufficient expressive power in the feature set, we develop a new framework to coordinate the two stages so that they can intimately cooperate. The whole procedure is introduced as follows:

---

1. Initialize an empty feature set $F \leftarrow \{\}$.
2. Learn the first feature $y^{(1)}$ by the infomax principle; $F \leftarrow F \cup \{y^{(1)}\}$.
3. Repeat the following steps until the stop criterion is met:
  (a) Extract the feature $y^{(t)}$ with the redundancy evaluated on F.
  (b) Add the new feature: $F \leftarrow F \cup \{y^{(t)}\}$.
  (c) Optimize the fusion weights **b**.
  (d) Discard the features with weights smaller than ϵ.

In each step of iteration, we keep monitoring the value of Eq.(8) and stop the loop when the objective function keeps basically unchanged for several iterations.

In the framework, the results of fusion stage are fed back to the extraction stage in order that the extractor can make use of it to evaluate the redundancies based on the fused set and produce an complementary feature as illustrated in fig.4. By eliminating the inactive features, the extractor can find new features adapted to the true demand of the fusion stage without being affected by the unused features, otherwise, the feature set will be gradually filled by the obsolete features and mislead the optimization process by the redundancy terms, thus seriously hinder the effective renewal.

6 Experiments

6.1 A Toy Problem

First, we design a toy problem to give an intuitive insight to the relation between class-relevant information and feature learning in pattern recognition as illustrated in Figure 5. In this experiment, two classes of Gaussian distributed samples are randomly generated, with each class having 500 samples. We extract a series of 1D features by linearly projecting the samples onto 64 different directions. The results clearly show that the class-relevant information, which is the difference between the total entropy and the class conditional entropy, closely

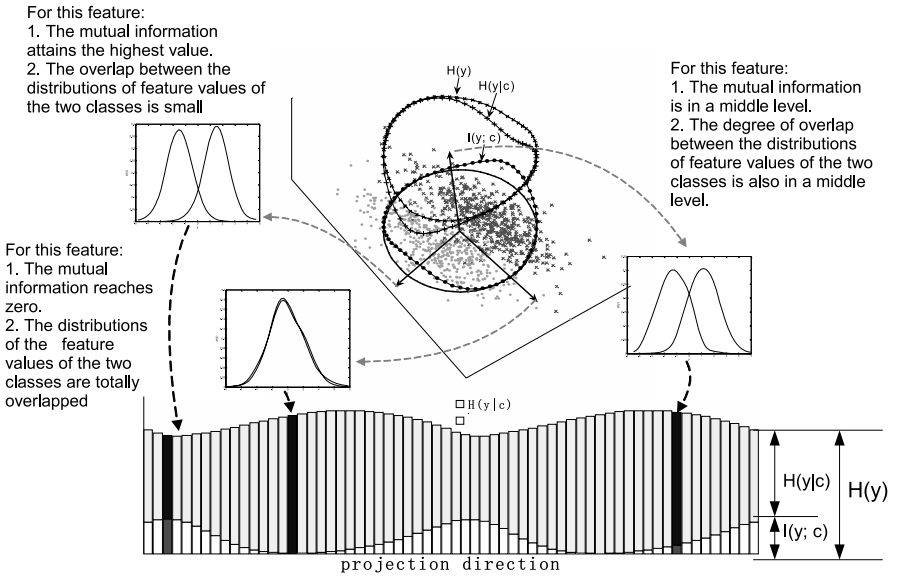

Fig. 5. The Toy Problem. The figure illustrates the relationship between information and feature distribution. The upper part shows a 2D feature space and the 1D distributions of feature values along 3 different directions. The lower part shows the values of the entropy, class-conditional entropy, and mutual information for the features along 64 consecutive directions.

relates to discrimination. From the figure, we can see that for the features with large information values, the distributions of the feature values of the two classes are well separated, while for the features with information values approximating zero, the distributions of the feature values are basically overlapped so that it is difficult to distinguish one class from the other based on that feature. Though the example is simple, it sufficiently exhibits the strong connections between information and classification.

6.2 Face Recognition

Experiment Settings. Face recognition problems is a challenging pattern recognition problem in computer vision, which is a good testbed to evaluate the practical performance of the feature extraction algorithms. To thoroughly test the algorithms, we compare our algorithms with other representative algorithms in face recognition literatures on three standard face databases: FERET[20], XM2VTS[21] and PURDUE AR[22]. To examine the generalization capabilities, for each database, we divide the selected samples into three disjoint datasets: the training set, the gallery set, and the probe set. The training set is for learning the features in the training stage. In the testing stage, every sample in the probe set is compared with each sample in the gallery set, and classified to the person whose gallery sample is most close to it in the feature space. We employ the

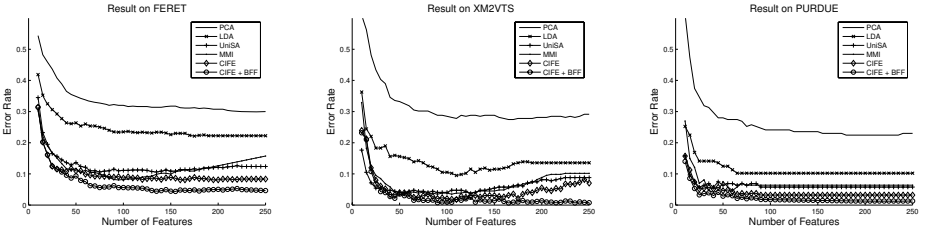

Fig. 6. The Face Recognition Performances for Linear Features

Table 1. The Best Performances of algorithms with Linear Features

Error rate	PCA	LDA	UniSA	MMI	CIFE	CIFE+BFF
FERET	0.299	0.175	0.087	0.079	0.065	0.044
XM2VTS	0.275	0.095	0.037	0.034	0.017	0.007
PURDUE	0.235	0.148	0.057	0.052	0.031	0.022

error rates to measure the performance of the algorithms. In detail, for FERET, we use all the 295 persons with $3-4$ samples for each person to form the training set, which has totally 995 samples. We then select another 800 persons for testing, where the gallery is composed of 800 (fa) samples from different persons, and the probe set is composed of 800 (fb) samples; For XM2VTS, the face images from 295 persons are captured in 4 different sessions. We assign the 295×3 samples captured in the session $1, 2, 3$ to the training set, the 295 samples from the session 1 to the client set, and the 295 samples from the session 4 to the probe set; For PURDUE, there are 90 persons who have the samples captured in all the 26 different conditions. We select 6 samples from each person with diverse expressions and illumination conditions to the training set, a sample captured in normal condition to the gallery set, and another 6 samples captured in different conditions to the probe set. The samples with extreme lighting condition and occlusion are not used in the experiment.

All face images are pre-processed. For each image, we first align it by affine transform to fix the positions of the eye centers and the mouth center, and crop it to the size of 64×72, and then perform histogram equalization to normalize the pixel values. After that, we use a mask to eliminate the background pixels. The remaining 4114 pixels are scanned in order to form the original vector representation of the face. To enhance efficiency and robustness, we use PCA to reduce the dimension and suppress the noise. 99% of the variational energy is preserved in the principal subspace after dimension reduction.

Linear Features. We compare our algorithms with other representative algorithms for feature extraction including PCA[1], LDA[2], Unified Subspace Analysis(UniSA)[4], Maximum Mutual Information(MMI) Algorithm proposed by Torkkola[15]. To clarify the contributions of different components of the framework, we test our algorithms in two different configurations. In a sim-

Table 2. The Best Performances of algorithms with Kernelized Features

Error rate	Kernels	PCA	LDA	UniSA	MMI	CIFE	CIFE+BFF
FERET	Poly 2	0.266	0.162	0.062	0.055	0.042	0.032
	Poly 4	0.267	0.150	0.051	0.052	0.042	0.027
	Sigmoid	0.271	0.142	0.057	0.051	0.037	0.022
	Gauss	0.265	0.134	0.051	0.055	0.032	0.017
XM2VTS	Poly 2	0.264	0.078	0.017	0.034	0.014	0.003
	Poly 4	0.264	0.075	0.017	0.014	0.014	0.013
	Sigmoid	0.254	0.085	0.017	0.014	0.014	0.003
	Gauss	0.258	0.064	0.014	0.007	0.000	0.000
PURDUE	Poly 2	0.241	0.139	0.056	0.035	0.022	0.017
	Poly 4	0.224	0.122	0.044	0.039	0.020	0.011
	Sigmoid	0.220	0.131	0.043	0.041	0.020	0.009
	Gauss	0.222	0.128	0.044	0.033	0.015	0.007

ple configuration, we merely use the Conditional Infomax Feature Extraction (CIFE) to extract features and simply use the Euclidean distance in the feature space to measure the dissimilarities between samples. In a full-functional configuration (CIFE + BFF), we further incorporate the Bayesian Feature Fusion scheme and follow the whole procedure of the integrated framework. The results obtained using different numbers of features are illustrated in Figure 6 and the best results for each algorithm are reported in the Table 1. We can see from the results that the algorithms based on infomax principle outperforms other ones. The CIFE consistently achieves better accuracies than the MMI. By incorporating the Maximum Information Fusion and dynamically discarding the obsolete features, both the accuracy and the robustness of the framework are further enhanced.

Kernelized Features. We also investigate the performances of the algorithms for nonlinear features based on their kernelized versions. The results are given in Table 2. The results of nonlinear feature extraction further validates the effectiveness of our framework. Moreover, we can see that with the adaptive weighting scheme employed, the CIFS + BFF framework has a desirable property that the performance will not degrade with the increasing of the feature numbers as in conventional approaches. The results also confirm the observation in previous works that kernelization can lead to better performance in real data, where the distributions are often nongaussian. By combining the kernel learning and infomax learning and incorporating an effective fusion stage, our framework achieves near perfect classification performance in all the 3 databases.

7 Conclusion

We have presented a novel information-theoretical perspective on the supervised learning and carefully studied the two key factors: class-relevance and redundancy. We introduced a new framework effectively unifying two novel algorithms:

Conditional Informative Feature Extraction and Bayesian Feature Fusion. The results of extensive experiments have sufficiently demonstrated the superiority of our framework over other state-of-the-art approaches.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region and a joint grant (N_CUHK409-03) from HKSAR RGC and China NSF. The work was done in The Chinese University of Hong Kong.

References

1. M. Turk, A. Pentland: Eigenfaces for Recognition. J. Cognitive Neuroscience **3**(1) (1991) 71–86
2. P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on PAMI **19**(7) (1997) 711–720
3. K. Etemad, R. Chellappa: Discriminant Analysis for Recognition of Human Face Images. J. Opt. Soc. Am. **14**(8) (1997) 1724–1733
4. X. Wang, X. Tang: A Unified Framework for Subspace Face Recognition. IEEE Trans. on PAMI **26**(9) (2004) 1222–1228
5. R.P.W. Duin, R. Haeb-Umbach: Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria. IEEE Trans. on PAMI **23**(7) (2001) 762–766
6. M. Loog, R.P.W. Duin: Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. IEEE Trans. on PAMI **26**(6) (2004) 732–739
7. T. M. Cover, J. A. Thomas: Elements of Information Theory. John Wiley Sons, Inc. (1991)
8. Y. Yang, J. O. Pedersen: A Comparative Study on Feature Selection in Text Categorization. In: ICML'97. (1997)
9. N. Kwak, C. Choi: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Trans. on PAMI **24**(12) (2002) 1667–1671
10. N. Vasconcelos: Feature Selection by Maximum Marginal Diversity. In: NIPS'02. (2002)
11. Y. Wu, A. Zhang: Feature Selection for Classifying High-Dimensional Numerical Data. In: CVPR'04. (2004)
12. H. Peng, F. Long, C. Ding: Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. on PAMI **27**(8) (2005) 1226–1238
13. P. Mitra, C.A. Murthy, S.K. Pal: Unsupervised Feature Selection Using Feature Similarity. IEEE Trans. on PAMI **24**(3) (2002) 301–312
14. N. Vasconcelos, M. Vasconcelos: Scalable Discriminant Feature Selection for Image Retrieval and Recognition. In: CVPR'04. (2004)
15. K. Torkkola, W. M. Campbell: Mutual Information in Learning Feature Transformations. In: ICML'00. (2000)
16. K. Torkkola: Feature Extraction by Non-Parametric Mutual Information Maximization. J. Machine Learning Research (2003) 1415–1438

17. M.A.T. Figueiredo: Adaptive Sparseness for Supervised Learning. IEEE Trans. on PAMI **25**(9) (2003) 1150–1159
18. B. Krishnapuram, A.J. Hartemink, L. Carin, M.A.T. Figueiredo: A Bayesian Approach to Joint Feature Selection and Classifier Design. IEEE Trans. on PAMI **26**(9) (2004) 1105–1111
19. A. Hyvarinen, E. Oja: Independent Component Analysis: Algorithms and Applications. Neural Networks **13**(4-5) (2000) 411–430
20. P. J. Phillips, H. Moon, S. A. Rizvi, P. J. Rauss: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Trans. PAMI **22**(10) (2000) 1090–1104
21. K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre: XM2VTSDB: The Extended M2VTS Database. In: Proc. of Int.l Conf. Audio- and Video-based Person Authentication. (1999)
22. A.M. Martinez, R. Benavente: The AR Face Database. CVC Technical Report 24, Purdue University (1998)