Lecture Notes

Sunday, September 20, 2015 6:58 PM

1.1 Pattern recognition and data mining

The variety of techniques used for pattern recognition are often refer to as *Data mining*. These techniques are drawn primarily from related fields of statistics and machine learning. *Data mining* is a term usually applied to techniques that can be used to find underlying structures and relationships in large amounts of data.

Data mining and pattern recognition

Terms "data mining" and "pattern recognition" are interchangeable, as both focus on the extraction of information or relationships from data.

1.2 Classification vs Estimation

Pattern recognition tasks assume highly varied forms. It is convenient to divide pattern recognition task into two categories:

Classification which involves associating an observation with one of several labels called *classes*; classification problems in biometrics entail decision-making in the face of uncertainty;

Estimation which entails generating an approximation of some desired numerical value based on an observation; estimates often are prone to uncertainty; the uncertainty in estimation can affect the chance an estimate is exactly correct.

Some formulations of classification and estimation problems are shown in the example below.

10/27/2015 8:50 PM - Screen Clipping

Statistical hypothesis

Many problems in biometric system design require that we decide whether to accept or reject a statement about some parameters. The statement is called a *hypothesis* and the decision-making procedure about the hypothesis is called *hypothesis testing*.

Statistical hypothesis

A ***statistical hypothesis*** is an assertion or conjecture concerning one or more populations. The truth or false of statistical hypothesis is never known with absolute certainty, unless we examine the entire population. This is impractical. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis (leads to rejection of the hypothesis). The decision procedure must be done with the awareness of the *probability of the wrong conclusion*. ***The rejection of a hypothesis implies that the sample evidence refutes it.*** In other words: ***The rejection means that there is a small probability of obtaining the sample information observed when, in fact, the hypothesis is true.***

The structure of hypothesis testing is formulated using the term *null hypothesis*. This refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an *alternative hypothesis*, denoted by H_1.

The structure of hypothesis testing is formulated using the term *null hypothesis*. This refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an *alternative hypothesis*, denoted by H_1.

Null and alternative hypothesis

The alternative hypothesis H_1 represents the question to be answered; its specification is crucial. The null hypothesis H_0 nullifies or opposes H_1 and is often the logical compliment to H_1. This results in one of the two following conclusions:

`Reject` H_0 : `In favor of` H_1 `because of sufficient evidence in the data`

`Fail to reject` H_0 : `because of insufficient evidence in the data`

10/27/2015 8:54 PM - Screen Clipping

$$\begin{aligned} \texttt{Null hypothesis } H_0 &: \mu = 50 \\ \texttt{One-sided alternative hypothesis } H_1 &: \mu < 50 \text{ or} \\ \texttt{One-sided alternative hypothesis } H_1 &: \mu > 50 \end{aligned}$$

Testing a statistical hypothesis

Let the null hypothesis be that the mean is $\mu = a$, and the alternative hypothesis be that $\mu \neq a$. That is, we wish to test:

$$\begin{aligned} \texttt{Null hypothesis } H_0 &: \mu = a \\ \texttt{Two-sided alternative hypothesis } H_1 &: \mu \neq a \end{aligned}$$

Suppose that a data sample of n is tested, and that the *sample mean* $\overline{x}$ is observed. The sample mean is an estimate of the true population mean $\mu = a$. A value of the sample mean $\overline{x}$, that falls close to the hypothesized value of μ, is the evidence that the true mean μ is really a; that is, such evidence supports the null hypothesis H_0. On the other hand, a sample mean $\overline{x}$ that is considerably different from a, is evidence in support of the alternative hypothesis H_1. Thus, the sample mean represents the test statistics.

10/27/2015 8:56 PM - Screen Clipping

Example 3: (Critical region and values.) The sample mean $\overline{x}$ can take on many different values. Suppose that if $48.5 \leq \overline{x} \leq 51.5$, we will not reject the null hypothesis $H_0 : \mu = 50$. If either $\overline{x} < 48.5$ or $\overline{x} > 51.5$, we will reject the null hypothesis in favor of the alternative hypothesis $H_1 : \mu \neq 50$. The values of $\overline{x}$ that are less than 48.5 and greater than 51.5 constitute the **critical region** for the test. The boundaries that define the critical regions (48.5 and 51.5) are called **critical values**.

10/27/2015 8:56 PM - Screen Clipping

Therefore, we reject H_0 in favor of H_1 if the test statistic falls in the critical region, and fails to reject H_0 otherwise. This decision procedure can lead to either of the two wrong conclusions:

Type I error or False reject rate (FRR): is defined as rejecting the null hypothesis H_0 when it is true. The type I error is also called the **significant level** of the test. The probability of making a type I error is

$$\alpha = P(\texttt{Type I error}) = P(\texttt{Reject } H_0 \texttt{ when } H_0 \texttt{ is true})$$

Type II error or False accept rate (FAR): is defined as failing to reject the null hypothesis when it is false. The probability of making a type II error is

$$\beta = P(\texttt{Type II error}) = P(\texttt{Fail to reject } H_0 \texttt{ when } H_0 \texttt{ is false})$$

Properties of type I (FRR) and type II (FAR) errors

Property 1: Type I error and type II error are related. A decrease in the probability of one generally results in an increase in the probability of the other

Property 2: The size of the critical region, and, therefore, the probability of committing a type I error, can always be reduced by adjusting the critical value(s)

Property 3: An increase in the sample size n will reduce α and β simultaneously

Property 4: If H_0 is false, β is maximum when the true value of a parameter approaches the hypothesized value. The greater the distance between the true value and the hypothesized value, the smaller β will be.

10/27/2015 8:56 PM - Screen Clipping

Recommendations for computing type I and II errors

Type I error. Generally, the designer controls the type I error probability α, called a **significance level**, when the critical values (the boundaries that define the critical region, see Example 3) are selected. Thus, it is usually easy for the designer to set the type I error probability at (or near) any desired value. Because the designer can directly control the probability of wrongly rejecting H_0, we always think of rejection of the null hypothesis H_0 as a **strong conclusion**.

Because we can control the probability of making a type I error, α, the problem is what value should be used. The type I error probability is a measure of risk, specifically, the risk of concluding the the null hypothesis is false when it really isn't. So, the value of α should be chosen to reflect the consequences (biometric data, device, system, etc.) of incorrectly rejecting H_0:

▶ **Smaller** values of α would reflect more serious consequences, and

▶ **Larger** values of α would be consistent with less severe consequences.

This is often hard to do, and what has evolved in much of biometric system design is to use the value $\alpha = 0.05$ in most situations, unless there is information available that indicates that this is an inappropriate choice.

Type II error. The probability of type II error, β, is **not a constant**. It depends on both the true value of the parameter and the sample size that we have selected. Because the type II error probability β is a **function** of both, the sample size and extent to which the null hypothesis H_0 is false, it is customary to think of the decision not to reject H_0 as a **weak conclusion**, unless we know that β is acceptably small. Therefore, rather than saying we "accept H_0" we prefer the terminology "**fail to reject H_0**".

> Failing to reject H_0 implies that we have not found sufficient evidence to reject H_0, that is, to make a strong statement. Failing to reject H_0 does not necessarily mean there is a high probability that H_0 is true. It may simply mean that more data are required to reach a strong conclusion. This can have important implications for the formulation of hypotheses.

The power of a statistical test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true. The power is computed as

$$\text{Power of a statistical test} = 1 - \beta$$

The power of a statistical test can be interpreted as the probability of correctly rejecting a false null hypothesis. The power of the test is very descriptive and concise measure of the sensitivity of a statistical test, where by sensitivity we mean the ability of the test to detect differences.

Example 4: (Type I and II errors.) The techniques for computing type I and II errors for given data sample is shown in Fig. 1.

10/27/2015 8:58 PM - Screen Clipping

Design example: Computing type I and II errors

Problem formulation:

Let be face features such as the regions of lips, mouth. nose, ears, eys, eyebrow, and other facial measureing be detected. Let biometric data corresponding to the lip topology is represented by a sample of size $n = 10$, while the mean and the standard deviation are $\mu = 50$ and $\sigma = 2.5$, respectively. This biometric data has a distribution for which the conditions of the central limit theorem apply, so the distribution of the sample mean is approximately normal with mean $\mu = 50$ and $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$. Find the probability of type I error.

Step 1: The probability of type I error

$\mu = 50$ and $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$. Find the probability of type I error.

Step 1: The probability of type I error

The probability of making type I error (or significance level of our test)

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$

is equal to the sum of the areas that have been shaded in the tails of the normal distribution. We may find this probability as

$$\text{Probability of type I error, } \alpha = \overbrace{P(\overline{X} < \underbrace{48.5}_{\overline{x}_1} \text{ when } \mu = 50)}^{\text{Left tail}} + \underbrace{P(\overline{X} > \overbrace{51.5}^{\overline{x}_2} \text{ when } \mu = 50)}_{\text{Right tail}}$$

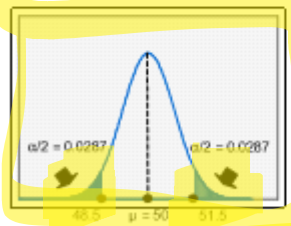
The z-values that correspond to the critical values 48.5 and 51.5 are calculated as follows:

$$z_1 = \frac{\overline{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{48.5 - 50}{0.79} = \boxed{-1.90} \quad \text{and} \quad z_2 = \frac{\overline{x}_2 - \mu}{\sigma/\sqrt{n}} = \frac{51.5 - 50}{0.79} = \boxed{1.90}$$

Therefore $\alpha = P(Z < -1.90) + P(Z > 1.90) = P(Z < -1.90) + (1 - P(Z < 1.90)) = 0.0287 + (1 - 0.9713) = \boxed{0.0574}$

Conclusion: This implies that 5.74% of all random samples would lead to rejection of the hypothesis $H_0: \mu = 50$ when the true mean is really 50.

Step 2: Reducing a type I error by decreasing the critical region



From inspection of the critical region for $H_0: \mu = 50$ versus $H_1: \mu \neq 50$ and $n = 10$, note that we can reduce α by pushing the critical regions further into the tails of the distribution. For example, if we make the critical values 48 and 52, the values of α is

$$\alpha = P(Z_2 < \frac{48 - 50}{0.79} + P(Z_2 > \frac{52 - 50}{0.79})$$

$$= P(Z < -2.53) + P(Z > 2.53) = 0.0057 + 0.0057 = \boxed{0.0114}$$

Fig. 1: Techniques for computing type I and II errors (Example 4).

10/27/2015 8:59 PM - Screen Clipping

Design example: Computing type I and II errors (Continuation)

Step 3: Reducing type I error by increasing the sample size

We could also reduce α by increasing the sample size, assuming that the critical values of 48.5 and 51.5 do not change. If $n = 16$, $\frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{16}} = 0.625$, and using the original critical region, we find

$$z_1 = \frac{48.5 - 50}{0.625} = \boxed{-2.40} \quad \text{and} \quad z_2 = \frac{51.5 - 50}{0.625} = \boxed{2.40}$$

Therefore $\alpha = P(Z < -2.40) + P(Z > 2.40) = 0.0082 + 0.0082 = \boxed{0.0164}$

Step 4: Design decision on type I error

An acceptable type I error can be chosen from the following possibilities:

Type I error from the original critical region $Z < -1.90, Z > 1.90$ is $\boxed{\alpha=0.0574}$

An acceptable type I error can be chosen from the following possibilities:

Type I error from the original critical region $Z < -1.90, Z > 1.90$ is $\boxed{\alpha=0.0574}$

The type I error reduced by decreasing the critical region from $Z < -1.90, Z > 1.90$ to $Z < -2.53, Z > 2.53$ is $\boxed{\alpha=0.0114}$

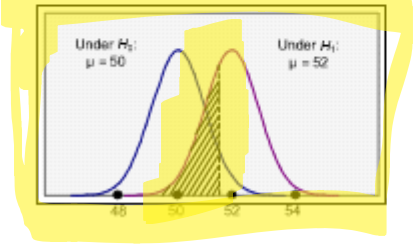
Type I error reduced by increasing the sample size from $n = 10$ to $n = 16$ is $\boxed{\alpha = 0.0164}$

Step 5: Specification of the probability of type II error

The probability of making type II error is

$$\beta = P(\text{Type II error}) = P(\text{Fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

To calculate β, we must have a specific alternative hypothesis; that is, we must have a particular value of μ. For example, suppose we want to reject the null hypothesis $H_0: \mu = 50$ whenever the mean μ is grater than 52 or less than 48. We could calculate the probability of type II error β for the values $\mu = 52$ and $\mu = 48$, and use this result to tell us something about how the test procedure would perform. Because of the symmetry of normal distribution function, it is only necessary to evaluate one of the two cases, say, find the probability of not rejecting the null hypothesis $H_0: \mu = 50$ when the true mean is $\mu = 52$.



- The normal distribution on the left (see figure on the left) is the distribution of the test statistic $\overline{X}$ when the null hypothesis $H_0: \mu = 50$ is true (this is what is meant by the expression "under $H_0: \mu = 50$")
- The normal distribution on the right is the distribution of the test statistic $\overline{X}$ when the alternative hypothesis is true and the value of the mean is 52 (or "under $H_1: \mu = 52$").

Design example: Computing type I and II errors (Continuation)

Step 5: (continuation)

Now the type II error will be committed if the sample mean $\overline{x}$ falls between 48.5 and 51.5 (the critical region boundaries) when $\mu = 52$. This is the probability that $48.5 \leq \overline{X} \leq 51.5$ when the true mean is $\mu = 52$, or the shaded area under the normal distribution on the right, that is

$$\beta = P(\text{Type II error}) = P(48.5 \leq \overline{X} \leq 51.5 \text{ when } \mu = 52)$$

Step 6: Computing of the probability of type II error

The z-values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{0.79} = \boxed{-4.43} \text{ and } z_2 = \frac{51.5 - 52}{0.79} = \boxed{-0.63}$$

Therefore,

$$\begin{aligned} \text{Probability of type II error, } \beta &= P(-4.43 \leq Z \leq -0.63) \\ &= P(Z \leq -0.63) - P(Z \leq -4.43) \\ &= 0.2643 - 0.0000 = \boxed{0.2643} \end{aligned}$$

<u>Conclusion</u>: If we are testing $H_0: \mu = 50$ against $H_1: \mu \neq 50$ with $n = 10$, and the true value of the mean is $\mu = 52$, the probability that we will fail to reject the false null hypothesis is $\boxed{0.2643}$. By symmetry (see graphical representation in Fig. 2), if the truth value of the mean

Conclusion: If we are testing $H_0: \mu = 50$ against $H_1: \mu \neq 50$ with $n = 10$, and the true value of the mean is $\mu = 52$, the probability that we will fail to reject the false null hypothesis is $\boxed{0.2643}$. By symmetry (see graphical representation in Fig. 2), if the truth value of the mean is $\mu = 48$, the value of β will also be $\boxed{0.2643}$.

Step 7: Analysis of a type II error

The probability of making a type II error β increases rapidly as the true value μ approaches the hypothesized value.

For example, consider the case when the true value of the mean is $\mu = 50.5$ and the hypothesized value is $H_0: \mu = 50$. The true value of μ is very close to 50, and the value for probability of type II error β is $\beta = P(48.5 \leq \overline{X} \leq 51.5)$ when $\mu = 50.5$.

The z-values corresponding to 48.5 and 51.5 when $\mu = 50.5$ are

$$z_1 = \frac{48.5 - 50.5}{0.79} = \boxed{-2.53} \text{ and } z_2 = \frac{51.5 - 50.5}{0.79} = \boxed{1.27}$$

Therefore $\beta = P(-2.53 \leq Z \leq 1.27) = P(Z \leq 1.27) - P(Z \leq -2.53) = 0.8980 - 0.0057 = \boxed{0.8923}$. This is higher than in case $\mu = 52$, that is, we are more likely to accept the faulty hypothesis $\mu = 50$ (fail to reject $H_0: \mu = 50$).

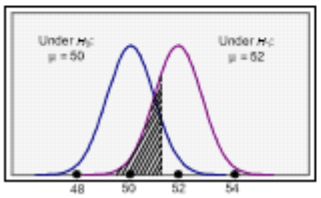
10/27/2015 8:59 PM - Screen Clipping

Design example: Computing type I and II errors (Continuation)

Step 7: (Continuation)

Conclusion: The type II error probability is much higher for the case in which the true mean is 50.5 than for the case in which the mean is 52. Of course, in many practical situations, we would not be as concerned with making a type II error if the mean were "close" to the hypothesized value. We would be much more interested in identifying the large differences between the true mean and the value specified in the null hypothesis.

Step 8: Reducing a type II error by increasing the sample size



The type II error probability also depends on the sample size n. Suppose that the null hypothesis is $H_0: \mu = 50$ and that the true value of the mean is $\mu = 52$. By letting the sample size increase from $n = 10$ to $n = 16$, we can compare them using the graphics on the left. The normal distribution on the left is the distribution of $\overline{X}$ when $\mu = 50$, and the normal distribution on the right is the distribution of $\overline{X}$ when $\mu = 52$. As shown in figure, the type II error probability is

$$\text{Probability of type II error } \beta = P(48.5 \leq \overline{X} \leq 51.5) \text{ when } \mu = 52$$

When $n = 16$, the standard deviation of $\overline{X}$ is $\delta/\sqrt{n} = 2.5/\sqrt{16} = 0.625$, and the z-values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{0.625} = \boxed{-5.60} \text{ and } z_2 = \frac{51.5 - 52}{0.625} = \boxed{-0.80}$$

Therefore

$$\begin{aligned} \text{Probability of type II error } \beta &= P(-5.60 \leq Z \leq -0.80) \\ &= P(Z \leq -0.80) - P(Z \leq -5.60) \\ &= 0.2119 - 0.0000 = \boxed{0.2119} \end{aligned}$$

This $\beta = 0.2119$ is smaller than $\beta = 0.2642$, so we decrease the probability of accepting the false hypothesis H_0 by increasing the sample size.

$= 0.2119 - 0.0000 = \boxed{0.2119}$

This $\beta = 0.2119$ is smaller than $\beta = 0.2642$, so we decrease the probability of accepting the false hypothesis H_0 by increasing the sample size.

Step 9: Design decision on type II error

An acceptable type II error can be chosen from the following possibilities:

The type II error for the original sample size $n = 10$ and $-4.43 \leq Z \leq -0.63$ is $\boxed{0.2643}$

The type II error, reduced by increasing the sample size from $n = 10$ to $n = 16$, is $\boxed{0.2119}$

10/27/2015 8:59 PM - Screen Clipping

Design example: Computing type I and II errors (Continuation)

Step 10: Computing the power of a test

Suppose that the true value of the mean is $\mu = 52$. When $n = 10$, we found that $\beta = 0.2643$, so the power of this test is

$$\texttt{Power of the test} = 1 - \beta = 1 - 0.2643 = \boxed{0.7357}$$

Conclusion: The sensitivity of the test for detecting the difference between a mean of 50 and 52 is $\boxed{0.7357}$. That is, if the true mean is really 52, this test will correctly reject $H_0: \mu = 50$ and "detect" this difference 73.57% of the time. If this value of power is judged to be too low, the designer can increase either α or the sample size n.

10/27/2015 9:00 PM - Screen Clipping

False match rate (FMR) is the expected probability that a sample will be falsely declared to match a single randomly-selected *non-self* template; that is, measurements from two different persons are interpreted as if they were from the same person.

False non-match rate (FNMR) is the expected probability that a sample will be falsely declared not to match a template of the same measure from the same user supplying the sample; that is, measurements from the same person are treated as if they were from two different persons.

Equal error rate (EER) is the value defined as EER=FMR=FNMR, that is, the point where false match and false non-match curves cross is called equal error rate or crossover rate. The EER provides an indicator of the system's performance: a lower EER indicates a system with good level of sensitivity and performance.

Difference between false match/non-match rates and false accept/reject rates is introduced in Fig. 8.

Example 10: (FMR and FNMR.) Let us assume that a certain commercial biometric verification system wishes to operate at 0.001% FMR. At this setting, several biometric systems, such as the state-of-the-art fingerprint and iris recognition systems, can deliver low than 1% FNMR. A FMR of 0.001% indicates that

commercial biometric verification system wishes to operate at 0.001% FMR. At this setting, several biometric systems, such as the state-of-the-art fingerprint and iris recognition systems, can deliver less than 1% FNMR. A FMR of 0.001% indicates that, if a hacker launches a brute force attack with a large number of different fingerprints, 1 out of 100 000 attempts will succeed on an average.

2.5 FAR computing

The FAR (type II error) is defined as the probability that a user making a false claim about his/her identity will be verified as that false identity. That is, FAR is the expected proportion of transactions with wrongful claims of identity (in a positive ID system) or non-identity (in a negative ID system) that are incorrectly confirmed. A transaction may consist of one or more wrongful attempts, depending upon the decision policy. Note that *acceptance* always refers to the claim of the user[3].

Example 9: (False accept.) If a person A_1 types the user ID of another person A_2 into the biometric login for the given terminal, this means that A_1 has just made a false claim that he or she is A_2. The person A_1 presents his biometric measurement for verification. If the biometric system matches A_1 to A_2, then there is a **false acceptance**. This could happen because the matching threshold is set too high, or it could be that biometric features of A_1 and A_2 are very similar.

Suppose the person A_1 was n times successfully authenticated as A_2 in the total number of attempts, N, then $\mathtt{FAR} = n/N$. The FRR is the mean (average) for K users of a system:

$$\mathtt{FAR} = \frac{1}{K}\sum_{i=1}^{K}\mathtt{FAR}_i$$

FAR and matching algorithm

The FAR characterizes the strength of the matching algorithm. The stronger the algorithm, the less likely that a false authentication will happen.

2.6 Matching errors

Matching algorithm errors, occurred while performing a single comparison of a submitted sample against a single enrolled template/model, are defined to avoid ambiguity within the system allowing multiple attempts or having multiple templates.

[3] It should be noted that conflicting definitions are implicit in literature. In access control literature, a false acceptance is said to have occurred when a submitted sample is incorrectly matched to a template enrolled by another user.

2.3 Decision error rates

Decision errors are due to matching errors and image acquisition errors. These errors are summed up and drive the decision process at various levels of the system, in particular, in situation when (a) one-to-one or one-to-many matching is required; (b) there is a positive or negative claim of identity; and (c) the system allows multiple attempts (the decision policy). Biometric performance has traditionally been stated in terms of the decision error rates.

2.4 FRR computing

The FRR (type I error) is defined as the probability that the user making a true claim about his/her identity will be rejected as him/herself. That is, the FRR is the expected proportion of transactions with truthful claims of identity (in a positive ID system) or non-identity (in a negative ID system) that are incorrectly denied. A transaction may consist of one or more truthful attempts dependent upon the decision policy. Note that *rejection* always refers to the claim of the user

> Example 8: (False reject.) If person A_1 types his/her correct user ID into the biometric login for the given terminal, this means that A_1 has just made a true claim that he/she is A_1. Person A_1 presents his/her biometric measurement for verification. If the biometric system does not match the template of A_1 to the A_1's presented measurement, then there is a **false reject**. This could happen because the matching threshold is too low, or the biometric features presented by a person A_1 are not close enough to the biometric template.

Suppose a person A_1 was denied his authentication (unsuccessfully authenticated) as A_1 n times, while the total number of attempts was N, then $\mathrm{FRR} = n/N$. Statistically, the more times something is done, the greater is the confidence in the result. The result is the mean (average) FRR for K users of the system:

$$\mathrm{FRR} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{FRR}_i$$

FRR and matching algorithm

The strength of the FRR is the **robustness** of the algorithm. The more accurate the matching algorithm, the less likely a false rejection will happen.

10/27/2015 9:19 PM - Screen Clipping

2.2 Decision rule

If the stored biometric template of a user I is represented by X_I and the acquired input for recognition is represented by X_Q, then the null H_0 and alternate H_1 hypotheses are:

If the stored biometric template of a user I is represented by X_I and the acquired input for recognition is represented by X_Q, then the null H_0 and alternate H_1 hypotheses are:

Null hypothesis H_0: Input X_Q does not come from the same person as the template X_I; the associated decision is: "Person I is not who he/she claims to be";

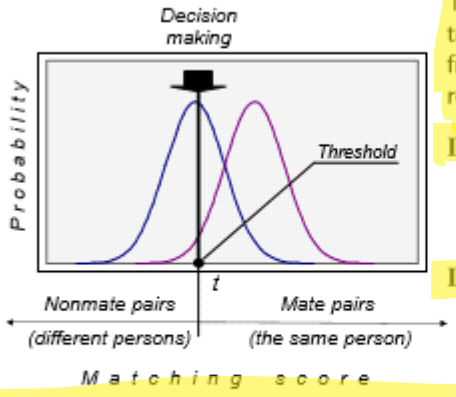
Alternate hypotheses H_1: Input X_Q comes from the same person as the template X_I; the associated decision is: "Person I is who he/she claims to be"

That is, we wish to test

$$\texttt{Null hypothesis } H_0: \; D = D_0$$
$$\texttt{Alternative hypothesis } H_1: \; D \neq D_0$$

The decision rule is as follows: if the matching score $S(X_Q, X_I)$ is less than the system threshold t, then decide H_0, else decide H_1.

Controlled decision making in a biometric system



The higher the score, the more certain the system is that the two biometric measurements come from the same person. The system decision is regulated by the **threshold** t:

Decision 1: Pairs of biometric samples generating scores *higher than* or *equal to* t are inferred as **mate pairs**, that is, the pairs belong to the same person.

Decision 2: Pairs of biometric samples generating scores *lower than* t are inferred as **nonmate pairs**, that is, the pairs belong to different persons.

10/27/2015 9:19 PM - Screen Clipping

Basic definitions and terminology

Sample: A biometric measure submitted by the user.

Template: A user's reference measure based on features extracted from the enrolment samples.

Matching score: A measure of the similarity between features derived from a presented sample and a stored template. A match/nonmatch decision may be made according to whether this score exceeds a decision threshold.

System decision: A determination of the probable validity of a users claim to identity/non-identity in the system.

Transaction: An attempt by a user to validate a claim of identity or non-identity by consecutively submitting one or more samples, as allowed by the system's decision policy.

Verification: The user makes a *positive* claim to an identity, requiring a one-to-one comparison of the submitted sample to the enrolled template for the claimed identity.

Identification: The user makes either no claim or an implicit negative claim to an enrolled identity, and a one-to-many search of the

template for the claimed identity.

Identification: The user makes either no claim or an implicit negative claim to an enrolled identity, and a one-to-many search of the entire enrolled database is required.

Positive claim of identity: The user claims to be enrolled in or known to the system. An explicit claim might be accompanied by a claimed identity in the form of a name, or personal identification number (PIN). Common access control systems are an example.

Negative claim of identity: The user claims not to be known to or enrolled in the system. For example, enrolment in social service systems open only to those not already enrolled.

Genuine claim of identity: A user making a truthful positive claim about identity in the system. The user truthfully claims to be him/herself, leading to a comparison of a sample with a truly matching template.

Impostor claim of identity: A user making a false positive claim about identity in the system. The user falsely claims to be someone else, leading to the comparison of a sample with a non-matching template.

10/27/2015 9:19 PM - Screen Clipping

2 Biometric system performance evaluation

Fig. 7 contains the basic definitions and terminology used in the design and testing of biometric systems. In thus design, the terms such as a sample of biometric data, user template, matching score, decision-making, decision rule and decision error rates are used in specific-application meaning.

2.1 Matching score

A broad category of variables, impacting the way in which the users inherent biometric characteristics, are displayed to the sensor. In many cases, the distinction between changes in the fundamental biometric characteristics and the presentation effects may not be clear.

Two samples of the same biometric characteristic from the same person are not identical due to imperfect imaging conditions, changes in the users physiological or behavioral characteristics, ambient conditions, and user's interaction with the sensor. Therefore, the response of a biometric matching system is the matching score $S(X_Q, X_I)$

$$\text{Response} = \text{Matching score } S(\underbrace{Input}_{X_Q}, \overbrace{Template}^{X_I})$$

that quantifies the similarity between the input X_Q and the template X_I representations. This similarity can be encoded by a single number.

10/27/2015 9:20 PM - Screen Clipping

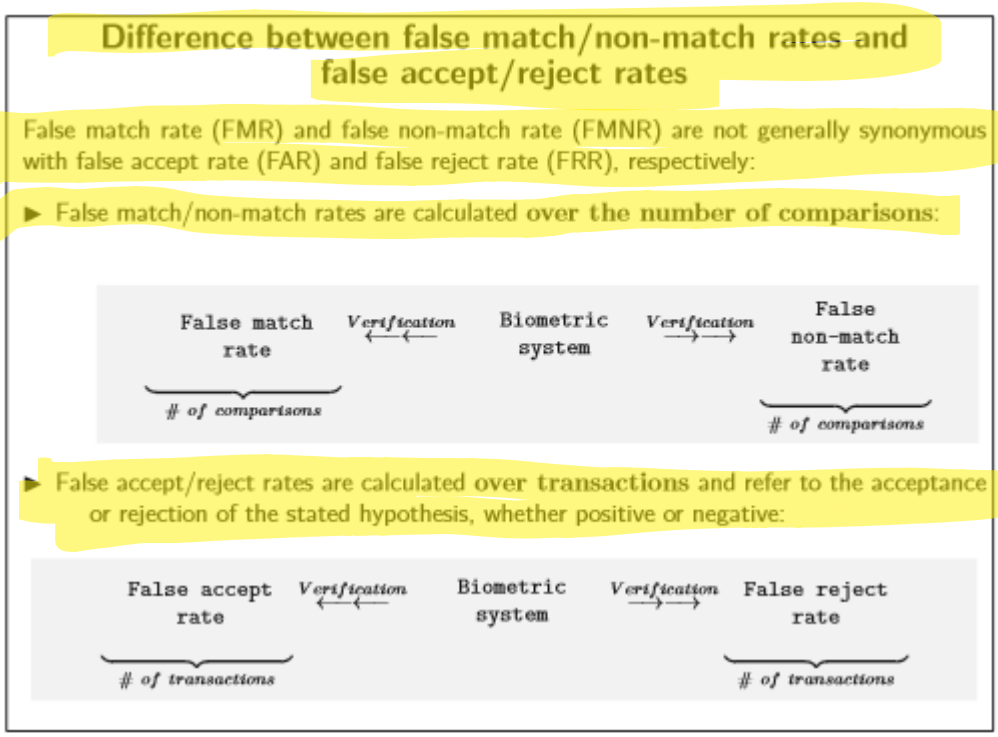
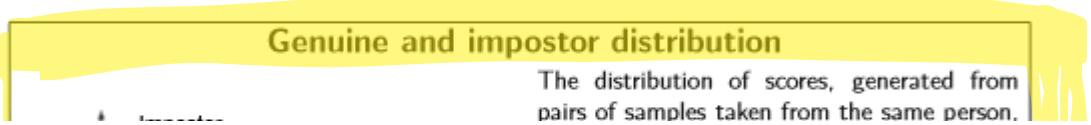


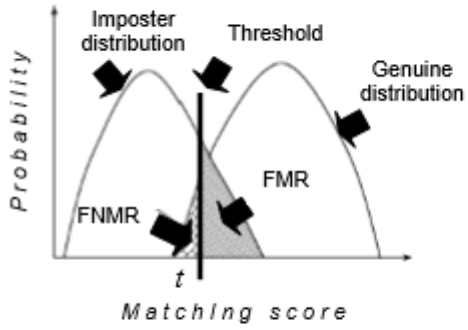
Fig. 8: Difference between false match/non-match rates and false accept/reject rates.

Example 11: (FMR and FNMR.) Consider that airport authorities are looking for the 100 criminals.
(a) Consider a verification system. The state-of-the-art fingerprint verification system operates at 1% FNMR and 0.001% FMR; that is, this system would fail to match the correct users 1% of the time and erroneously verify wrong users 0.001% of the time.
(b) Consider an identification system. Assume that the identification FMR is still be 1%, the FNMR is 0.1%. That is, while the system has a 99% chance of catching a criminal, it will produce large number of false alarms. For example, if 10,000 people may use a airport in a day, the system will produce 10 false alarms.

The EER (equal error rate) is defined as the crossover point on a graph that has both the FAR and FRR curves plotted.



Genuine and impostor distribution



The distribution of scores, generated from pairs of samples taken from the same person, is called the **genuine distribution**. The distribution of scores while the samples are taken from different persons, is called the **impostor distribution**.

FMR and FNMR for a given threshold t are displayed over the genuine and impostor score distributions; FMR is the percentage of non-mate pairs whose matching scores are greater than or equal to t, and FNMR is the percentage of mate pairs whose matching scores are less than t.

The FMR (FAR) and FNMR (FRR) are related and must be balanced (Figure 9). For example, in access control, perfect security would require denying access to everyone. Conversely, granting access to everyone would mean no security. Obviously, neither extreme is reasonable, and biometric system must operate somewhere between the two.

10/27/2015 9:20 PM - Screen Clipping

2.7 FTE computing

The FTE (failure to enroll) is defined as the probability that a user, attempting to biometrically enroll, will be unable to do so. The FTE is usually defined by a minimum of three attempts. The FTE can be calculated as follows. Let unsuccessful enrollment event occurs if a person A_1, on his third attempt, is still unsuccessful. Let n be the number of unsuccessful enrollment events, and N be the total number of enrollment events. Then $\text{FTE} = n/N$. The mean (average) FTE for K users of a system is

$$\text{FTE} = \frac{1}{K}\sum_{i=1}^{K}\text{FTE}_i$$

10/27/2015 9:21 PM - Screen Clipping

False match rate (FMR) or False accept rate (FAR)

- A FMR and FAR occurs when a system incorrectly matches an identity; A FMR (FAR) is the probability of individuals being wrongly matched.
- False matches may occur because there is a high degree of

False non-match rate (FNMR) or False reject rate (FRR)

- A FNMR and FRR occurs when a system rejects a valid identity; A FNMR (FRR) is the probability of valid individuals being wrongly not matched.
- False non-matches occur because there is not a sufficiently strong similarity between individuals' enrollment and trial templates, which could be caused by any

- False matches may occur because there is a high degree of similarity between two individuals' characteristics.
- In a verification and positive identification system, unauthorized people can be granted access to facilities or resources as a result of an incorrect match.
- In a negative identification system, the result of a false match may be to deny access.

similarity between individuals' enrollment and trial templates, which could be caused by any number of conditions. For example, an individual's biometric data may have changed as a result of aging or injury.

- In verification and positive identification system, people can be denied access to some facility or resource as a result of a system's failure to make a correct match.
- In a negative identification system, the result of a false non-match may be that a person is granted access to resources to which he/she should be denied.

Balance of FMR (FAR) and FNMR (FRR)

FMR (FAR) and FNMR (FRR) are related and must, therefore, always be assessed in tandem, and acceptable risk levels must be balanced with the disadvantages of inconvenience.

10/27/2015 9:21 PM - Screen Clipping

3 Receiver operating characteristic (ROC) curves

The standard method for expressing the technical performance of a biometric device for a specific population in a specific application is the **Receiver Operating Characteristic** (ROC) curve.

3.1 Applications of biometric system in terms of a ROC

The system performance at all the operating points (thresholds) can be depicted in the form given in Fig. 10.

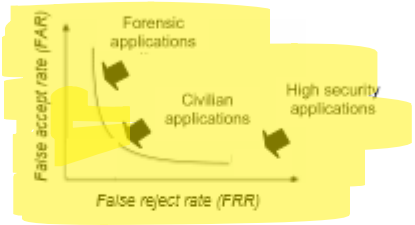


Fig. 10: Typical operating points of different biometric applications are displayed on the ROC curve.

An ROC curve plots, parametrically as a function of the decision threshold $t = T$, the rate of "false positives" (i.e. impostor attempts accepted) is shown on the X-axis, against the corresponding rate of "true positives" (i.e. genuine attempts accepted) on the Y-axis.

against the corresponding rate of "true positives" (i.e. genuine attempts accepted) on the Y-axis.

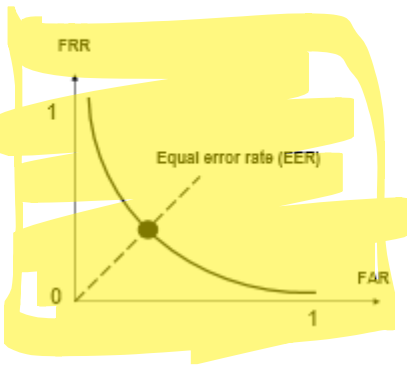
3.2 Equal error rate (EER) in terms of ROC

Graphical interpretation of the EER is given in Figure 11. The FMR, FNMR, and EER behavior is expressed in terms of a ROC. The FMR and FNMR can be considered as the functions of the threshold $t = T$. These functions give the error rates when the match decision is made at some threshold T.

3.3 Comparison the performance of biometric systems

ROC curves allow to compare the performance of different systems under similar conditions, or of a single system under differing conditions.

10/27/2015 9:21 PM - Screen Clipping



- *When the threshold T is set low, the FMR is high and the FNMR is low; when T is set high, the FMR is low and the FNMR is high.*
- *For a given matcher, operating point (a point on the ROC) is often given by specifying the threshold T.*
- *In biometric system design, when specifying an application, or a performance target, or when comparing two matches, the operating point is specified by choosing FMR or FNMR.*
- *The equal error operating point is defined as the EER. The matcher can operates with highly unequal FMR and FNMR; in this case, the EER is unreliable summary of system accuracy.*

Fig. 11: The relationship between FRR, FAR, and EER.

Example 12: (Comparison two matchers.) Various approaches can be used in matcher design. The matches must be compared using criteria of operational accuracy (method and algorithm) and operational time (computing platform). In Figure 12, the technique for comparison two matches is introduced using criterion of operational accuracy.

10/27/2015 9:22 PM - Screen Clipping

3.4 Confidence intervals for the ROC

Each point on the ROC curve is calculated by integrating "genuine" and "impostor" score distributions between zero and some threshold, $t = T$. Confidence intervals for

Each point on the ROC curve is calculated by integrating "genuine" and "impostor" score distributions between zero and some threshold, $t = T$. Confidence intervals for the ROC at each threshold, t, have been found through a summation of the binomial distribution under the assumption that each comparison represents a Bernoulli trial[4]

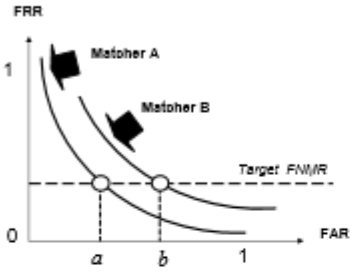
The confidence, β, given a non-varying probability p, of k sample/template comparison scores, or fewer, out of n independent comparison scores being in the region of

[4] An experiment can be represented by n repeated Bernoulli trials, each with two outcomes that can be labeled **success**, with a probability p, or **failure**, with probability $1 - p$. The probability distribution of the binomial random variable X, that is, the number of successes in n independent trials, is $b(x; n, p) = \left(\frac{n!}{(n-x)!}\right) p^x q^{n-x}$, $x = 0, 1, \ldots n$. For example, for $n = 3$ and $p = 0.25$, the probability distribution of X can be calculated as $b(x; 3, 0.25) = \left(\frac{3!}{(3-x)!}\right)(0.25)^x(0.75)^{3-x}$, $x = 0, 1, \ldots 3$.

10/27/2015 9:22 PM - Screen Clipping

Design example: Comparison two matchers using the ROC curves

Problem formulation:



In biometric system design, two types of matches are specified, type A and type B matcher. These matches are described by the ROC curves. Figure in the left shows the corresponding ROCs for these matches and their operating points for some specified target FNMR. The problem is to choose the better matcher.

Step 1: Understanding of the initial data

The ROCs of two matches are plotted in the form acceptable for comparison (the same type of ROC and scaling factors). The ROC shows the trade-off between FMR and FNMR with respect to the threshold T. For a given operational matcher, the operational point is specified by the particular threshold T.

Step 2: Comparison of two matchers

It follows from the ROC characteristics of the matchers, that:

- For every specified FMR it has a lower FNMR;
- For every specified FNMR it has a lower FMR;

Conclusion

The matcher A is better than matcher A for all possible thresholds T.

10/27/2015 9:22 PM - Screen Clipping

$$\text{Confidence intervals } 1 - \beta = P(i \leq k) = \sum_{i=0}^{k} b(i; n, p) \qquad (4)$$

Confidence intervals $1 - \beta = P(i \leq k) = \underbrace{\sum_{i=0}^{k} b(i; n, p)}_{\text{Available from Table}}$ (4)

e binomial sums are available and given table for different values of n and p is given.

10/27/2015 9:23 PM - Screen Clipping

The required number of comparison scores (and test subjects) cannot be predicted prior to testing. To deal with this, *Doddingtons Law* is to test until 30 errors have been observed.

Example 14: (Doddingtons law.) If the test is large enough to produce 30 errors, we will be about 95% sure that the *true* value of the error rate for this test lies within about 40% of that measured.

If the test is large enough to produce 30 errors, we will be about 95% sure that the *true* value of the error rate for this test lies within about 40% of that measured, provided that Equation 4 is applicable. The comparison of biometric measures will not be Bernoulli trials and Equation 4 will not be applicable if:

10/27/2015 9:24 PM - Screen Clipping

sured.

If the test is large enough to produce 30 errors, we will be about 95% sure that the *true* value of the error rate for this test lies within about 40% of that measured, provided that Equation 4 is applicable. The comparison of biometric measures will not be Bernoulli trials and Equation 4 will not be applicable if:

(*a*) Trials are not independent, and
(*b*) The error probability varies across the population.

Example 15: (Equation 4 is not applicable.) Trials will not be independent if users stop after a successful use and continue after a non-successful use.

10/27/2015 9:24 PM - Screen Clipping

3.6 Test size

The size of an evaluation, in terms of the number of volunteers and the number of attempts made (and, if applicable, the number of fingers/hands/eyes used per person) will affect how accurately we can measure error rates. The larger the test, the more accurate the results are likely to be.

Rules such as the *Rule of 3* and *Rule of 30*, detailed below, give lower bounds to the number of attempts needed for a given level of accuracy. However, these rules are

accurate the results are likely to be.

Rules such as the *Rule of 3* and *Rule of 30*, detailed below, give lower bounds to the number of attempts needed for a given level of accuracy. However, these rules are overoptimistic, as they assume that error rates are due to a single source of variability, which is not generally the case with biometrics. Ten enrolment-test sample pairs from each of a hundred people is not statistically equivalent to a single enrolment-test sample pair from each of a thousand people, and will not deliver the same level of certainty in the results.

The Rule of 3 addresses the question: *What is the lowest error rate that can be statistically established with a given number N of (independent identically distributed) comparisons?* This value is the error rate p for which the probability of errors in N trials is zero, purely by chance. It can be, for example, 5%.

The Rule of 3

$$\texttt{Error rate } p \approx \frac{3}{N} \texttt{ for a 95\% confidence level} \quad (5)$$

$$\texttt{Error rate } p \approx \frac{2}{N} \texttt{ for a 90\% confidence level} \quad (6)$$

Example 17: (Rule of 3.) A test of 300 independent samples can be said with 95% confidence to have an error rate of $\frac{3}{300} = \boxed{1\%}$ or less.

10/27/2015 9:25 PM - Screen Clipping

The "Rule of 30" Doddington[5] proposes the *Rule of 30* for helping determine the test size: ***To be 90% confident that the true error rate is within ±30% of the observed value, we need at least 30 errors.***

The rule below generalizes different proportional error bands:

The Rule of 30

To be 90% confident that the true error rate is within

±10% of the observed value, we need at least 260 errors
±30% of the observed value, we need at least 30 errors
±50% of the observed value, we need at least 11 errors

Example 18: (Rule of 30.) If we have 30 false non-match errors in 3,000 independent genuine trials, we can say with 90% confidence that the true error rate is $\frac{30}{3000} = 0.01 \pm 30\%$, that is,

$$1\% - 0.3 \leq \texttt{ True error rate } \leq 1\% + 0.3$$

$$\boxed{0.7\%} \leq \texttt{ True error rate } \leq \boxed{1.3\%}$$

10/27/2015 9:25 PM - Screen Clipping

10/27/2015 9:25 PM - Screen Clipping

1.1 Basic definitions

An image is defined as a two-dimensional (2D) function $f(x, y)$, where x and y are spatial coordinates, and the amplitude of f at any pair of coordinates (x, y) is the *intensity* of the image at that point. An image can be *continuous* or *discrete* with respect to the coordinates and amplitude. An image can be converted to digital form using digitized coordinates and amplitude. Digitizing the coordinate values is called *sampling*. Digitizing the amplitude values is called *quantization*. *Digital image* is defined when x, y, and the amplitude values of f are discrete quantities. The result of sampling and quantization is a matrix of real numbers. Color images are formed by a combination of individual 2D images.

> Example 1: (Color images.) In the RGB color system, a color image consists of three (red, green, and blue) individual component images.

The field of *digital image processing* refers to processing digital images using digital devices and systems. Digital image is composed of a finite number of elements called *pixels*. A digital images can be considered in any band of electromagnetic spectrum, ranging from gamma to radio waves. That is, images can be generated by sources that humans are not accustomed to associating with images (humans are limited to the visual band).

10/27/2015 9:27 PM - Screen Clipping

1.2 Histograms

A *histogram* is a compact summary of a 2D image. To construct a histogram for continuous data, we must divide the range of the data into intervals of equal width. Some judgments must be used in selecting the number of intervals so that a reasonable display can be developed. A histogram that uses either too few or too many intervals will not be informative. Once the number of intervals and the lower and upper boundary of each interval have been determined, the data are sorted into the intervals, and count is made of the number of observations in each interval. To construct the histogram, use the horizontal axis to represent the measurement scale for the data and the vertical scale to represent the counts, or *frequencies*.

10/27/2015 9:28 PM - Screen Clipping

1.3 2D discrete Fourier transform

1.3 2D discrete Fourier transform

Let $f(x, y)$ for $x = 0, 1, 2, \ldots, M-1$ and $y = 0, 1, 2, \ldots, N-1$, denote an $M \times N$ image.

Direct discrete Fourier transform. The 2D discrete Fourier transform (DFT) of f, denoted by $F(u, v)$, is given by the equation

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi \frac{u \times x}{M} + \frac{v \times y}{N}}$$

for $u = 0, 1, 2, \ldots, M-1$ and $v = 0, 1, 2, \ldots, N-1$.

10/27/2015 9:28 PM - Screen Clipping

The *frequency domain* is the coordinate system spanned by $F(u, v)$ with frequency variables u and v. This is analogous to the *spatial domain*, which coordinate system is spanned by $f(x, y)$ with spatial variables x and y.

Inverse discrete Fourier transform. Given DFT $F(u, v)$ of the image $f(x, y)$, find the image $f(x, y)$ can be restored using the inverse transform. The inverse DFT is given by the equation

$$f(x,y) = \frac{1}{M \times N} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u,v) e^{j2\pi \frac{u \times x}{M} + \frac{v \times y}{N}}$$

for $x = 0, 1, 2, \ldots, M-1$ and $y = 0, 1, 2, \ldots, N-1$.

10/27/2015 9:29 PM - Screen Clipping

1.4 Properties of DFT

If $f(x, y)$ is real, its Fourier transform is conjugate symmetric about the origin, that is,

$$F(u,v) = F^*(-u,-v)$$

This implies that the Fourier spectrum also is symmetric about the origin, $|F(u, v)| = |F^*(-u, -v)|$.

The DFT is infinitely periodic in both the u and v directions, with periodicity determined by M and N:

$$F(u,v) = \underbrace{F(u+M,v)}_{\text{Periodic in } u} = \underbrace{F(u,v+N)}_{\text{Periodic in } v} = \underbrace{F(u+M,v+N)}_{\text{Periodic in } u \text{ and } v}$$

Periodicity is also a property of the inverse DFT:

$$f(x,y) = \underbrace{f(x+M,y)}_{\text{Periodic in } x} = \underbrace{f(x,y+N)}_{\text{Periodic in } y} = \underbrace{f(x+M,y+N)}_{\text{Periodic in } x \text{ and } y}$$

Example 5: (Computing the 2D DFT.) The DFT of an $M \times N$ image array `f` and its inverse can be computed using the fast Fourier transform MATLAB function `fft2`, that is, `F=fft2(f)`. This function returns a result represented by the matrix of size $M \times N$. The Fourier spectrum can be computed using `abc` function: `S=abc(F)`, which compute the magnitude of each element of the array.

1.5 Filtering in frequency domain

The basic for linear filtering in both the spatial and frequency domains is the *convolution theorem* as follows. Convolution of two spatial functions $f(x,y)$ and $h(h,y)$ can be computed by the multiplication of the Fourier transforms of these functions, $H(u,v)$ and $F(u,v)$, respectively:

$$\underbrace{f(x,y) * h(x,y)}_{\text{Convolution}} \Leftrightarrow \underbrace{H(u,v) \times F(u,v)}_{\text{Product}}$$

where symbol " $*$ " denotes convolution of the two functions; symbol " $\Leftrightarrow$ " constitutes the correspondance. Conversely,

$$\underbrace{f(x,y) \times h(x,y)}_{\text{Product}} \Leftrightarrow \underbrace{H(u,v) * F(u,v)}_{\text{Convolution}}$$

In terms of filtering, filtering in the spatial domain consists of convolving an image $f(x,y)$ with a filter mask $h(x,y)$. The same result can be obtained in the frequency domain by multiplying $H(u,v)$ by $F(u,v)$. The Fourier transform $H(u,v)$ is called the *filter transfer function*.

10/27/2015 9:29 PM - Screen Clipping

1.6 Image restoration

The goal of restoration is to improve a given image. Restoration techniques are based on modeling the degradation and applying the inverse process in order to recover the original

The goal of restoration is to improve a given image. Restoration techniques are based on modeling the degradation and applying the inverse process in order to recover the original image. Given the image $f(x, y)$, degraded image $g(x, y)$ can be modeled as follows

$$g(x, y) = \underbrace{H[f(x, y)]}_{Degradation\ function} + \underbrace{\eta(x, y)}_{Noise}$$

where $H[f(x, y)]$ is the degradation function and $\eta(x, y)$ is the additive noise. If H is a liner, spatially invariant process, the degraded image can be described in the ***spatial domain*** as follows

$$g(x, y) = \underbrace{h(x, y) * f(x, y)}_{Degradation\ function} + \underbrace{\eta(x, y)}_{Noise}$$

An equivalent computing in ***frequency domain*** is as follows

$$G(u, v) = \underbrace{H(u, v) * F(u, v)}_{Degradation\ function} + \underbrace{N(u, v)}_{Noise}$$

Note that spatial noise values are random numbers, characterized by a *probability distribution function*. Various distributions are used in noise modeling, for example, the uniform, normal (Gaussian), exponential, and others.

Example 6: (Noise). In MATLAB, the function `g = imnoise(f,type,parameters)` models the corruption of the input image `f` with noise. Using this function, the noise with Gaussian, Poisson, and others distributions can be added. For example, `g = imnoise(f,'gaussian',m,var)` adds Gaussian noise with mean `m` and the variance `var` to image `f`.

The simplest approach to restoring a degraded image is to form an estimate as follows:

$$\hat{F}(u, v) = \frac{G(u, v)}{H(u, v)} = F(u, v) + \frac{N(u, v)}{H(u, v)}$$

Taking the inverse Fourier transform of $\hat{F}(u, v)$, the corresponding estimate of the image can be obtained. It follows from this equation that even if we knew $H(u, v)$ exactly, we could not recover $F(u, v)$ because $N(u, v)$ (the noise component) is not known. That is why another approaches, such as Winer filtering, are needed for restoring images. More on these and other image processing techniques can be found in the books [3, 5].

could not recover $I(a, b)$ because $N(a, b)$ (the noise component) is not known. That is why another approaches, such as Winer filtering, are needed for restoring images. More on these and other image processing techniques can be found in the books [3, 5].

10/27/2015 9:30 PM - Screen Clipping

1 Pattern recognition

Pattern recognition is the association of an observation to past experience or knowledge. Humans continuously perform perceptual pattern recognition – from understanding spoken languages, to recognizing faces of friends and foes, or to distinguish between odors and perfumes. However, much remains a mystery in the task of recognizing spoken languages or faces; and we still do not know much about how to describe a formal method for the task we perform so effortlessly.

1.1 Pattern recognition and data mining

The variety of techniques used for pattern recognition are often refer to as *Data mining*. These techniques are drawn primarily from related fields of statistics and machine learning. *Data mining* is a term usually applied to techniques that can be used to find underlying structures and relationships in large amounts of data.

Data mining and pattern recognition

Terms "data mining" and "pattern recognition" are interchangeable, as both focus on the extraction of information or relationships from data.

1.2 Classification vs Estimation

Pattern recognition tasks assume highly varied forms. It is convenient to divide pattern recognition task into two categories:

Classification which involves associating an observation with one of several labels called *classes*; classification problems in biometrics entail decision-making in the face of uncertainty;

Estimation which entails generating an approximation of some desired numerical value based on an observation; estimates often are prone to uncertainty; the uncertainty in estimation can affect the chance an estimate is exactly correct.

Some formulations of classification and estimation problems are shown in the example below.

10/27/2015 9:31 PM - Screen Clipping

There are two potential sources of errors in estimation:

Modeling error, due to differences between the model used and an optimal model, and

Uncertainty, due to the nature of biometric data; uncertainty can be classified into

Modeling error, due to differences between the model used and an optimal model, and

Uncertainty, due to the nature of biometric data; uncertainty can be classified into two types: uncertainty due to missing variables (problem features) and inherently random noise.

10/27/2015 9:31 PM - Screen Clipping

1.3 Pattern recognition as a process

In biometric systems, the pattern recognition regards image or other biometric signal (such as audio record) recognition. For these objects, pattern recognition can be viewed as the process of assigning a *label* to an *observation*. In biometric systems, this is implemented in the pattern recognition module (Fig. 1).

Pattern recognition module of a biometric system

Input: The input of a pattern recognition module is an encoded observation.
Output: A model may estimate the values of one or more variables given an input vector. The output is the label assigned to the observation.
Mapping: An estimation model is defined by a *mapping* from the input space to the output space. A mapping is a function assigning one output vector to each possible input vector.

Pattern recognition requires the tools providing detailed instructions, that is, implemented mathematical equations characterizing the relationship between inputs and desired outputs of a pattern recognition module. Formulating these equations, or building a *model*, is the central problem in the pattern recognition task in biometric systems. The development process for building a model is referred to as *modeling*.

10/27/2015 9:31 PM - Screen Clipping

Preprocessing

The removal of irrelevant information and extraction of key features to simplify a pattern recognition problem is called *preprocessing*. The original encoding of the observation used by the preprocessor is referred to as a *raw input vector*. Preprocessing aims at simplifying the relationships to be inferred by a model. There are various types of preprocessing, in particular:

- Reducing the number of input variables (the size of the input space).
- Normalizing each of the input variables.
- Smoothing the relationships by transforming a problem so that the relationships of the resulting problem are simpler.

10/27/2015 9:31 PM - Screen Clipping

1.3.2 Feature extraction

The first stage in any task of pattern recognition of any objects, including biometric ones (fingerprints, hand geometry, iris, etc.), is usually referred to as *feature extraction*. Feature extraction is nothing more than a process of measurement, but this process can often be so complicated that it constitutes the main work of a pattern recognizer. The result of feature extraction stage is a set of numbers that are fed to the classifier (the classification, or decision, stage of the recognizer) (Fig. 2).

Example 3: (Feature extraction.) Fingerprint template (feature vector) can range in size from 200 to over 1,000 bytes. Facial templates vary in size from less than 100 bytes to over 3K. Signature templates can range in size from 1K to 3K. The largest templates are associated with behavioral biometrics.

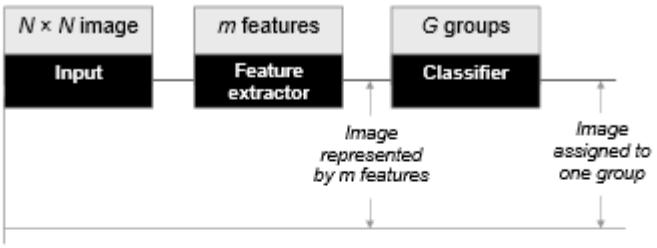


Fig. 2: Image feature extraction and classification.

A *local feature* is a subset of pixels at a particular location within an image which form a recognizable object in their own right.

Detecting a local feature within an image may be sufficient in itself to classify the entire image, or it may have to be used in combination with other features as part of the classification rule.

Example 5: (Local feature.) The features of the signature can be detected by an algorithm for feature extraction (Figure 3).

1.3.3 Feature reduction and Data normalization

There are many ways to normalize or scale data. In particular, Min-max normalization performs a linear transformation of the original input range into a newly specified data range. The minimum from the original image, *Min1*, is mapped to the new minimum.

There are many ways to normalize or scale data. In particular, Min-max normalization performs a linear transformation of the original input range into a newly specified data range. The minimum from the original image, $Min1$, is mapped to the new minimum, $Min2$. Similarly, the original maximum, $Max1$, is mapped to the new maximum, $Max2$. All points are linearly mapped to the new scale using the following formula:

$$\texttt{New value } y' = \frac{\texttt{Original value } y - Min1}{Max1 - Min2}(Max2 - Min2) + Min2 \tag{1}$$

10/27/2015 9:33 PM - Screen Clipping

One of the advantages of min-max normalization is that it ***preserves all relationships*** of the data values exactly. It does not introduce any potential biases into the data (the shape of the histogram is preserved).

In cases where the actual minimums and maximums of the input variables are not known, another normalization should be used. In particular, the input variable data (original image y) can be translated into the mean value,y', so that ***the mean is zero and the variance is one***. This is done by the following transformation:

$$\texttt{New value } y' = \frac{\texttt{Original value } y - \texttt{Mean } \mu}{\texttt{Standard deviation } \sigma} \tag{2}$$

where `Mean` is the population mean (μ) and `Standard deviation` is a standard deviation of the population (σ).

Example 8: (Normalization.) Equation 2 can be used for normalization of a normal random variables. Let X be a normal random variable with mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. Then the random variable $Z = \frac{X-\mu}{\sigma}$ is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$; that is, Z is a **standard** normal random variable.

10/27/2015 9:34 PM - Screen Clipping

1.3.4 Classification

The results of the feature extraction, the features of the objects, are further used for the object classification. The chosen model for classification is called a *classifier*, and is characterized as follows:

The classifier

- The classifier is aimed at dividing a pool of object into classes while minimizing the classification error.
- The input to the classifier algorithm is the **features** of the objects.
- The feature vectors are called **patterns**.

classification error.

- The input to the classifier algorithm is the **features** of the objects.
- The feature vectors are called **patterns**.
- The feature space is divided by a classifier into regions that correspond to **classes**.
- The patterns (feature vectors) whose true classes are known and which are used for the design of the classifier, represent the **training patterns**.
- A classifier can be designed based on a set of training data being available (a priori known information). This classifier is said to have a **supervised** pattern recognition.
- If training data is not available, a set of feature vectors is used to unravel the underlying similarities, and perform grouping of similar vectors together. This is known as **unsupervised** pattern recognition. Such a grouping is also called **clustering**.

10/27/2015 9:34 PM - Screen Clipping

1.4 Classification models for pattern recognition

There are various strategies and the corresponding models for classification

- Supervised classification, or matching (the training objects are labeled to belong to some classes),
- Unsupervised classification (no labeled objects are available).

The supervised classifiers include:

- Simple correlation techniques,
- Linear discriminant analysis based on discrimination rule,
- Statistical methods based on Bayes rule,

10/27/2015 9:34 PM - Screen Clipping

The unsupervised classifiers include:

- Data clustering based on *Gaussian model*,
- *Gaussian mixture model* with Expectation-maximization (EM) algorithm, or *K-means* algorithm
- Principal Component Analysis (PCA),
- Independent Component Analysis (ICA).

10/27/2015 9:35 PM - Screen Clipping

1.4.1 Similarity and correlation as the basis of template matching

Given a *prototype* of each of the groups into which you are trying to classify an image, a useful feature to use in a classification rule would be any measure of *similarity* between the object and each of the prototypes. This idea of using prototypes for each group is often referred to as *template matching*. A suitable measure of similarity is the *correlation coefficient*:

Correlation coefficient

Let two groups, X_1 and X_2, be given. The correlation between these group is expressed in terms of correlation coefficients:

$$\rho_{X_1X_2} = \frac{E(X_1X_2) - \mu_1\mu_2}{\sqrt{\sigma_1^2\sigma_1^2}}, \tag{3}$$

where $E(X_1, X_2)$ is the expected value (men) of the data from both groups; μ_1 and μ_2 are the means of the two populations, and σ_1^2 and σ_2^2 are their covariances, respectively. The value of the correlation coefficient, $\rho_{X_1X_2}$, varies between 0 and 1, and is 1 when there is a perfect match between two images. That is, when the correlation between two images is equal to 1, the two images are identical apart from brightness and contrast. As the value of the correlation coefficient begins to fall towards 0, the two images become increasingly different.

Thus, the value of the correlation coefficient can be used to classify an image by the simple process of computing the correlation between the image and prototype from each group; the image is then assigned to the group with which it has the largest correlation.

Simple regression model

For a problem with one input variable and one output variable, linear regression can only produce mappings that are represented by the equation

$$y = \beta_0 + \beta_1 x_1 + \text{Error} \tag{7}$$

Given a set of example patterns for an ***estimation problem*** with a single input x and a single output y, linear regression selects parameters so that the output values generated by the model are close to the desired output values provided by the example patterns, that is, the line (Equation 7) is chosen to best *fit* the data.

Simple linear regression model

In the **simple linear regression model**, the dependent variable, or **response**, is related to one independent variable, or **regressor variable**, as follows

$$y = \beta_0 + \beta_1 x_1 + \text{Error}$$

The parameters β_0 and β_1 are called **regression coefficients**.

Suppose that the true relationship between Y and x is represented by a straight line, and that observation Y at each level of x is a random variable. The expected value of Y for each value of x is

$$E(Y|x) = \beta_0 + \beta_1 x_1 \tag{8}$$

where the intercept β_0 and the slope β_1 are unknown regression coefficients. We assume

$$E(Y|x) = \beta_0 + \beta_1 x_1 \tag{8}$$

where the **intercept** β_0 and the **slope** β_1 are unknown regression coefficients. We assume that each observation, Y, can be described by the model (7) where a random error is characterized by the mean zero and variance σ^2.

The **fitted** or **estimated regression line** is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 \tag{9}$$

The **least squares estimates** of the intercept $\widehat{\beta}_0$ and slope $\widehat{\beta}_1$ are well-known in statistics.

10/27/2015 9:37 PM - Screen Clipping

The Bayes rule

Assign an object to the group G_i with the largest conditional probability of membership, $P(G_j|\mathbf{X})$, where $P(G_i|\mathbf{X}) > P(G_j|\mathbf{X})$ for all $i \neq j$.

10/27/2015 9:37 PM - Screen Clipping

The Bayes formula: $P(G_i|\mathrm{X}) = \dfrac{P(\mathbf{X}|G_i)P(G_i)}{\sum_j P(\mathrm{X}|G_i)P(G_i)}$

10/27/2015 9:37 PM - Screen Clipping

This idea of a classification rule dividing up the sample space of possible values of $\mathbf{X}$ into regions belonging to each group, can be extended to more than two groups and more than two measurements.

The Bayes classifier is **optimal** with respect to minimizing the classification error probability. The mean value and the variation (or standard deviation) play the role of **features**. The feature are treated as random variables, and the feature vectors are called **patterns**.

Linear classifiers

The case of ***normally distributed*** groups each produces a particular simple division of the sample space in that the decision surfaces can be quadratic (parabolas, hyperbolas, etc.). In case if the the group are not only normally distributed, bur have the same covariance matrices, the dividing surface between groups can be linear. Linear classifiers are the most common type of classification rule used in pattern recognition.

covariance matrices, the dividing surface between groups can be linear. Linear classifiers are the most common type of classification rule used in pattern recognition.

Example 12: (Classification.) Two well separated groups can be classified using a liner division shown in Fig. 5b.

1.4.4 Gaussian model for unsupervised classification

Classification theory is essentially a statistical theory, for it makes use of conditional probabilities as ways to summarize our knowledge of some phenomenon or object after incorporation of what we know about it. The statistical model that uses the conditional probability to assign an object to a group is known as Gaussian model, and is based on the Bayes rule.

Unimodal Gaussian model

The unimodal Gaussian algorithm is a relatively simple parametric model for pattern classification. The basic for this model relies on the assumption that the probability distribution for input vectors of each class is Gaussian. By using Bayes rule, the probability of a particular class $C - j$ given an input pattern X is

$$P(C_j|X) = P(X|C_j)\frac{P(C_j)}{P(X)} \tag{10}$$

where $P(C_j|X)$ is the probability that an input vector X belongs to class C_j; $P(X|C_j)$ is the probability density function of an input vector X if the class was known be j; it tells the distribution of feature vectors in the feature space inside a particular class, i.e.

10/27/2015 9:38 PM - Screen Clipping

Gaussian mixture model

Similar to the Gaussian model, the **Gaussian mixture** estimates probability density functions for each class, and then performs classification. However, in actual model, the probability density can be multimodal. For example, if we are searching for different face parts on a facial image, and there are several basic types of the eyes (for different races, perhaps), the unimodal Gaussian model would describe a wide mixture of all eye types, including patterns that might not look like an eye at all.

Gaussian model and Gaussian mixture model

Similar to the unimodal Gaussian model, the **Gaussian mixture** estimates probability density functions for each class.

Unlike the unimodal Gaussian model, which assumes $P(X|C_j)$ to be in the form of a Gaussian, the Gaussian mixture model estimates $P(X|C_j)$ as a weighted average of multiple Gaussians.

Gaussian mixture model utilizes a mixture of several Gaussian distributions and can, therefore, represent sub-classes inside one class. Given enough Gaussian components, this model can approximate an arbitrary distribution. Specifically, in the Gaussian mixture model, the estimate of the probability density function for each class j takes on the form of a weighted sum of Gaussians:

model, the estimate of the probability density function for each class j takes on the form of a weighted sum of Gaussians:

$$P(X|C_j) = \sum_{k=1}^{N_c} w_k G_k \tag{11}$$

where w_k is the weight of the k-th Gaussian G_k, and the weights sum to one. Each Gaussian component, G_k, is defined from Gaussian distribution function with mean M_k

(free parameters) and the variance in the form of covariance matrix (free parameters) (details can be found in textbooks on statistics).

Estimation of gaussian mixture (finding the free parameters) for one class can be considered as unsupervised training of the case, where samples are generated by individual components of the mixture distribution and without the knowledge of which sample was generated by which component. The free parameters can be adjusted using a procedure, that aims at maximizing the **log likelihood function**, L, for each class j,

$$L = \prod_{i=0}^{N_{Train}} lnP(X_i|C_j), \tag{12}$$

where N_{train} is the number of datasets available for training. Thus, this model requires some training prior to classification.

Training using the Gaussian mixture model

Step 1: (*a*) Initialize the initial Gaussian means μ_i, $i = 1, 2, \ldots, G$.
(*b*) Initialize the covariance matrices.
(*c*) Initialize the the weights $\pi_i = 1/G$, so that all Gaussians are equally likely.

Step 2: Present each pattern X of the training set and model each of classes K as a weighted sum of Gaussians (Equation 11)

Step 3: Recompute and iteratively update the weights, means, and covariances. Stop training if weights are less than some threshold value. Otherwise, continue the iterative updates.

Classification using the Gaussian mixture model

Step 1: Present each input pattern X and compute the confidence for each class j as follows:

$$\text{Confidence} = P(C_j)P(X|C_j)$$

where $P(C_j) = N_{C_j}/N$ is the prior probability of class C_j estimating by counting the number of training patters.

Step 2: Classify the pattern X as the class with the highest confidence.

Step 2: Classify the pattern X as the class with the highest confidence.

10/27/2015 9:38 PM - Screen Clipping

The purpose of the training is to estimate the unknown parameters of the Gaussian mixture while maximizing the likelihood estimations, which is equivalent to reducing variance to 0. Some optimization algorithm must be employed to perform this task. In the mixed models, an approach called the *Expectation Maximization (EM)* is used. The likelihood function in the EM keeps increasing until a maximum is reached and the algorithm converges.

10/27/2015 9:39 PM - Screen Clipping

1.4.5 Factor Analysis

Factor analysis is a statistical method in examining the data. Factorial designs are used in pattern recognition involving several factors (variables, or inputs), where it is necessary to study the joint effect of the factors on a response (output of a pattern recognition module). The effect of a factor is defined as the change in response produced by a change in the level of the factor. The difference in response between the levels of one factor is not the same at all levels of the other factors. When this occurs, there is an interaction between the factors. Factorial experiments are the only way to discover interactions between variables.

The are two techniques widely used in image recognition:

- Principle component analysis (PCA), which is classified as a supervised recognition model, and
- Independent component analysis (ICA), which is classified as an unsupervised recognition model.

PCA is a data transform technique that decorrelates the input data by exploiting pairwise second-order dependencies (e. g. covariance) and produces linearly independent variables, called *components* or *eigenvectors* [4].

PCA attempts to identify an m-dimensional subspace of the n-dimensional input space ($m < n$) that seems most "significant," and then project the data onto this subspace. In doing so, the number of input variables is reduced from n to m. Hence, the size of the input space may be reduced in a sensible and systematic way by eliminating the latter principle components.

Principle component analysis (PCA)

Input: PCA begins by computing n orthonormal vectors, known as the **principle components**, which provide a basis for the input space.

- Principle components are ordered in a sequence from the most to the least significant. Since these unit vectors each point in a direction perpendicular to the others, they can be viewed as a new set of axes for the input space.
- This new set of axes possesses useful properties that are lacking in the original set of axes (the original input variables). In particular, when the input data is viewed using this new set of axes, ***the variance of the data is maximal along the first axis***, next highest along the second axis, etc.
- All input variables are **normalized** within the same range of data values.

Output: The first principal components, i.e., those along which data exhibits highest variance, are the **most significant**. Only the variables of relatively low variance should be eliminated.

2 Pattern recognition module design

The pattern recognition module is a key component of any biometric system. Design of a pattern recognition module consists of 10 main steps which are introduced in Fig. 6. The design step are explained in details in Fig. 7 and Fig. 8.

Step 1: Defining the pattern recognition problem

At the first step, the pattern recognition problem is specified as a classification or an estimation problem. The following questions might be to considered:

Question 1: What level of accuracy is needed to achieve and what level of accuracy would be considered successful?

Question 2: How to benchmark the performance?

Question 3: What existing alternatives can be compared against?

Question 4: What kind of data will you use to evaluate the various models?

10/27/2015 9:39 PM - Screen Clipping