Classification

Vladimir Curic
Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
Outline

- An overview on classification
- Basics of classification
- How to choose appropriate features (feature set)
- How to perform classification
- Classification methods
Where are we right now?

**Figure:** Classification in image processing.
What is classification?

- Distinguish different types of objects
- We would like to create an intelligent system that can draw conclusions from our image data
- Determine which parts of the image belong to the object of interest
- No classification (pattern recognition) is possible without some kind of knowledge
- Often used as last stage in automated image analysis task
Classification process

**Figure:** Classification process.
Some important concepts

- Descriptors are called features (feature vectors) $\mathbf{x} = (x_1, \ldots, x_d)$
- Classification is grouping of patterns (samples) into classes
- Patterns are placed in classes of objects that share common properties
Classification example 1

Figure: Classification of leaves.
Classification example 2

Figure: Histological bone implant images.
Features

- Features are the individual measurable heuristic properties of the phenomena
- Discriminating (effective) features
- Independent features
- Features:
  - area, perimeter
  - texture
  - color
  - ...

What are good features?

- Each pattern is represented in terms of $d$ features $\mathbf{x} = (x_1, \ldots, x_d)$
- The goal is to choose those features that allow pattern vectors belonging to different classes to occupy compact and disjoint regions
- It is application dependent
- You might try many, many features, until you find the right ones
- Often, people compute 100s of features, and put them all in a classifier
  - "The classifier will figure out which ones are good"
  - This is wrong!!!
Peaking phenomenon (Curse of dimensionality)

- Additional features may actually degrade the performance of a classifier.
- This paradox is called peaking phenomenon (curse of dimensionality).

**Figure:** Curse of dimensionality.
Feature set 1

- Colour?
- Area?
- Perimeter?
- ...

[Images of various shapes and silhouettes]
Feature set 2

- Colour?
- Area?
- Perimeter?
- ...

[Image of coins]

Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
Feature set 3

- Colour?
- Area?
- Perimeter?
- ...

Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
Dimensionality reduction

- Keep the number of features as small as possible
  - Measurements cost
  - Accuracy
- Simplify pattern representation
- The resulting classifier will be faster and will use less memory
- On the other hand, a reduction in the number of features may lead to a loss in the discriminatory power and thereby lower the accuracy
Feature extraction

- Feature extraction is the process of generating features to be used in the selection and classification tasks.
- Reduce dimensionality by (linear or non-linear) projection of $D$-dimensional vector onto $d$-dimensional vector ($d < D$).
- May have not clear physical meaning.

**Figure:** Feature extraction.
PCA?

Figure: Feature extraction.
Independent Component Analysis (ICA)
Feature selection

- Feature selection methods choose features from the original set based on some criteria
- Reduce dimensionality by selecting subset of original features
- Have clear physical meaning

Figure: Feature selection.
Choice of a criterion function

- The main issues in dimensionality reduction is the choice of a criterion function
- A most commonly used is the classification error of a feature subset
- Example: A measure of distance between two distributions $f_1$ and $f_2$
- Mahalanobis distance: Main assumption is that Gaussian distributions have equal covariance matrix $\Sigma$

$$D_M(f_1, f_2) = (m_1 - m_2)^T \Sigma^{-1} (m_1 - m_2)$$

$m_1$ is a mean of objects in class 1
$m_2$ is a mean of objects in class 2
Feature selection methods

- Exhaustive search
- Best individual features
- Sequential forward selection
- Sequential backward selection
Exhaustive search

- Evaluate all \( \binom{d}{m} \) possible subsets
- Guaranteed to find the optimal subset
- Very expensive!
- Example: If \( d = 50 \), \( m = 5 \) then there are 2,118,760 possible subset (classifications) to evaluate
Best individual features

- Evaluate all $m$ features individually
- Select the best $l \leq m$ individual features
- Simple and not likely to lead to an optimal subset
- Example:
  - Feature 1 is best
  - Feature 2 is best
  - Maybe features 3 and 4 outperform features 1 and 2!
Sequential forward selection

- Selects the best feature and then add one feature at a time
- Once a feature is retained, it cannot be discarded
- Computationally fast (for a subset of size 2 examine $d - 1$ possible subsets)
Sequential backward selection

- Starts with all $d$ features and successive delete one feature at a time
- Once a feature is deleted, it cannot be brought back into the optimal subset
- Require more computation time than sequential forward selection
Search algorithms

- The presented algorithms are suboptimal
- When should we use forward selection?
- When should we use backward selection?
- Is there any optimal algorithms?
  - Branch and bound (divide and conquer method)
  - Genetic algorithms
  - Simulated annealing
Supervised vs. Unsupervised classification

- **Supervised**
  - First apply knowledge, then classify

- **Unsupervised**
  - First classify, then apply knowledge

**Figure:** Supervised and unsupervised classification.
Supervised classification

- Training set - finds the appropriate features for representing the patterns and classifier is trained to partition the feature space.
- Classification mode - the trained classifier assigns the input pattern to one of the pattern classes under consideration based on measured features.
- A given pattern \( \mathbf{x} \) is to be assigned to one of the \( K \) classes \( C_1, C_2, \ldots, C_k \) based on the vector of \( d \) feature values \( \mathbf{x} = (x_1, \ldots, x_d) \).
- Decision rules (Bayes decision, maximum-likelihood) to decision boundary.
Decision boundary

Figure: Decision boundary.
Linear and quadratic classifier

Figure: Linear (left) and quadratic (right) classifier.
Training data

- Training data can be obtained from available training samples
- Classify patterns which are not used during the training stage
- The performance of classifier depends on the number of available training samples as well as the specific values of the samples
- Number of training samples should not be too small!
Training data

- Conditional densities are unknown and they have to be learned from available training data
- Maybe density is known (for example multivariate Gaussian)
- A common strategy is to replace unknown parameters by their estimated values
- Less available information - the difficulty of classification problem increases
How to choose appropriate training set?

Figure: Original image and training image.
How to choose appropriate training set?

Figure: Original image and training image.
Once a feature selection finds a proper representation, a classifier can be designed using a number of possible approaches.

The performance of classifier depends on the interrelationship between sample size, number of features and classifier complexity.

The choice of a classifier is a difficult problem!!!

It is often based on which classifier(s) happen to be be available or best known to the user.
Most commonly used classifiers

- The user can modify several associated parameters and criterion function
- There exist some classification problem for which they are the best choice
- Classifiers:
  - Nearest neighbour classifier
  - Support vector machine
  - Linear and quadratic discriminant analysis
Nearest neighbour

- Stores all training samples
- Assigns pattern to majority class among $k$ nearest neighbour
\(k\)-nearest neighbour

- Assigns pattern to majority class among \(k\) nearest neighbour
$k$—nearest neighbour

- Metric dependent
- Might be slow
- The function is approximated only locally
- The best choice of $k$ depends on data
- Sensitive to outliers
- Larger values of $k$ reduce effects of noise, but make boundaries between the classes less distinct
Support vector machine

- Linear classifier
- Based only on samples close to boundary
Support vector machine

- It is primarily a two class classifier.
- Maximize the width of the margin (empty area) between the classes by selecting a minimum number of support vectors.
Support vector machine

- Support vectors define the classification function (maximizing the margin ⇒ the number of support vector is minimized)
- Metric dependent
Discriminant analysis

- Classes are normally distributed
- Covariance matrices $\Sigma_1, \Sigma_2$ for classes $C_1$ and $C_2$, respectively
- Mean values $m_1, m_2$ for classes $C_1$ and $C_2$, respectively
- Prediction that point $x$ belong to the second class $C_2$

$$\begin{align*}
(x - m_1)^T \Sigma_1^{-1} (x - m_1) + \ln |\Sigma_1| - (x - m_2)^T \Sigma_2^{-1} (x - m_2) - \ln |\Sigma_2| < T
\end{align*}$$

- Threshold $T$?
Discriminant analysis

- Simple and fast
- Linear Discriminant Analysis (LDA)
  - Covariance matrices of the considered classes are equal, i.e., $\Sigma_1 = \Sigma_2$
- Quadratic Discriminant Analysis (QDA)
  - Covariance matrices of the considered classes are not equal, i.e., $\Sigma_1 \neq \Sigma_2$
Linear and Quadratic discriminant analysis

- LDA can only separate the classes with straight lines (in two dimensions) or, more generally, (hyper)-planes in higher dimensions.
- QDA is more sophisticated since it can separate classes with circles, lines, ellipses, parabolas or hyperbolas.

**Figure:** LDA (left) and QDA (right).
Combine classifiers!

- Each of the classifiers can be developed in a different context and for entirely different representation/description (identification of persons by their voice, face as well as hand writing)
- More than a single training set is available
Combine classifiers!

- Different classifiers trained on the same data may only differ in their global performances
- May show strong local differences
- Combination is very useful if the individual classifiers are independent
Combination

- Parallel (ensemble)
  - All classifiers are evaluated independently
  - Their results are combined (first weight then combine!!!)

- Serial - linear sequence
  - The number of possible classes for a given pattern are gradually reduced as more classifiers in the sequence has be involved
  - First consider simple classifiers (low computational and measurements cost)
  - More accurate and expensive classifiers apply after

- Hierarchical (tree-like)
  - Individual classifiers are combined into a structure (a tree-like structure)
  - Advantage high efficiently and flexibility
Unsupervised classification

- Difficult, expensive or even impossible to rely on label training sample with its true category (It is not possible to obtain ground truth)
- Patterns within a cluster are more similar to each other than are patterns belonging to different clusters
Unsupervised classification

- Difficult
- How to determine the number of clusters $K$?
- The number of clusters often depends on resolution (fine vs. coarse)
**$K$—means**

- Usually, a number of clusters is predetermined
- Often starts at random guess
- Tries to minimize some type of error criterion $C = \{C_1, ..., C_k\}$ classes
  
  $d$ distance measure

  $$
  \arg \min_C \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j - m_i)
  $$

- Stops when some stopping criteria is satisfied (the cluster are stabilized)
$K$—means

- **Step 1.** Select an initial partition with $K$ clusters. Repeat steps 2 through 4 until the cluster membership stabilizes
- **Step 2.** Generate a new partition by assigning each pattern to its closest cluster center
- **Step 3.** Compute new cluster centres as the centroids of the clusters
- **Step 4.** Repeat steps 2 and 3 until an optimum value of the criterion function is found
$K$-means
$k$-means
$K$—means - disadvantages

- Different initialization can result in different final clusters
- Fixed number of clusters can make it difficult to predict what $K$ should be
- It is helpful to rerun the classification using the same as well as different $K$ values, to compare the achieved results
- 10 different initializations for 2D data
- For $N$-dimensional data 10 different initializations is often not enough!
How to determine $K$?

- Davies-Bouldin index (DB index)
- It works for spherical clusters
- Intra-cluster and inter-cluster measure
The following eight points \( A_1(2, 10), A_2(2, 5), A_3(8, 4), A_4(5, 8), A_5(7, 5), A_6(6, 4), A_7(1, 2), A_8(4, 9) \) should be classified into three clusters using \( K \)-means clustering. Initial cluster centres are: \( A_1(2, 10), A_4(5, 8) \) and \( A_7(1, 2) \).

Find the three cluster centres after the first iteration. The distance function between two points \( A(x_a, y_a) \) and \( B = (x_b, y_b) \) is defined as \( d(A, B) = |x_a - x_b| + |y_a - y_b| \).
Fuzzy $K$–means (fuzzy $C$–means)

- In the hard clustering process, each data sample is assigned to only one cluster.
- In practice, there are many cases in which the clusters are not completely disjointed.
- Membership values are assigned for all points for each of the clusters.
Summary and conclusions

- Classification is needed in image processing
- It is highly application dependent
- Features and classifiers
- Using more features does not guarantee a better result!