

Feature Selection for Gender Classification

No Author Given

No Institute Given

Abstract. Most existing feature selection methods focus on ranking features based on an information criterion to select the best K features. However, several authors find that the optimal feature combinations do not give the best classification performance [6],[5]. The reason for this is that although individual features may have limited relevance to a particular class, when taken in combination with other features it can be strongly relevant to the class. In this paper, we derive a new information theoretic criterion that called multidimensional interaction information (MII) to perform feature selection and apply it to gender determination. In contrast to existing feature selection methods, it is sensitive to the relations between feature combinations and can be used to seek third or even higher order dependencies between the relevant features. We apply the method to features delivered by principal geodesic analysis (PGA) and use a variational EM (VBEM) algorithm to learn a Gaussian mixture model for on the selected feature subset for gender determination. We obtain a classification accuracy as high as 95% on 2.5D facial needle-maps, demonstrating the effectiveness of our feature selection methods.

1 Introduction

High-dimensional data pose a significant challenge for pattern recognition. The most popular method for reducing dimensionality are variance based subspace methods such as PCA. However, the extracted PCA feature vectors only capture sets of features with a significant combined variance, and this renders them relatively ineffective for classification tasks. Hence it is crucial to identify a smaller subset of features that are informative for classification and clustering. The idea underpinning feature selection is to select the features that are most relevant to classification while reducing redundancy. Mutual information (MI) provides a principled way of measuring the mutual dependence of two variables, and has been used by a number of researchers to develop information theoretic feature selection criteria. For example, Batti [2] has developed the Mutual Information-Based Feature Selection (MIFS) criterion, where the features are selected in a greedy manner. Given a set of existing selected features S , at each step it locates the feature x_i that maximize the relevance to the class $I(x_i; C)$. The selection is regulated by a proportional term $\beta I(x_i; S)$ that measures the overlap information between the candidate feature and existing features. The parameter β may significantly affect the features selected, and its control remains an open problem. Peng et al [9] on the other hand, use the so-called Maximum-Relevance

Minimum-Redundancy criterion (MRMR), which is equivalent to MIFS with $\beta = \frac{1}{n-1}$. Yang and Moody’s [11] Joint Mutual Information (JMI) criterion is based on conditional MI and selects features by checking whether they bring additional information to an existing feature set. This method effectively rejects redundant features. Kwak and Choi [8] improve MIFS by developing MIFS-U under the assumption of a uniform distribution of information for input features. It calculates the MI based on a Parzen window, which is less computationally demanding and also provides better estimates.

However, there are two limitations for the above MI feature selection methods. Firstly, they assume that every individual relevant feature should be dependent with the target class. This means that if a single feature is considered to be relevant it should be correlated with the target class, otherwise the feature is irrelevant [4]. So only a small set of relevant features is selected, and larger feature combinations are not considered. Hence, although a single feature may not be relevant, when combined with other features it can become strongly relevant. The second weakness is that most of the methods simply consider pairwise feature dependencies, and do not check for third or higher order dependencies between the candidate features and the existing features. For example, [1] there are four features X_1, X_2, X_3, X_4 , the existing selected feature subset is $\{X_1, X_4\}$, assume $I(X_2, C) = I(X_3, C)$, $I(X_2, X_1|C) = I(X_3, X_1|C)$, $I(X_2, X_4|C) = I(X_3, X_4|C)$ and $I(X_1, X_4, X_2) \gg I(X_1, X_2) + I(X_4, X_2)$, which indicates that X_2 has strong affinity with the joint subset $\{X_1, X_4\}$, although has smaller individual affinity to each of them. So in this situation, X_2 may be discarded, and X_3 is selected, although the combination $\{X_1, X_4, X_2\}$ can produce a better cluster than $\{X_1, X_4, X_3\}$.

To overcome this problem in this paper, we introduce the so called multidimensional interaction information (MII) to select the optimal subset of features. This criterion is capable of detecting the relationships between higher order feature combinations. In addition, we apply a mixture Gaussian model and the variational EM algorithm to the selected feature subset to detect clusters. We apply the method to gender classification based on facial needle-maps extracted from the Max-Planck database of range images.

2 Information-Theoretic Feature Selection

Mutual Information: In accordance with Shannon’s information theory [10], the uncertainty of a random variable C can be measured by the entropy $H(C)$. For two variables X and C , the conditional entropy $H(C|X)$ measures the remaining uncertainty about C when X is known. The mutual information (MI) represented by $I(X; C)$ quantifies the information gain about C provided by variable X . The relationship between $H(C), H(C|X)$ and $I(X; C)$ is $I(X; C) = H(C) - H(C|X)$.

For training a classifier, we prefer features which can minimize the uncertainty on the output class set C . If $I(X; C)$ is large, this implies that feature vector X and output class set C are closely related. When X and C are inde-

pendent, the MI of X and C goes to zero, and this means that the feature X is irrelevant to class C . As defined by Shannon, the initial uncertainty in the output class C is expressed as:

$$H(C) = - \sum_{c \in C} P(c) \log P(c) . \quad (1)$$

where $P(c)$ is the prior probability over the set of class C . The remaining uncertainty in the class set C if the feature vector X is known is defined by the conditional entropy $H(C|X)$

$$H(C|X) = - \int_x p(x) \left\{ \sum_{c \in C} p(c|x) \log p(c|x) \right\} dx . \quad (2)$$

where $p(c|x)$ denotes the posterior probability for class c given the input feature vector x . After observing the feature vector x , the amount of additional information gain is given by the mutual information (MI)

$$I(X; C) = H(C) - H(C|X) = \sum_{c \in C} \int_x p(c, x) \log \frac{p(c, x)}{p(c)p(x)} dx . \quad (3)$$

Conditional Mutual Information: Assume that S is the set of existing selected features, X is the set of candidate features, $S \cap X = \emptyset$, and C is the output class set. The next feature in X to be selected is the one that maximizes $I(C; x_i|S)$, i.e. the conditional mutual information (CMI) which can be represented as $I(C; x_i|S) = H(C|S) - H(C|x_i, S)$, where C is the output class set, S is the selected feature subset, X is the candidate feature subset, and $x_i \in X$. From information theory, the conditional mutual information is the expected value of the mutual information between the candidate feature x_i and class set C when the existing selected feature set S is known. It can be also rewritten as

$$I(C; x_i|S) = E_S(I(C; x_i|S)) = \sum_S \sum_{c \in C} \sum_{x_i \in X} P(x_i, S, c) \log \frac{P(S)P(x_i, S, c)}{P(x_i, S)P(S, c)} . \quad (4)$$

Multidimensional Interaction Information for Feature Selection: The conditioning on a third random variable may either increase or decrease the original mutual information. That is, the difference $I(X; Y|Z) - I(X; Y)$, referred to as the interaction information and represented by $I(X; Y; Z)$, can measure the difference between the original mutual information $I(X; Y)$ when a third random variable is taken into account or not. The difference may be positive, negative, or zero, but it is always true that $I(X; Y|Z) \geq 0$.

Given the existing selected feature set S , the interaction information between the output class set and the next candidate feature x_i can be defined as $I(C; x_i; S) = I(C; x_i|S) - I(C; x_i)$. From this equation, the interaction information measures the influence of the existing selected feature set S on the amount of information shared between the candidate feature x_i and the output class set C , i.e. $\{C, x_i\}$. A zero value of $I(C; x_i; S)$ means that the information

contained in the observation x_i is not useful for determining the output class set C , even when combined with the existing selected feature set S . A positive value of $I(C; x_i; S)$ means that the observation x_i is independent of the output class set C , so $I(C, x_i)$ will be zero. However, once x_i is combined with the existing selected feature set S , then the observation x_i immediately becomes relevant to the output class set C . As a result $I(C; x_i|S)$ will be positive. As a result the interaction information is capable of solving *XOR gate* type classification problems. A negative value of $I(C; x_i; S)$ indicates that the existing selected feature set S can account for or explain the correlation between $I(C; x_i)$. As a result the shared information between $I(C; x_i)$ is decreased due to the additional knowledge of the existing elected feature set S .

According to the above definition, we propose the following multidimensional interaction information for feature selection. Assume that S is the set of existing selected feature sets, X is the set of candidate features, $S \cap X = \emptyset$, and C is the output class set. The objective of selecting the next feature is to maximize $I(C; x_i|S)$, defined by introducing the multidimensional interaction information:

$$I(C; x_i|S) = I(C; x_i) + I(\{x_i, s_1 \dots s_m, C\}) . \quad (5)$$

where

$$I(C; x_i, S) = I(x_i, s_1, \dots, s_m; C) = \sum_{s_1, \dots, s_m} \sum_{c \in C} P(x_i, s_1, \dots, s_m; c) \times \log \frac{P(x_i, s_1, \dots, s_m; c)}{P(x_i, s_1, \dots, s_m)P(c)} . \quad (6)$$

Consider the joint distribution $P(x_i, S) = P(x_i, s_1, \dots, s_m)$. By the chain rule of probability, we expand $P(x_i, S)$, $P(x_i, S; C)$ as

$$\begin{aligned} P(x_i, s_1, \dots, s_m) &= P(s_1)P(s_2|s_1)P(s_3|s_2, s_1) \times P(s_4|s_3, s_2, s_1) \cdots P(x_i|s_1, s_2 \dots s_m) \quad (7) \\ P(x_i, S; C) &= P(x_i, s_1, s_2 \dots s_m; C) = P(C)p(s_1|C)P(s_2|s_1, C)P(s_3|s_1, s_2, C) \\ &\quad \times P(s_4|s_1, s_2, s_3, C) \cdots P(x_i|s_1, \dots, s_m, C) . \end{aligned} \quad (8)$$

There are two key properties of our proposed definition in Equation 5. The first is that the interaction information term $I(\{x_i, s_1 \dots s_m, C\})$ which can be zero, negative and positive. It can deal with a variety of cluster classification problems including the *XOR gate*. When it taken on a negative value, it can help to select optimal feature sets. The second benefit is its multidimensional form, compared to most existing MI methods which only check for pairwise feature interactions. Our definition can be used to check for third and higher order dependencies between features.

We use a greedy algorithm to locate the k optimal features from an initial set of n features. Commencing with a complete set X and an empty set S , we select a feature which gain a large value of $I(x, C)$. We then apply the greedy strategy to select the feature that maximizes the multidimensional mutual information between the features and the output class set, i.e. the first feature

X'_{max} that maximizes $I(X', C)$, the second selected feature X''_{max} that maximizes $I(X'', X', C)$, the third feature X'''_{max} that maximizes $I(X''', X'', X', C)$. We repeat this procedure until $|S| = k$.

3 Experimental Results

2.5D Facial Needle-maps: The data set used to test the performance of our proposed algorithm is the facial needle-maps extracted from the Max-Planck database of range images, see Fig. 1. It comprises 200 (100 females and 100 males) laser scanned human heads without hair. The facial needle-maps (fields of surface normals) are obtained by first orthographically projecting the facial range scans onto a frontal view plane, aligning the plane according to the eye centers, and then cropping the plane to be 256-by-256 pixels to maintain only the inner part of the face. The surface normals are then obtained by computing the height gradients of the aligned range images. We then apply the principal geodesic analysis (PGA) method to the needle maps to extract features. Whereas PCA can be applied to data residing in a Euclidean space to locate feature subspaces, PGA applies to data constrained to fall on a manifold. Examples of such data are direction fields, and tensor-data.

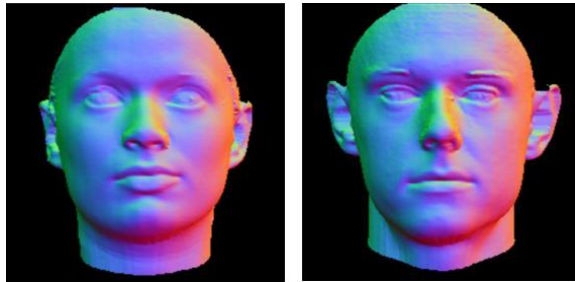


Fig. 1. Examples of Max - Planck needle - maps

Gender Classification Results: Using the feature selection algorithm outlined above, we first examine the gender classification performance using different sized feature subsets for the leading 10 PGA features. Then, by selecting a different number of features from the leading 30 PGA feature components, we compare the gender classification results from our proposed method Multidimensional Interaction Information (MII) with those obtained using alternative existing MI-based criterion methods, namely a) Maximum-Relevance Minimum-Redundancy (MRMR), b) Mutual Information Based Feature Selection (MIFS). The data used are the PGA feature vectors extracted from the 200 Max-Planck facial needle maps.

First, we explore the leading 10 PGA feature components. We have 10 selected feature subsets of size d ($d = 1, \dots, 10$) shown in Fig. 2 (b). For each d ,

we apply the variational EM algorithm [3] to fit a mixture of Gaussians model to the d -dimensional feature vectors of the 200 needle maps. After learning the mixture model, we use the a posteriori probability to perform gender classification. The gender classification accuracies obtained on different feature subsets are shown in Fig. 2 (a). From Fig. 2, it is clear that the best classification accuracy for the facial needle-maps is achieved when 5 features are selected. Using fewer or more features both deteriorate the accuracy. The main reason for this is that we select the gender discriminating features only from the leading 10 PGA features. On the base of cumulative reason, there are probably additional non leading features are important for gender discrimination. Therefore, in the following experiments, we extend our attention to the leading 30 features.

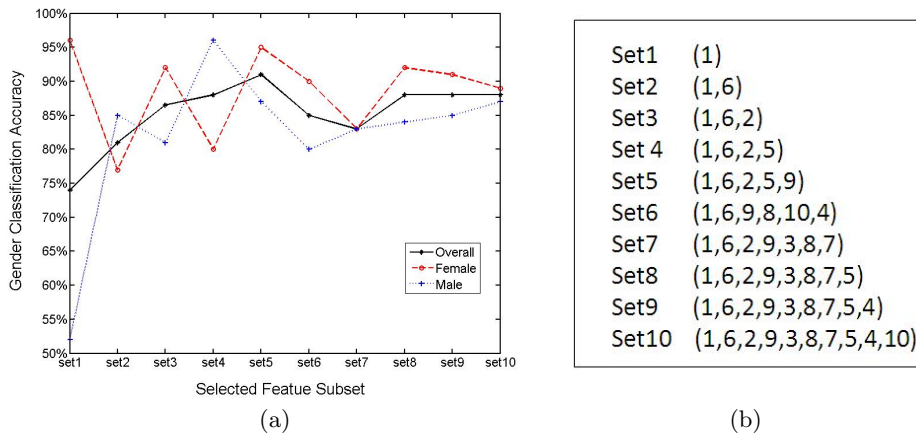


Fig. 2. Gender Classification on Facial needle-maps using different selected feature subsets

In Fig. 3, we compare the the performance of the three criterion functions. At small dimensionality there is little difference between the different methods. However, at higher dimensionality, the features selected by MII clearly have a higher discriminability power than the features selected by MRMR and MIFS. The ideal size of the feature subsets for MIFS and MRMR is 4 or 5, where an accuracy 87% and 92.5% respectively are achieved. However, for MII we obtain the highest accuracy of 95% when nearly 10 features are selected. This means that with both both MRMR and MIFS there is a tendency to overestimate the redundancy between features, since they neglect the conditional redundancy term $I(x_i, S|C)$. As a result some important features can be discarded, which in turn leads to information loss.

Our proposed method therefore leads to an increase in performance at high dimensionality. By considering higher order dependencies between features, we avoid premature rejection on the basis of redundancy, a well documented problem

known to exist in MRMR and MIFS [7]. This gives feature sets which are more informative to our classification problem.

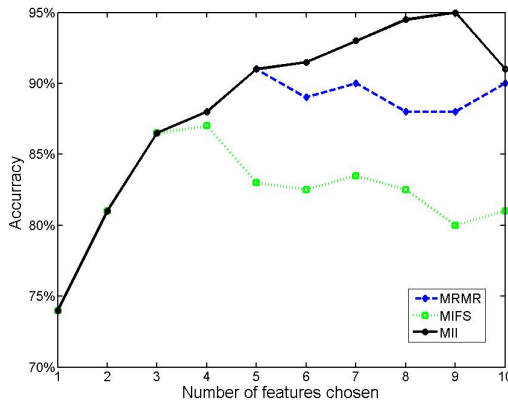


Fig. 3. Average classification accuracies: MII show significant benefit compared to criteria of MRMR and MIFS measuring only pairwise dependencies

We illustrate the learning process and the classification results of using the selected feature subset. The learning process of the EM step in the VBEM algorithm is visualized in Fig. 4 by showing the models learned, a) on initialization, b) on convergence. Each of the two stages is visualized by showing the distribution of the 1st and 6th PGA feature components as a scatter plot. From Fig. 4, it is clear that on convergence of the VBEM algorithm (after 50 iterations), the data are well clustered according to gender. We see that after convergence, only two components survive. In other words, there is an automatic trade-off in the Bayesian model between fitting the data and the complexity of the model, in which the complexity penalty arises from components whose parameters are forced away from their prior values.

4 Conclusions

This paper presents a new information criteria based on Multidimensional Interaction information for feature selection. The proposed feature selection criterion offers two major advantages. First, the MII criteria takes into account high-order feature interactions, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved. Second, the variational EM algorithm and a Gaussian mixture model are applied to the selected feature subset improving the overall classification.

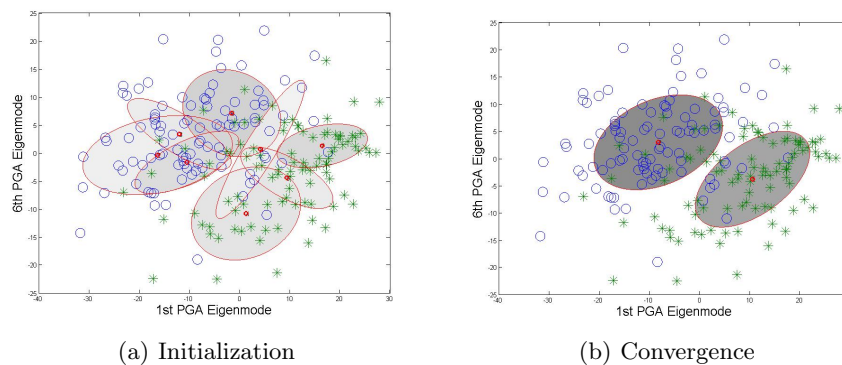


Fig. 4. VBEM of $K = 8$ learning process visualized on *1st* and *6th* PGA feature components, in which the ellipses denote the one standard-deviation density contours for each components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The initial number of component is $K = 8$, during the learning process, some components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted

References

1. Balagani, K., Phoha, V., Iyengar, S., Balakrishnan, N.: On Guo and Nixon's Criterion for Feature Subset Selection: Assumptions, Implications, and Alternative Options. *IEEE TSMC-A: Systems and Humans* 40(3), 651–655 (2010)
2. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks* 5(4), 537–550 (2002)
3. Bishop, C.: *Pattern Recognition and Machine Learning*, vol. 4. Springer New York (2006)
4. Cheng, H., Qin, Z., Qian, W., Liu, W.: Conditional Mutual Information Based Feature Selection. In: *IEEE International Symposium on Knowledge Acquisition and Modeling*. pp. 103–107 (2008)
5. Cover, T.: The Best Two Independent Measurements Are Not The Two Best. *IEEE TSMC* 4(1), 116–117 (2010)
6. Cover, T., Thomas, J., Wiley, J.: *Elements of Information Theory*, vol. 1. Wiley-Interscience (1991)
7. Gurban, M., Thiran, J.: Information Theoretic Feature Extraction for Audio-visual Speech Recognition. *IEEE Transactions on Signal Processing* 57(12), 4765–4776 (2009)
8. Kwak, N., Choi, C.: Input Feature Selection by Mutual Information Based on Parzen Window. *IEEE TPAMI* 24(12), 1667–1671 (2002)
9. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE TPAMI* 27(8), 1226–1238 (2005)
10. Shannon, C.: A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1), 3–55 (2001)
11. Yang, H., Moody, J.: Feature Selection Based on Joint Mutual Information. In: *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*. pp. 22–25 (1999)