Dynamic Feature Fusion Trees:
a New, Simple and Flexible Approach to Statistical Pattern Matching

Nigel Sedgwick

Cambridge Algorithmica Limited
9 Oakdene
Beaconsfield
Buckinghamshire
United Kingdom HP9 2BZ

Tel: +44 (0)1494 678989
URL: http://www.camalg.co.uk
Fax: +44 (0)1494 678990
Email: ncs@camalg.co.uk

Overview of Presentation

1. Introduction
2. Pattern Matching is becoming Increasingly Pervasive
3. A Scientific Model of Pattern Matching: Examples (UCI Wine Dataset)
4. How do we Measure Performance: the ROC Curve
5. Definition of Detection Gain; the Maths (Naïve Bayes)
6. Normalising Single Features (building PDFs)
7. Pairwise Feature Fusion
8. Why am I doing this: Some Problems Seen
9. Dynamic Feature Fusion Trees (features raw and fused)
10. Dealing with Noise and Missing Data
11. Parallels with Neural Networks (?and the Brain)
12. Discussion and Conclusions
13. Thoughts on Future Work

Pattern Matching is Pervasive

Simple old things: Oil Warning Light in Car

Newer things: Car Number-Plate Recognition

...

Biometrics: are you who you claim to be?

Automatic Speech Recognition

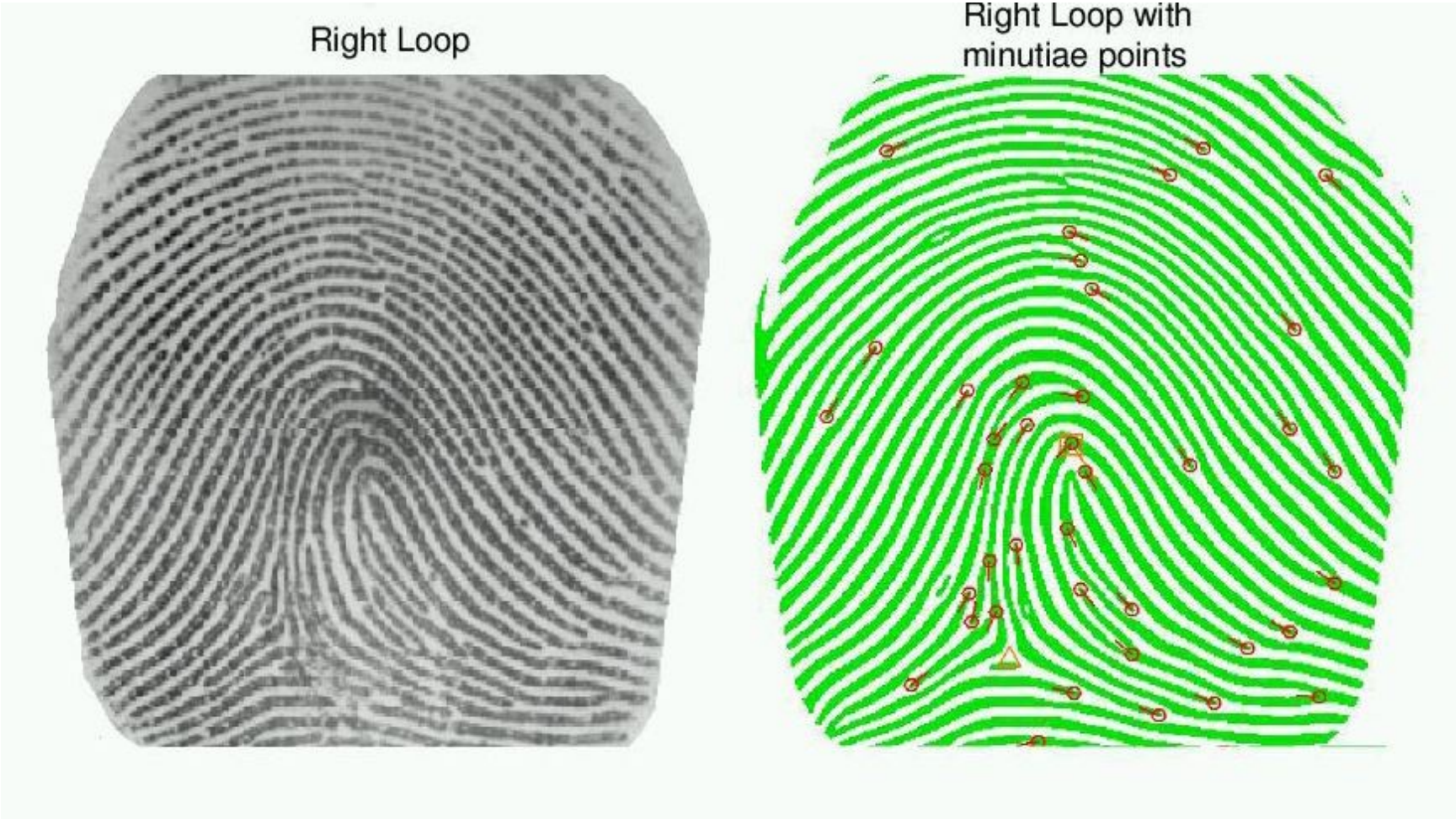
Equipment Condition Monitoring

Military Target Detection and Tracking (eg Radar, Sonar, IR, Visible)

...

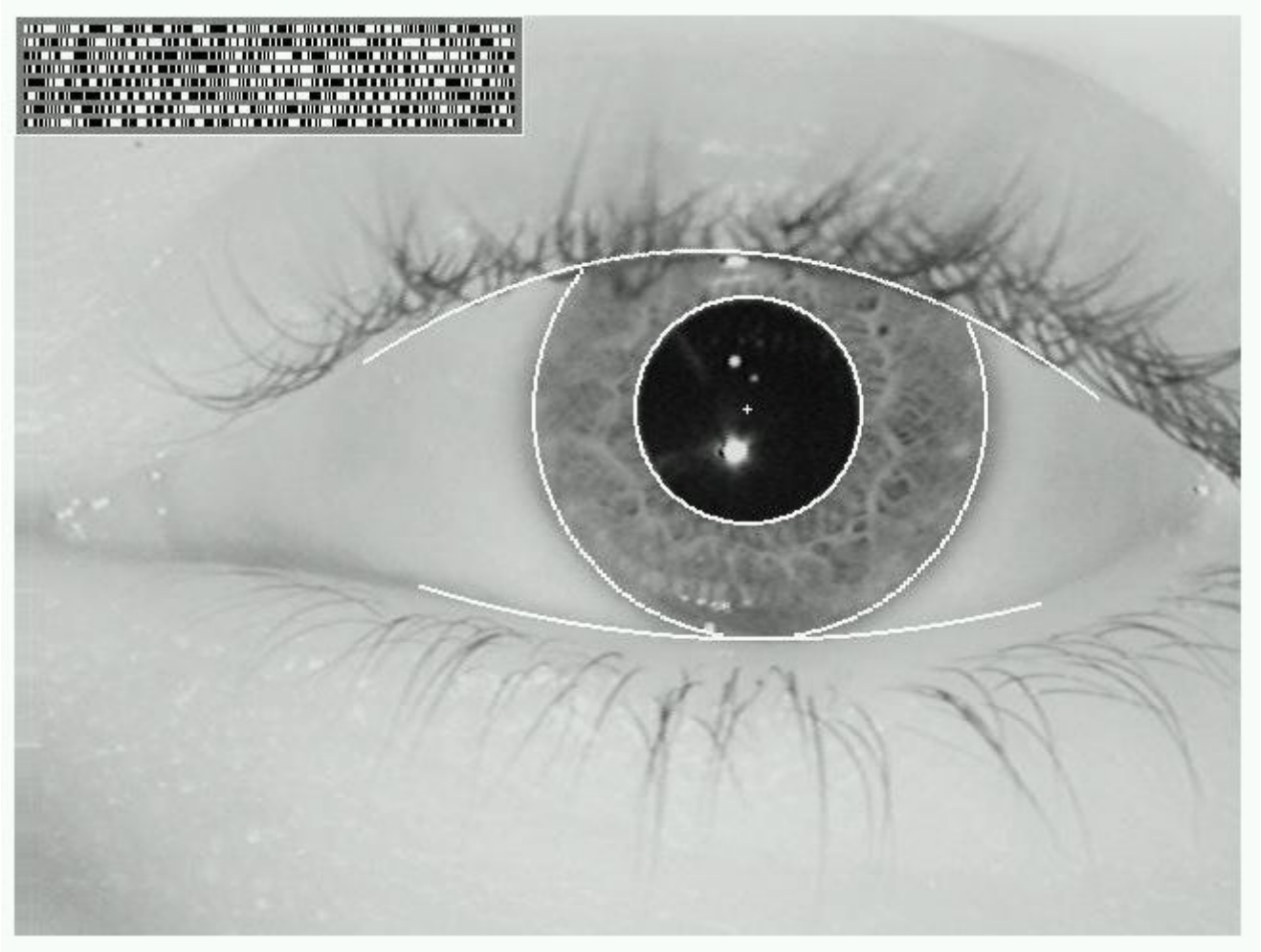
Detection of Communications Signals (even demodulation)

Example of Fingerprint Biometric Sample



Reproduced with permission from: "An evaluation of fingerprint image quality across an elderly population vis-Ã -vis an 18-25 year old population", Nathan Sickler and Stephen J. Elliott, IEEE International Carnahan Conference On Security Technology, Las Palmas De Gran Canaria, Spain, 2005:
http://www.carnahan2005.ulpgc.es/programme/presentaciones_pdf/12_Miercoles/2a/2.-Sickler%20Carnahan%202005%20Presentation.pdf

Example of Iris Biometric Sample



Reproduced with permission from Professor John Daugman of the Cambridge University Computer Laboratory: http://www.cl.cam.ac.uk/users/jgd1000/iriscode.jpg

Types of Pattern Matching

Statistical Pattern Matching: doing it with numbers

This presentation is about the statistical approach. Maths is difficult; many people find it very troublesome. But actually, measuring things accurately is very very useful.

Syntactic Pattern Matching: sequence matters

Examples are language and Automatic Speech Recognition. Word order matters, and there are multiple layers of interpretation. In physical systems, change with time matters too.

Neural Pattern Matching: we all do it

In many ways, brains do it better. But how? And particularly how did that ability evolve with a biological mechanism. Some people think doing it the same way is *the way*: IMHO probably not, but it is very interesting (and probably important).

Statistical Pattern Matching

Classification

The output is a decision: eg is this thing a dog, a cat, a rabbit, a crocodile or a fox. Input features can be discrete labels (eg fur or not; main colour: brown, grey, red, mixed; light, dark, etc) or continuous measurements (eg size, number of toes per foot, relative size of mouth to head).

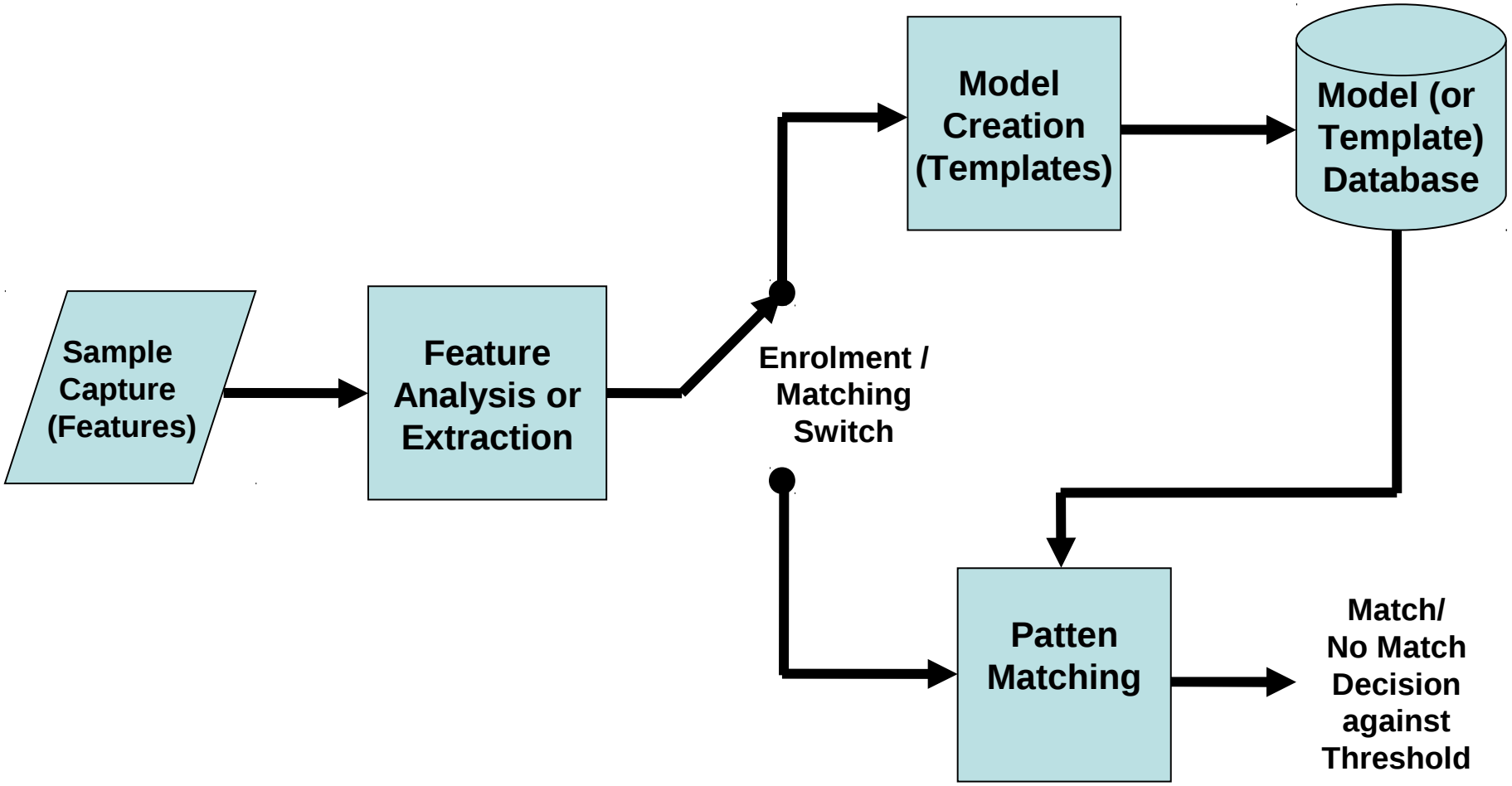
Regression

The output is a continuous variable (or several) as a combination of the input numbers: eg what is the expected residual life of this piece of equipment; what is the signal-to-noise-ratio. Input features can again be discrete labels or continuous measurements

Models and Parameterisation

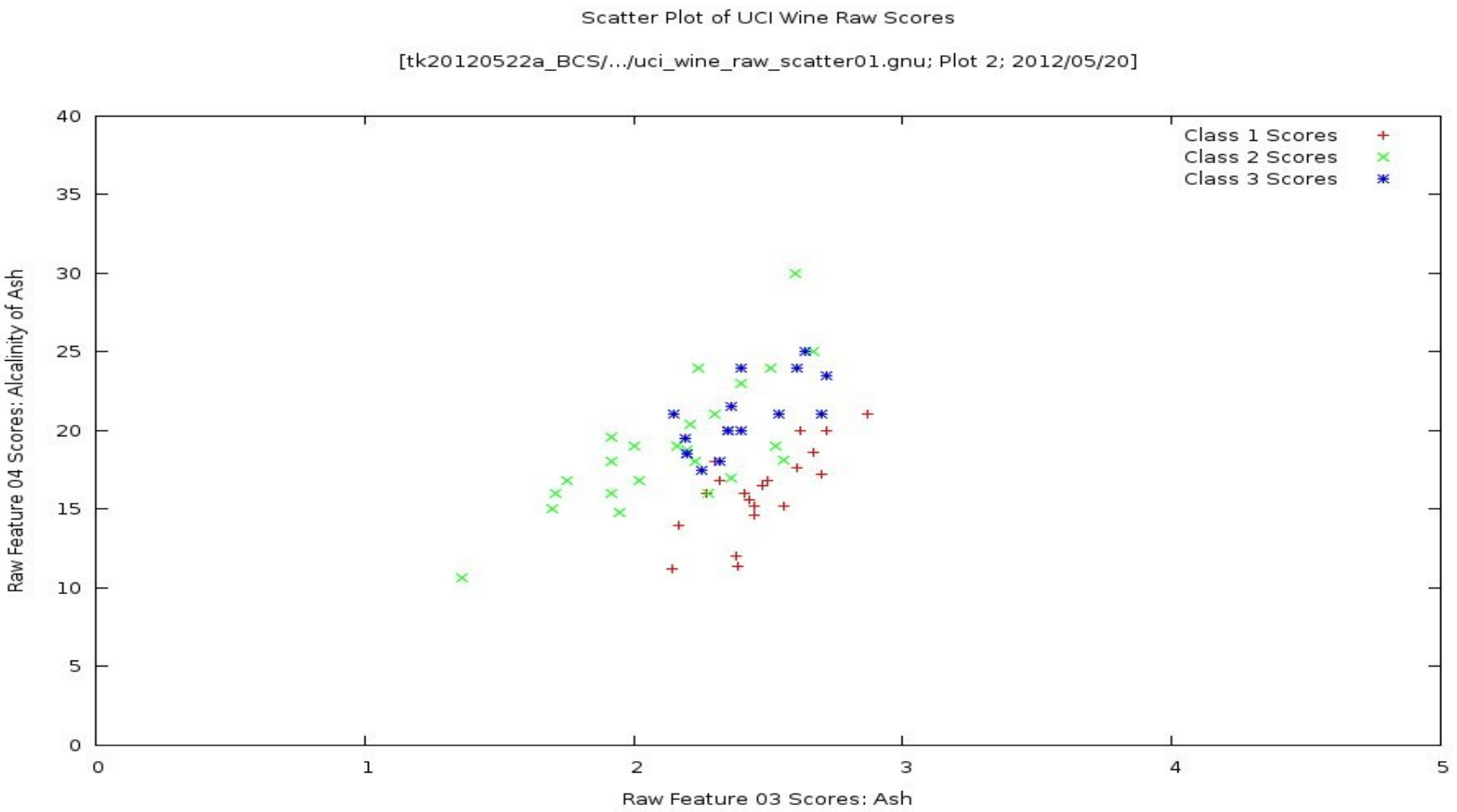
Both classification and regression use mathematical models, with parameters which are usually learned from examples. Supervised learning has examples with defined outputs; unsupervised learning (eg clustering) also must 'learn' the output classifications, instead of being told them (or less often, the regression output variables).

Structure of a Pattern Matching System for Classification



Example of UCI Wine Data (1)

Scatter Plot of 2 Features and 3 Classes: some correlation



Example of UCI Wine Data (2)

Summary of the Dataset

Wine of several vinyards, from 3 grape varieties

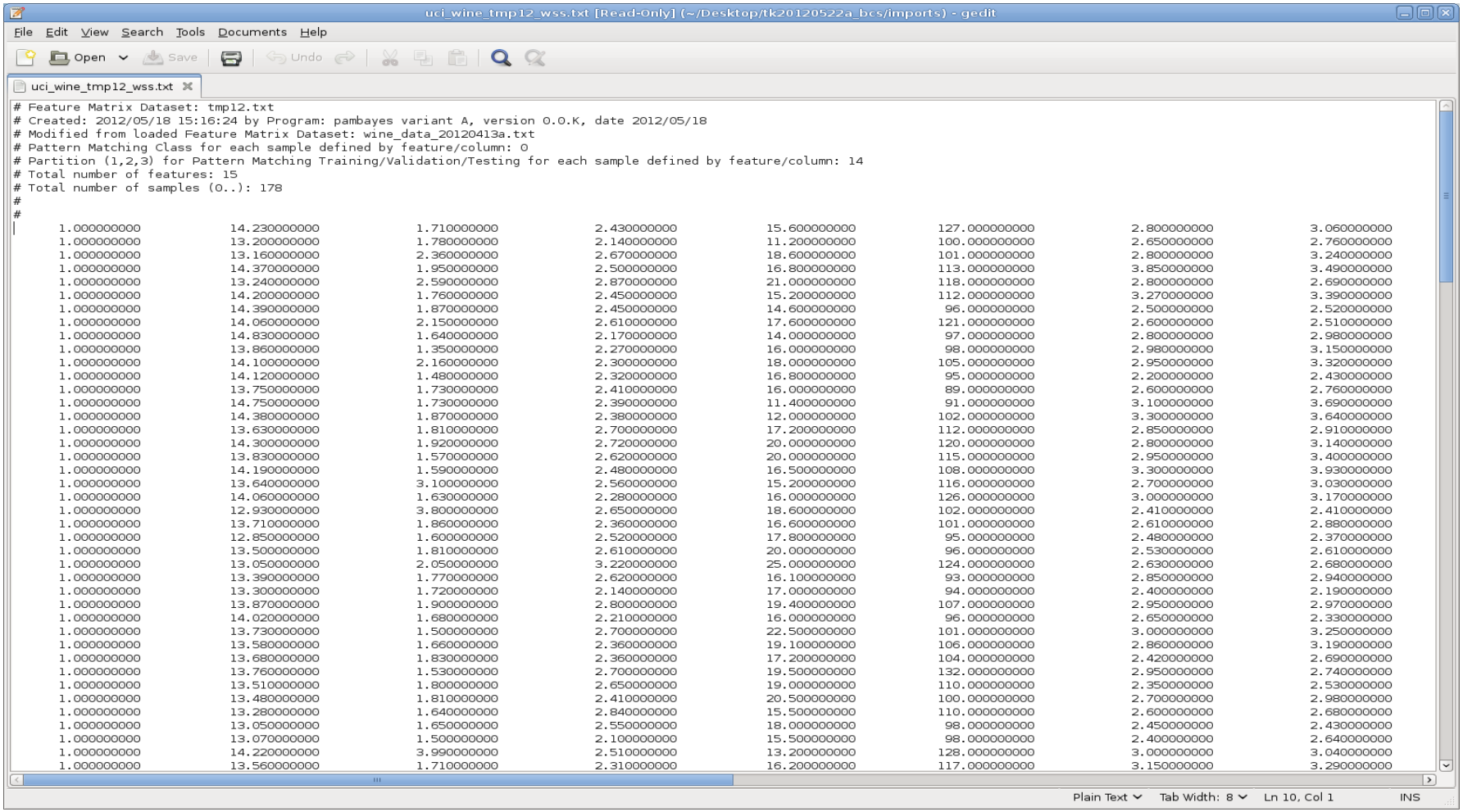
PM Class 1: 59 samples
PM Class 2: 71 samples
PM Class 3: 48 samples

Number of Features (Attributes, or Measurement Types): 13

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

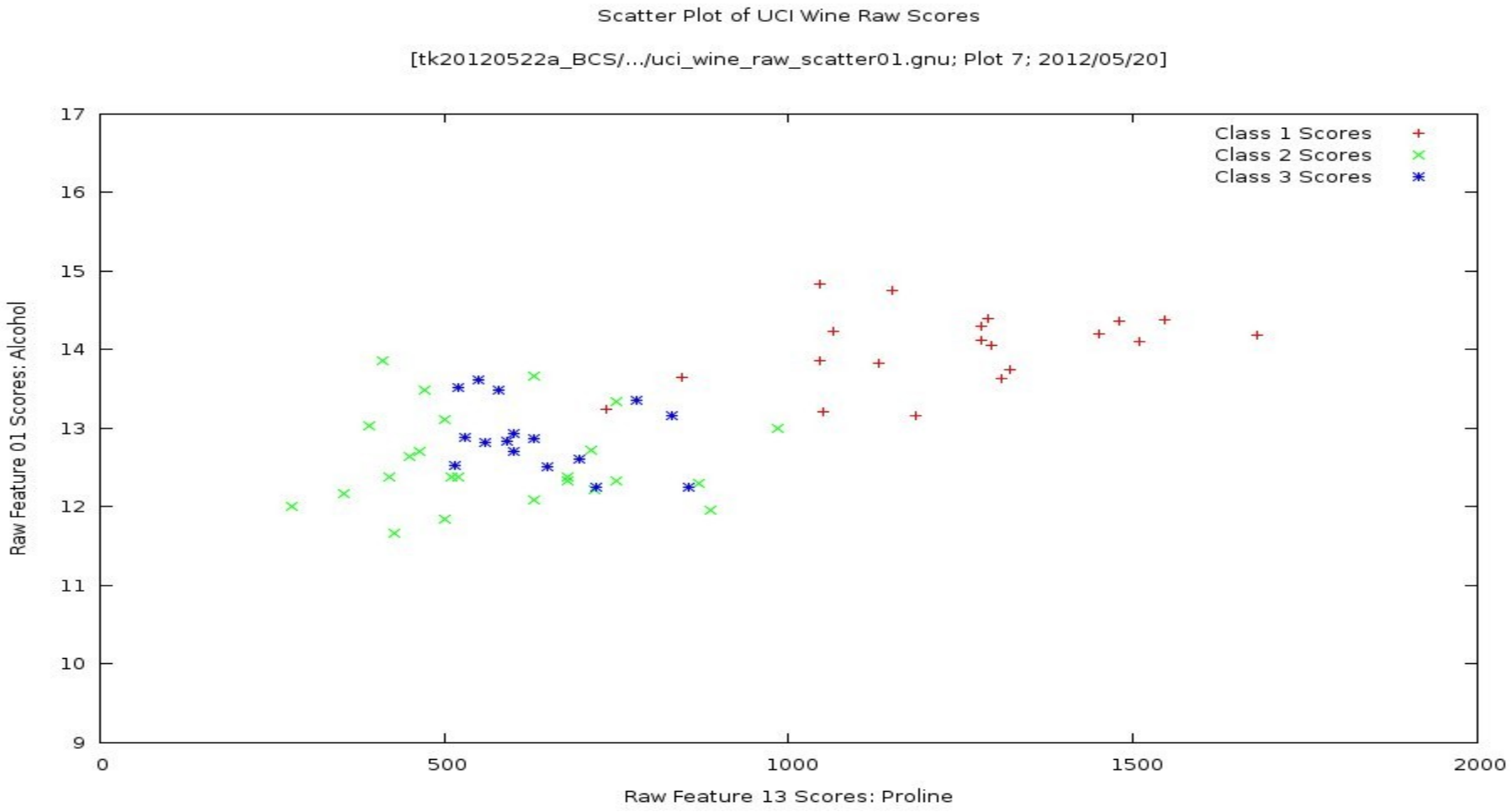
Example of UCI Wine Data (3)

Raw Data File: Header Comments and Features



Example of UCI Wine Data (4)

Scatter Plot: examples of good and bad separation, partial overlap



How do we Measure Performance

Many people, especially on biometrics (more in the public eye) talk of the *error rate*, but this single number has insufficient meaning.

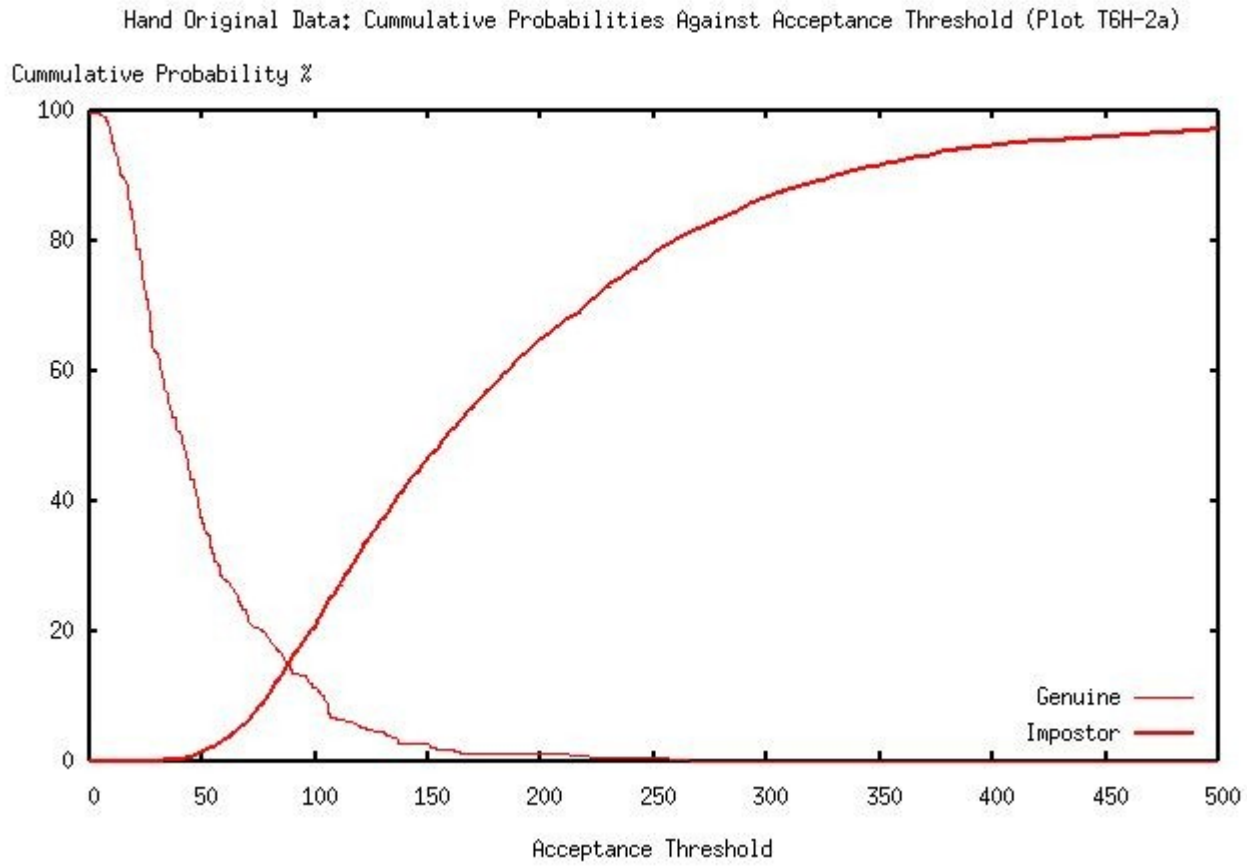
Every sensibly designed Pattern Matching System has a big control (it should be on the front panel) marked detection threshold (or operating point).

This changes the trade-off between False Alarms and Misses.

Any operator can change this value to get the best (for useful purpose) from the PM system. But only the implementors can change the trade-off curve, getting a better machine (usually for more money) or a lower-price one (usually with a worse trade-off).

The real thing is called the Receiver Operating Characteristic (ROC) curve. **[A term originally, I think, from radar just before WW2.]**

Plots of FNMR and FMR against Score

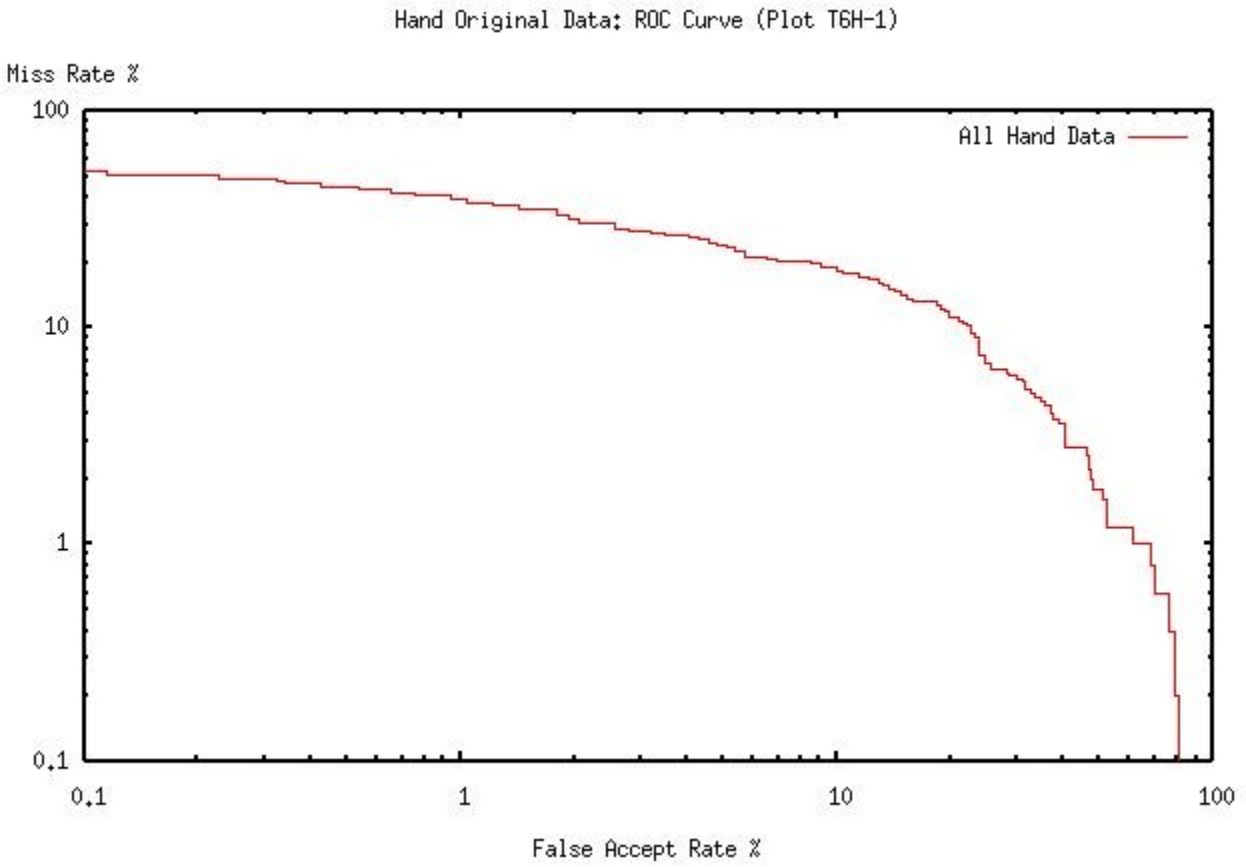


FNMR: **False Non-Match Rate (Verification Miss Rate)**

FMR: **False Match Rate (Verification False Alarm Rate)**

EER: **Equal Error Rate (FMR==FNMR)**

Receiver Operating Characteristic (ROC) Curve



Best plotted on a log-log scale, to give wide dynamic range.
Also known as the Detection/Error Trade-Off (DET) Curve.

Definition of Detection Gain

Detection Gain is the ratio of: the probability of the target being present given the evidence and prior knowledge, to the probability of it being present given just the prior knowledge.

Definition:

$$DG(e) = \frac{P(T|e)}{P(T)}$$

Consequence:

from Bayes Rule and from probabilities summing to 1.0

$$DG(e) = \frac{1.0}{P(T) + \frac{P(\sim T)}{LRtnt(e)}}$$

using Definition:

of Likelihood Ratio, Target to Non-Target

$$LRtnt(e) = \frac{P(e|T)}{P(e|\sim T)}$$

Approximation:

with P(T)<<1.0

$$DG(e) \approx LRtnt(e)$$

With Multiple Features

With evidence from a multiplicity of features:

$$LRtnt(e) = LRtnt(e1, e2, e3, \ldots)$$

Expanding in terms of generative PDFs:

$$LRtnt(e) = \frac{P(e1, e2, e3, \ldots | T)}{P(e1, e2, e3, \ldots | \sim T)}$$

With the assumption of independence of features:

$$LRtnt(e) = LRtnt(e1) \,.\, LRtnt(e2) \,.\, LRtnt(e3) \ldots$$

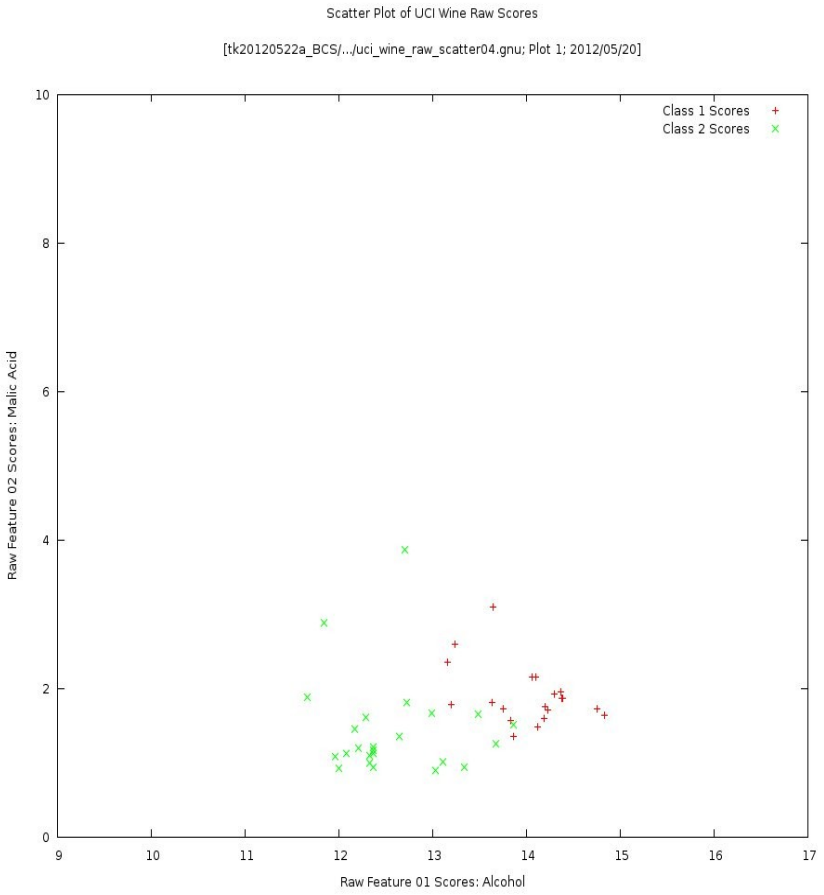
Note that, under the independence assumption, fusion of feature scores is independent of the *a priori* probabilities. Even if P(T) is not very small, the effect of *a priori* probabilities is taken into account after feature fusion.

Score Normalisation

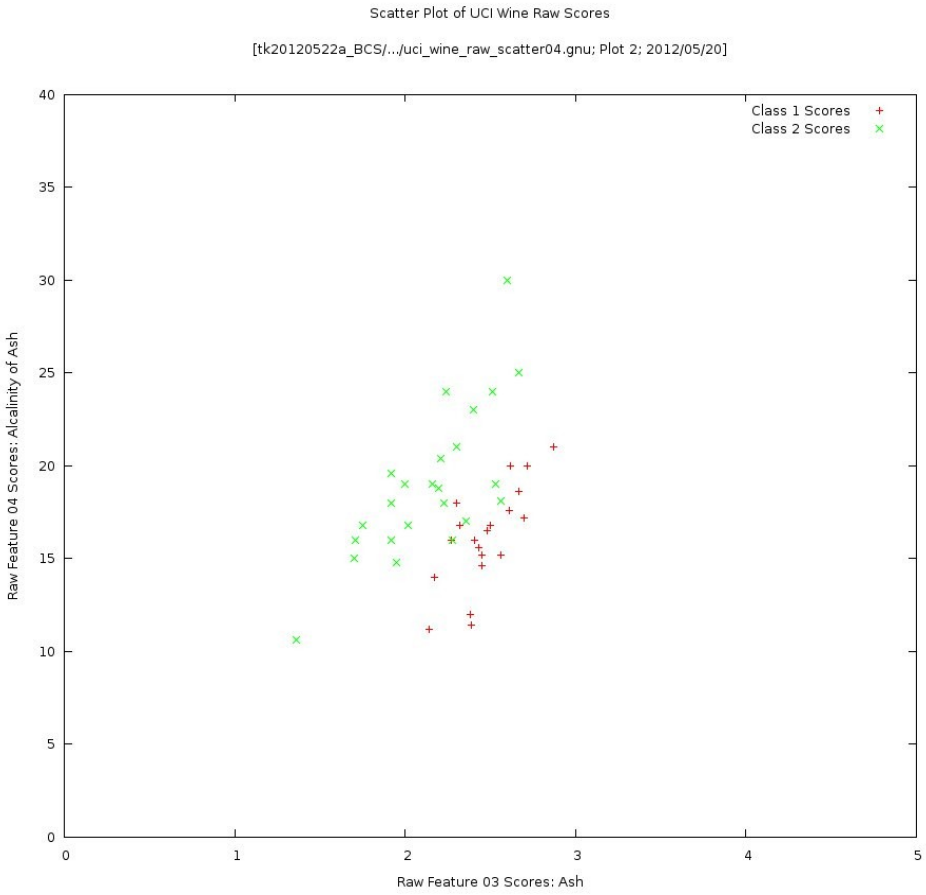
- Raw scores can be on arbitrary, device-dependent scales.
- It is meaningless to combine scores from different arbitrary scales.
- Score normalisation applies an appropriate transformation to scores from each modality/instance/algorithm, so that all normalised scores are on the same scale.

- Probability ordered scales: high scores match better.
- Distance ordered scales: low scores match better.
- Scores closely related to linear probabilities are usually best combined by multiplication.
- Scores closely related to log probabilities are usually best combined by addition.

Raw Feature Scatter Plots with Probability and Distance Scoring

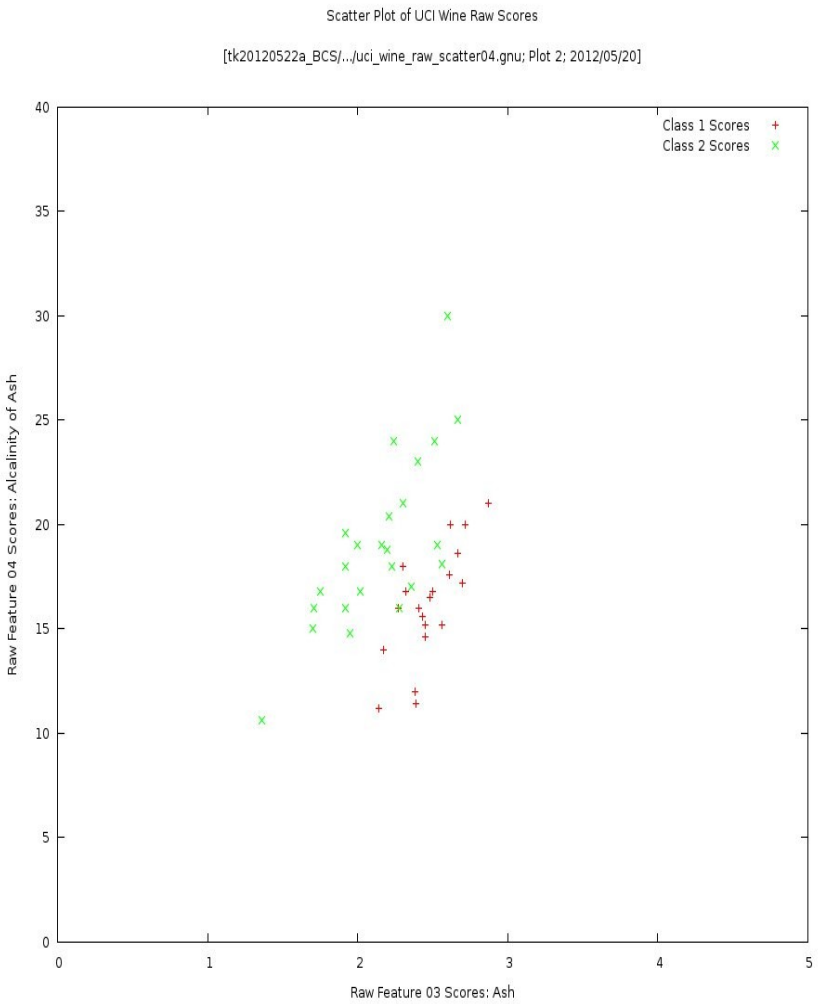


Wine Raw Features 1 and 2

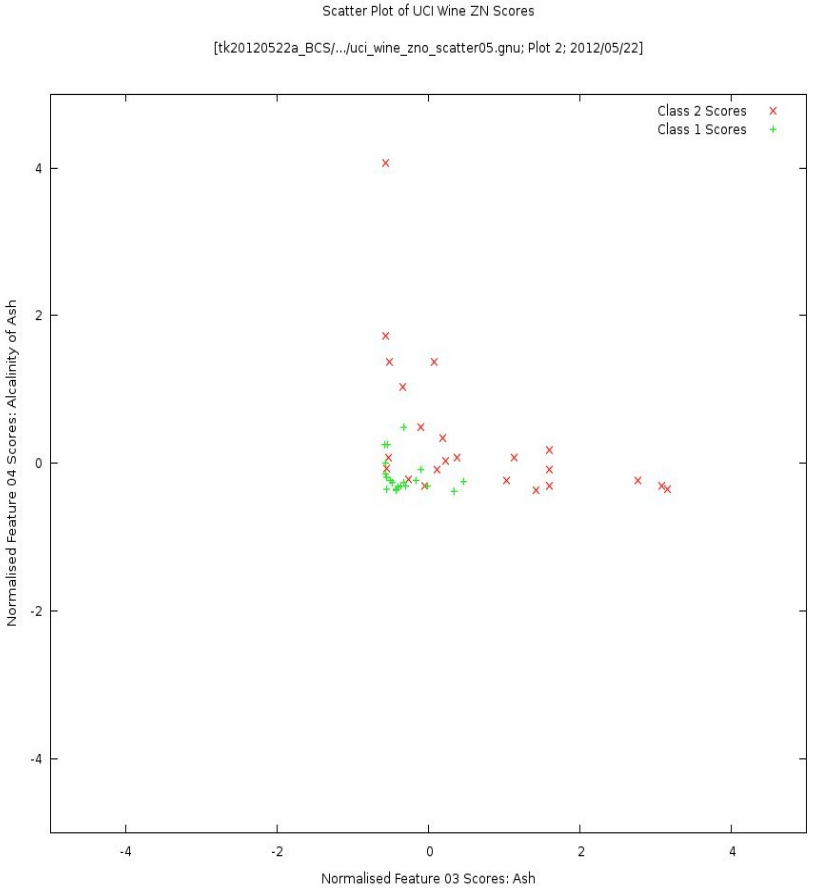


Wine Raw Features 3 and 4

Raw and Normalised Features

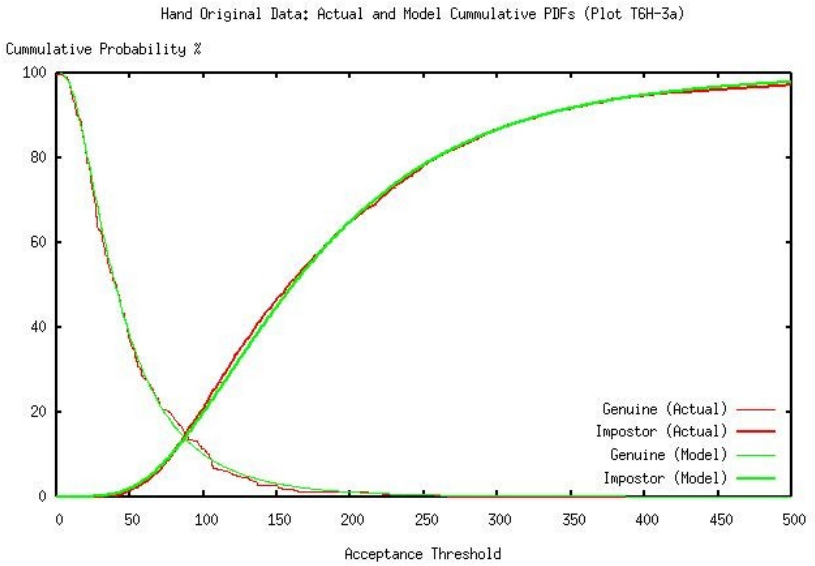


Raw Features 3 and 4

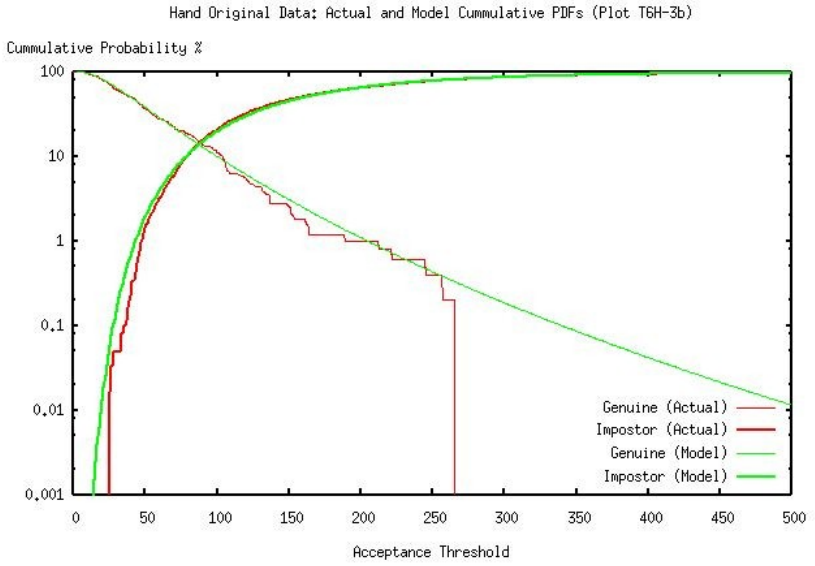


Normalised Features 3 and 4

Normalisation by Fitting of Parametric PDFs

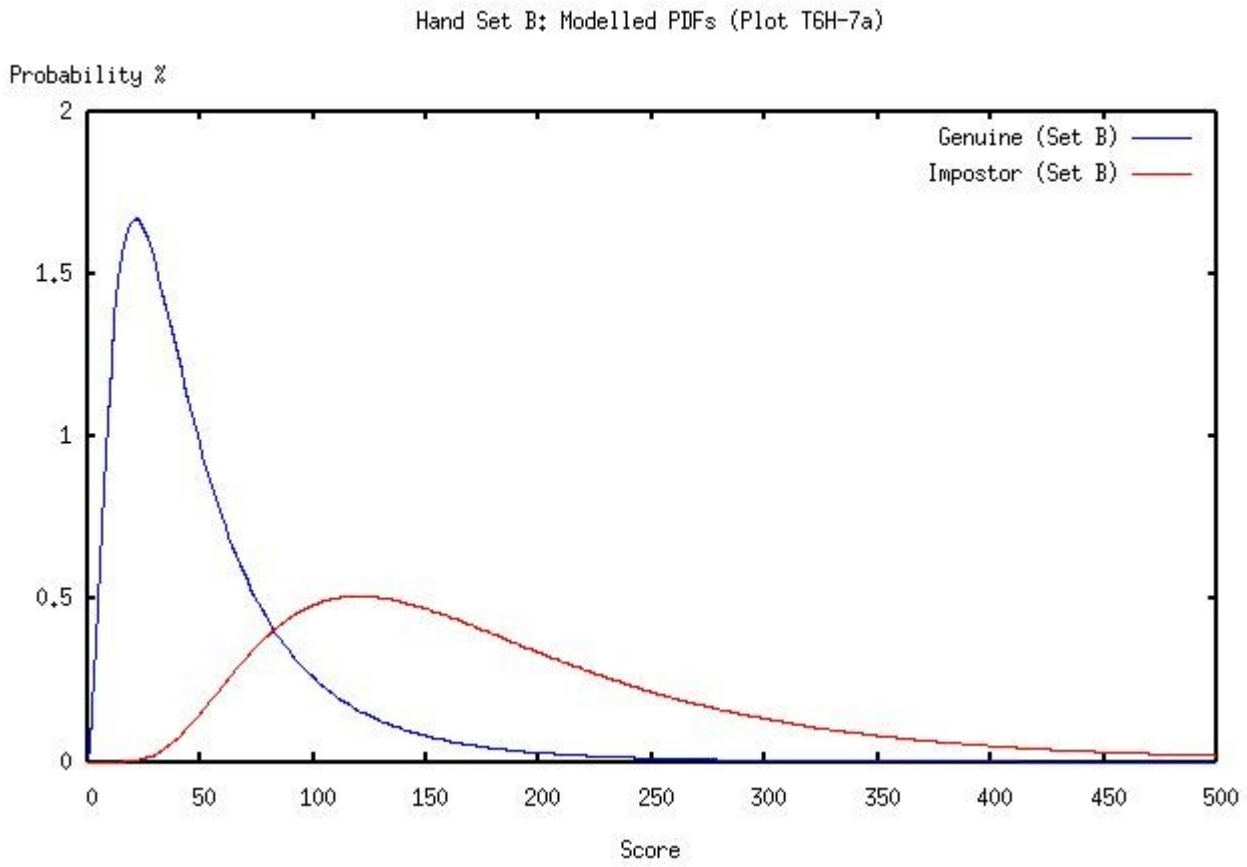


Hand Biometric Score Cumulative Linear Frequencies and Model PDFs



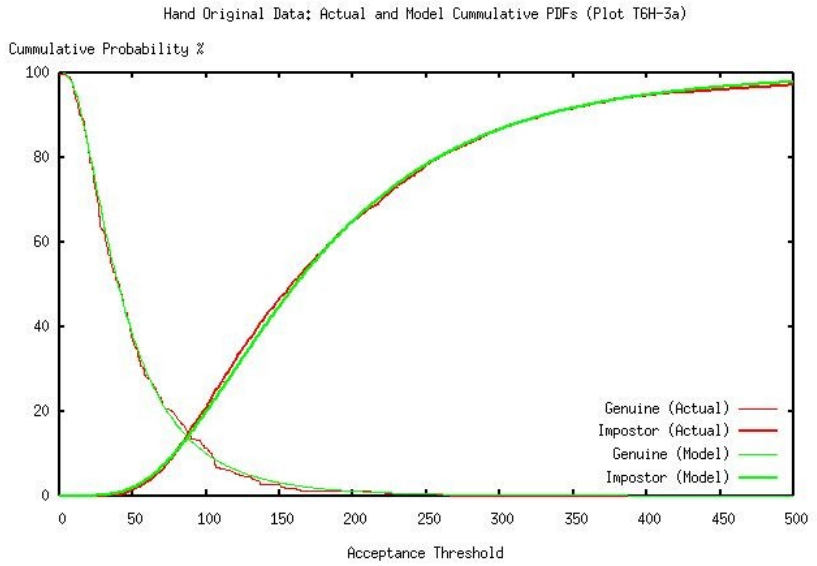
Hand Biometric Score Cumulative Log Frequencies and Model PDFs

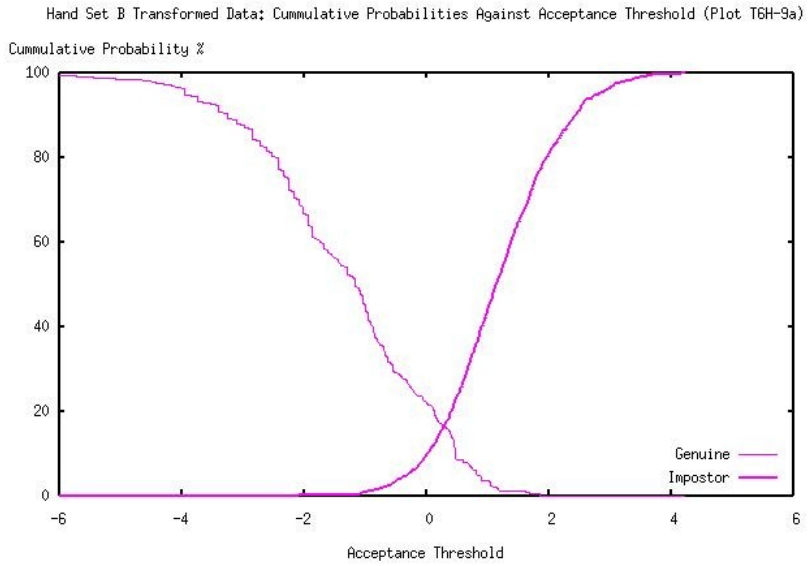
Non-Cumulative PDFs



Hand Biometric Non-Cumulative PDFs

Cumulative Frequencies Before and After Normalisation

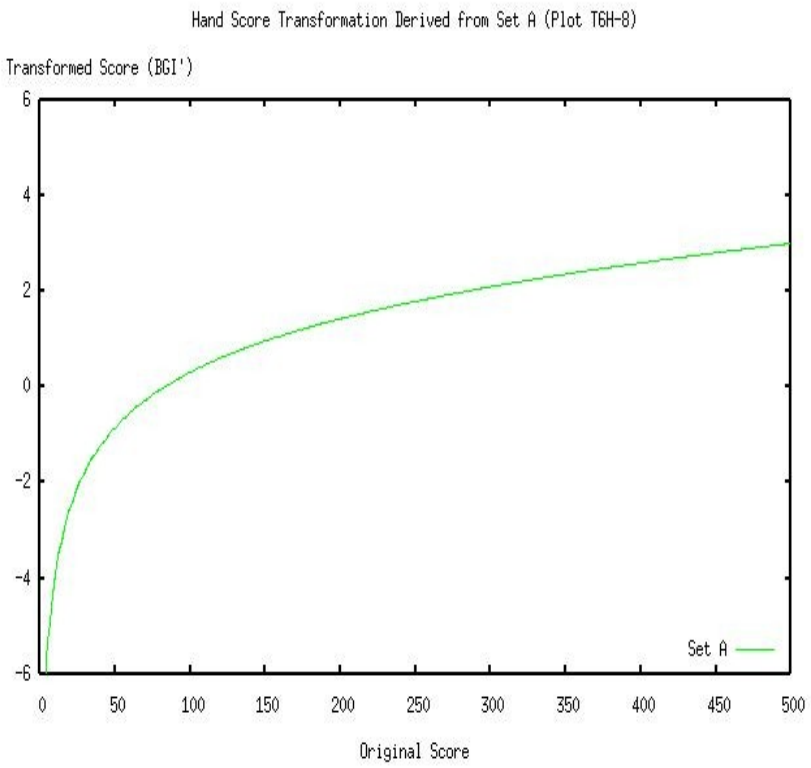


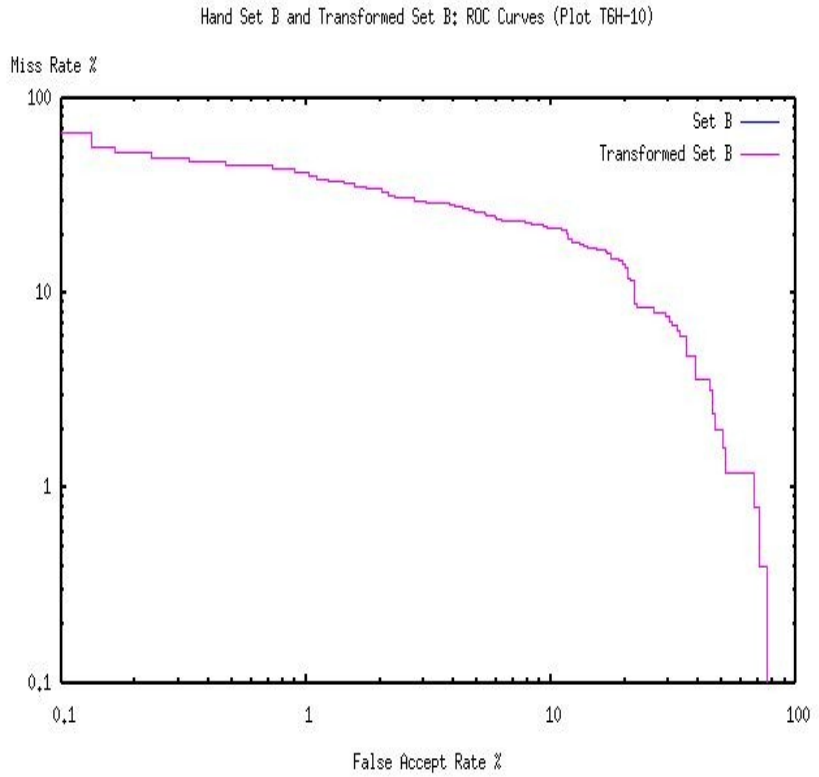


Hand Biometric Score Cumulative Linear Frequencies and Model PDFs

Cumulative Frequencies against Normalised Score

Normalisation Function and Monotonicity

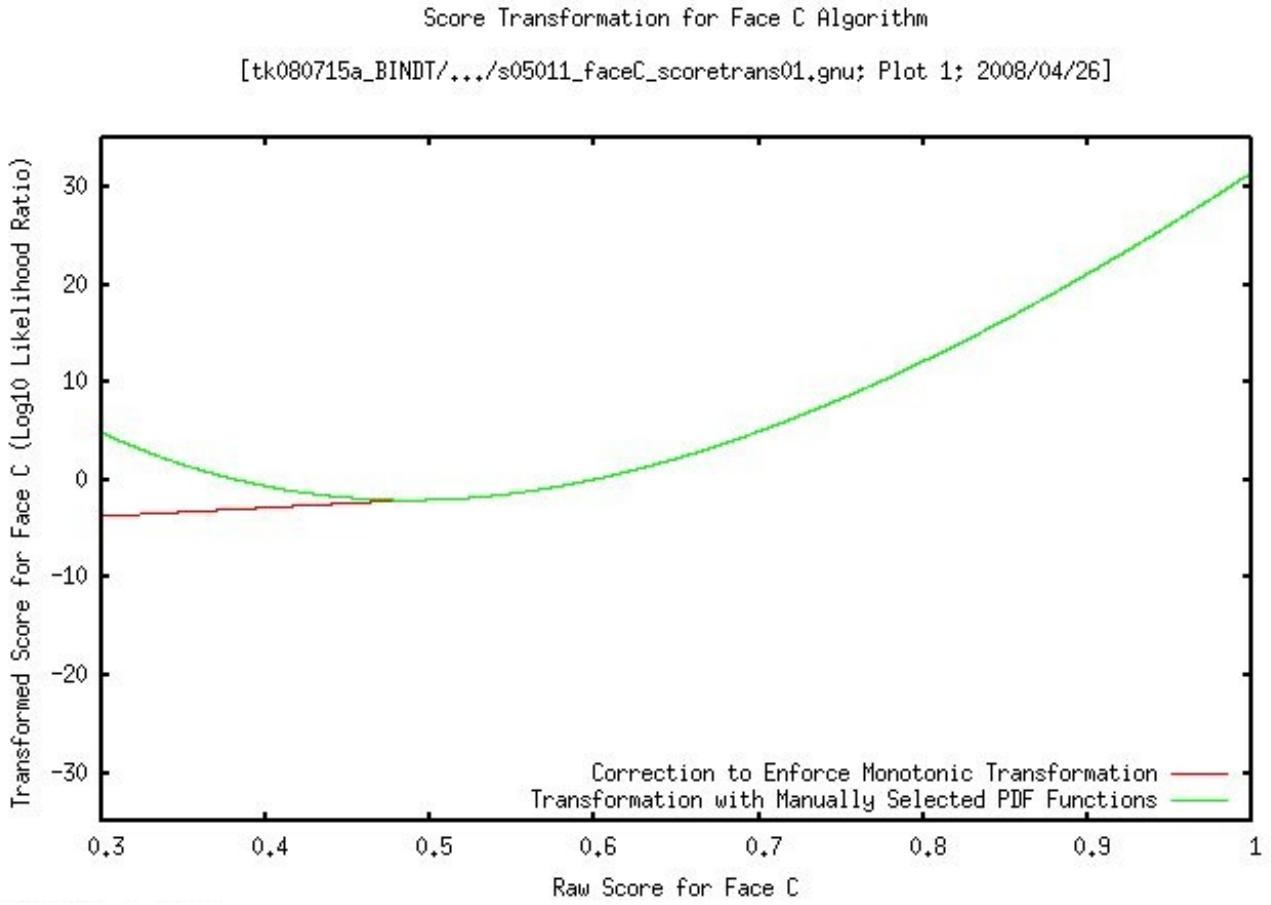




Hand Biometric Raw to Normalised Score Transformation

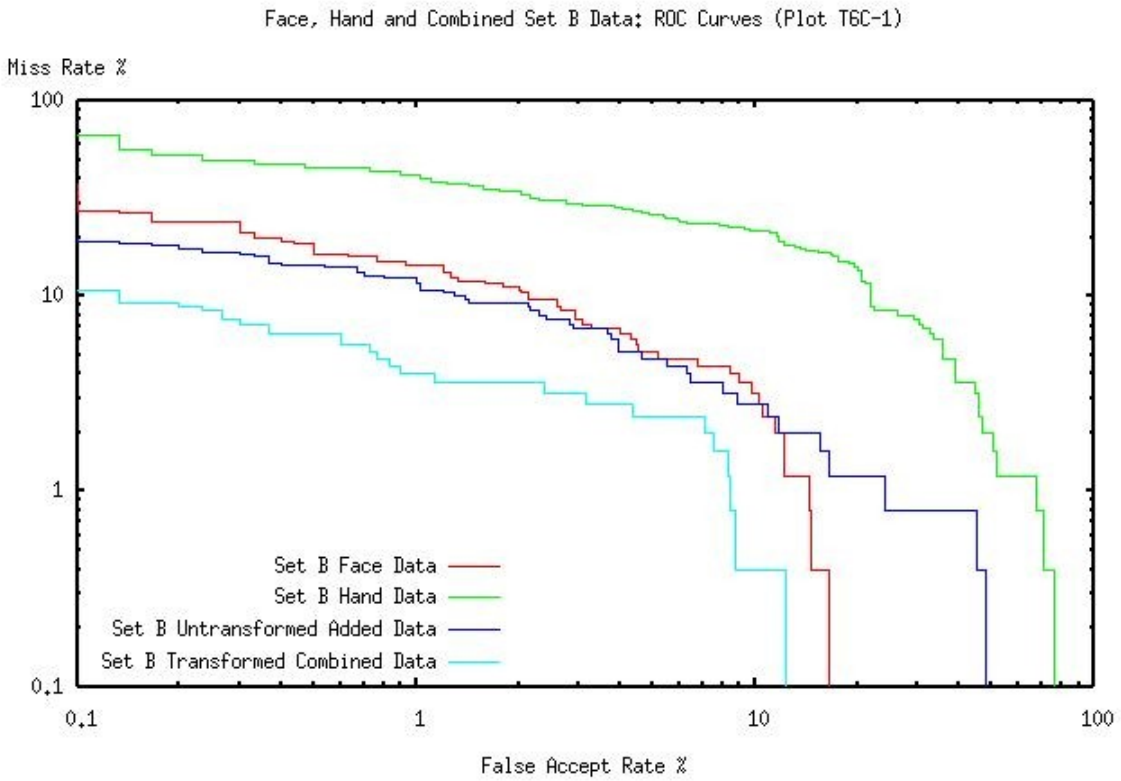
ROC Curves for Hand Biometric (Unchanged)

Normalisation Function with Semi-Automatic Correction



Score Transformation for Face Biometric C, with Correction

ROC Curves Before and After Multi-Modal Biometric Fusion



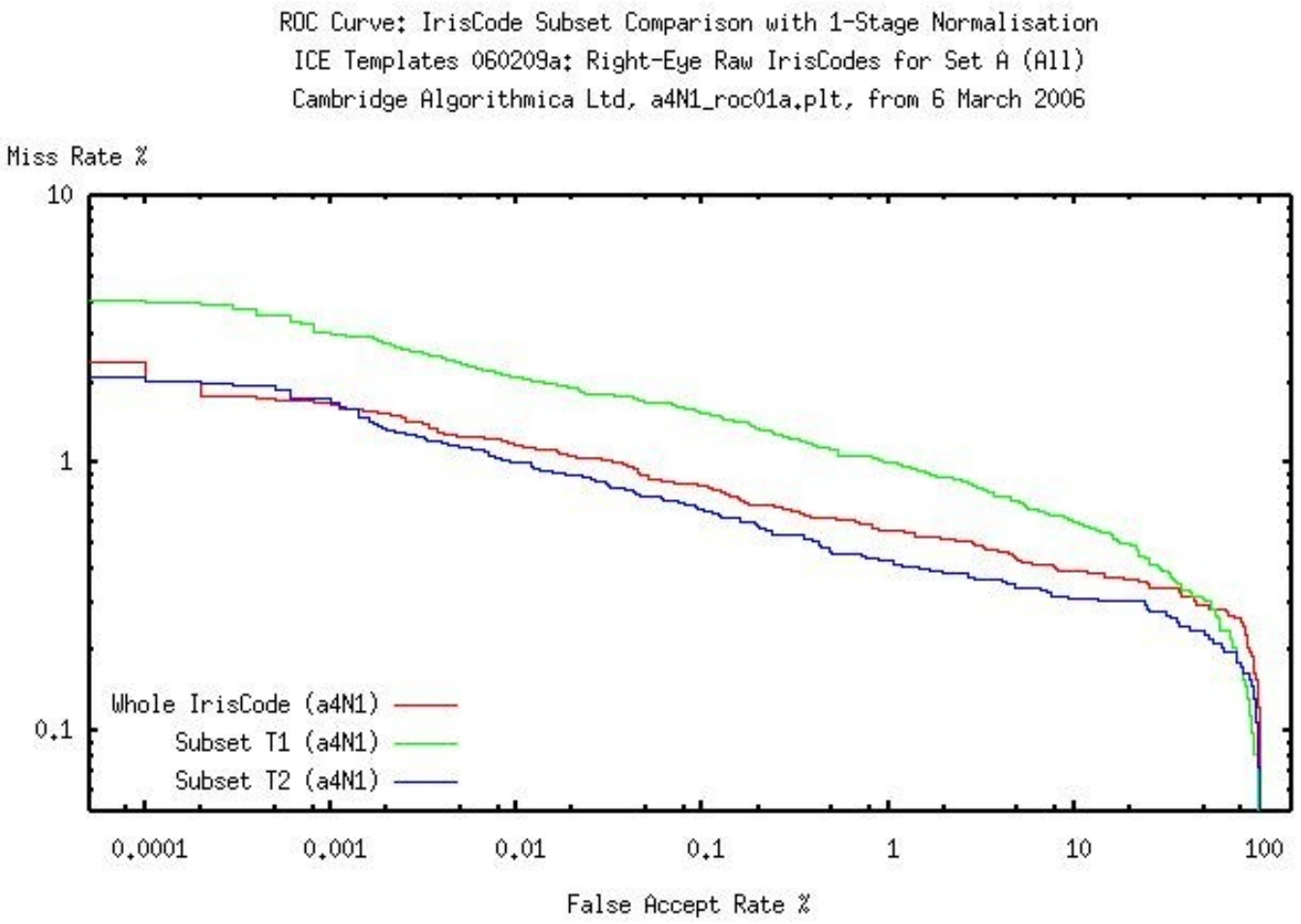
ROC Curves for Hand and Face Biometrics Individual and Fused Features

Why am I Doing This:
Some Problems Seen

- Involvement in Pattern Matching since 1974
- See increasing need for better performance in advanced applications (eg ASR and biometrics)
- 2003 – clear failure of the biometrics community to see the way forward on multi-modal biometric fusion
- Saw commonality with advanced work on multi-algorithmic demodulation of radio signals
- Saw other problems in PM approaches when got into the detail: an example follows (and one precedes)
- Vast improvements in computational power and data storage ability makes everything much easier to do – so why not!
- It is very interesting, and there are many benefits to business and society

Initial ROC Curve Comparison of T1 and T2 Subsets, and Whole IrisCode

Note unexpected ranking of T2 Subset as better than whole IrisCode.

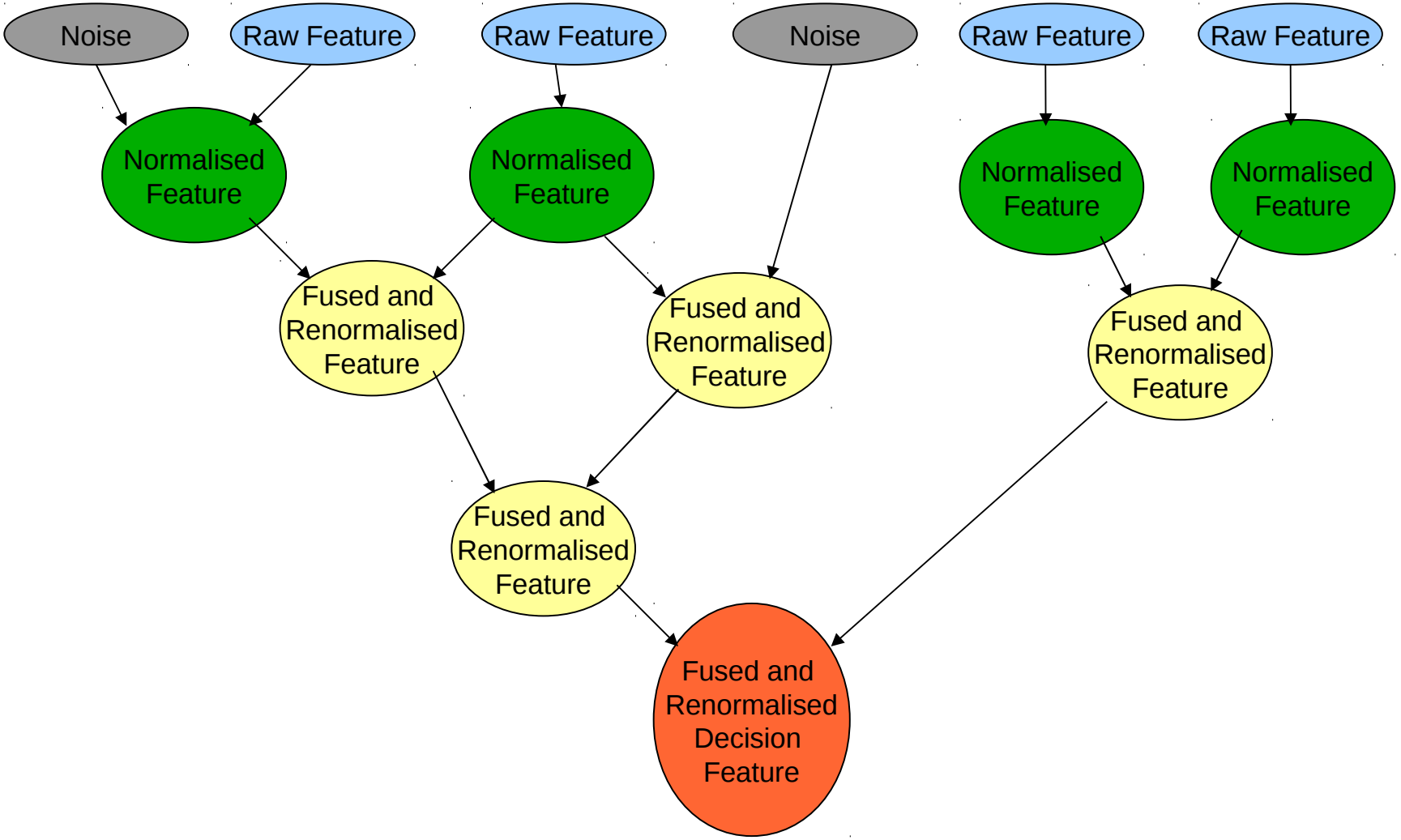


Dynamic Feature Fusions Trees (1)

- Statistical Pattern Matching, as a field, knows that nothing comes from nothing. Though we know what is optimal (multi-dimensional PDFs), we often (usually) don't know them: not enough training data; operational circumstances differ too much from training scenarios
- But, IMHO, we should do the right thing as much as possible, rather than pretending that the wrong thing is OK enough (well, yes perhaps – but that won't be so tomorrow or the next day)

- **Pairwise fusion of normalised features is close to the best thing, for the pair (especially with bell-shaped PDFs)**
- **If you add an extra feature, why not fuse what you were doing before, with the extra feature after normalisation: it may well be good enough**
- **Do better by seeing if that extra feature should be fused earlier in the process**

Dynamic Feature Fusions Trees (2)



Dealing with Noise and Missing Data

Classical Methods have some Problems. Many classical feature extraction methods have a problem with noise, and even more of a problem with missing data.

Using a Noisy Feature might make things worse than leaving it out.

Do we want to be Inventing Feature Values? If Principle Components Analysis (PCA), or something similar is being used and a feature is missing, what is to be done. During matching, most often, a value is 'invented' for the feature, that is at least somewhat consistent with the other (present) features.

Why not Normalise PDFs with Known Noise Levels? One can parameterise the class PDFs so that, with knowledge of the noise, they are 'broadened'. The normalised feature remains a true representation of the (log) likelihood ratio.

Why not just ignore the Missing Features? For Dynamic Feature Fusion Trees (DFFTs), with missing features, one leaves them out, and copies through the other feature.

Why not Parameterise Normalisations Assuming a Modest Number of Missing Features? Where copying through adversely affects subsequent normalisations, one can compute a (modest) number of normalisations: one for each combination of missing features. Again, the log likelihood ratios remain meaningful.

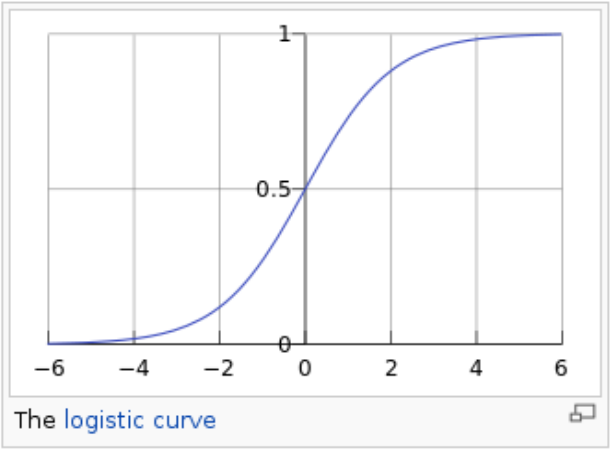
Parallels with Neural Networks (1) (?and the Brain)

A few decades ago (the 1980s most strongly), there was a great interest Artificial Neural Networks (ANNs) that were 'similar' to the operation of human and animal brains. ANNs have had some success, with a wide variety of different detailed approaches. However, they have not outstripped other PM algorithms

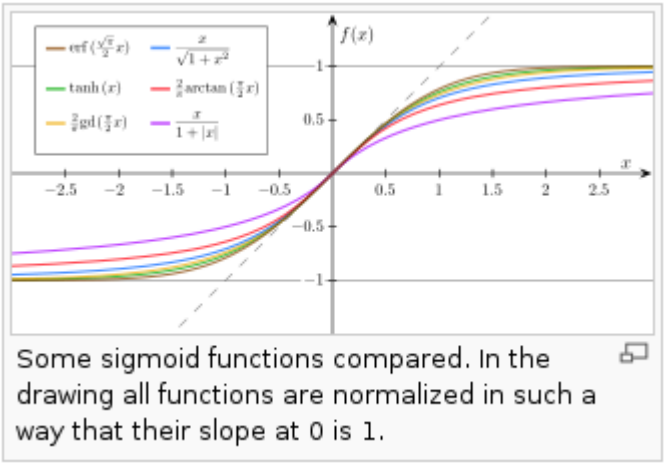
DFFTs and AANs have some noticeable similarities:

- **They are both feed-forward networks**
- **They apply a (usually monotonic) non-linear (eg sigmoid) transformation to the sum of inputs**
- **They apply a threshold to the output (or apply the equivalent to the input)**

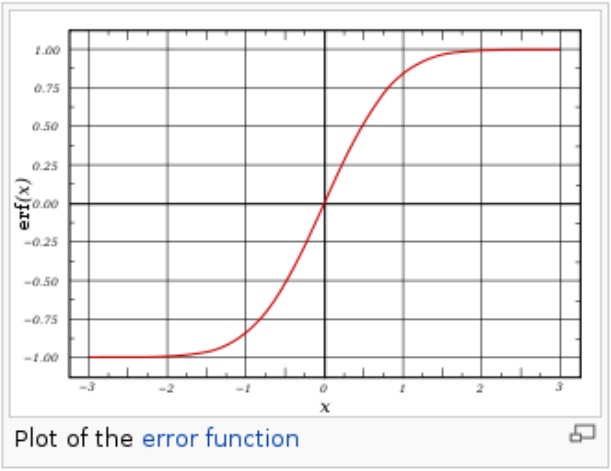
Parallels with Neural Networks (2) (?and the Brain)



The logistic curve



Some sigmoid functions compared. In the drawing all functions are normalized in such a way that their slope at 0 is 1.



Plot of the error function

Various sigmoid functions that are used for ANNs (and other things).

Parallels with Neural Networks (3) (?and the Brain)

The differences between DFFTs and ANNs are, however, important.

ANNs use, in the main, empirically specified sigmoid functions. These may be parameterised as part of the ANN training process. However, that is by no means certain, as the connection weights are the main thing to learn. If sigmoid parameters are learned, it is usually done as an offset and as a scaling of the width of the sigmoid.

DFFTs use, instead of a sigmoid, a data-driven transformation that is actually the ratio of two PDFs. This theoretic basis, from Bayes Rule and other aspects of statistics is, IMHO, quite different and quite important.

An interesting question: do human/animal brains acquire transformations as well as connections and weights? How?

Discussion

1. **Why sometimes does it work so well to 'ignore' the correlation of features? Certainly it is advantageous, as it avoids or reduces "the curse of dimensionality"**
2. **Possible answer: consider the case of 2 features that are identical; thus they are perfectly correlated. The second one adds no information.**
3. **Fusion of scores by multiplication of likelihood ratios just effectively squares the score of the 2 individual original features**
4. **This is a monotonic feature transformation, so does not change the ROC curve; but it does render unnormalised, the fused score**
5. **So one aspect of 'naïve' fusing of normalised but correlated (log) likelihood ratios is that the fused score is unnormalised**

Are DFFTs Dynamic? New? Simple? Flexible?

Conclusions

There are many aspects of DFFTs that draw on previous work and knowledge common to the field of statistical pattern matching.

However, there seem to be some other aspects that offer scope for improved performance, with somewhat simpler requirements on understanding (and the level of maths).

A key issue is selection of the PDF family, particularly for raw features. This can be done by comparison, using the Bayes Information Criterion (BIC, or similar), that allows ranking of PDF models with different numbers of parameters.

The increasing availability of computer power and massive storage makes easier the use of complicated algorithms; it also helps with experimentation and evaluation.

Directions of Future Work

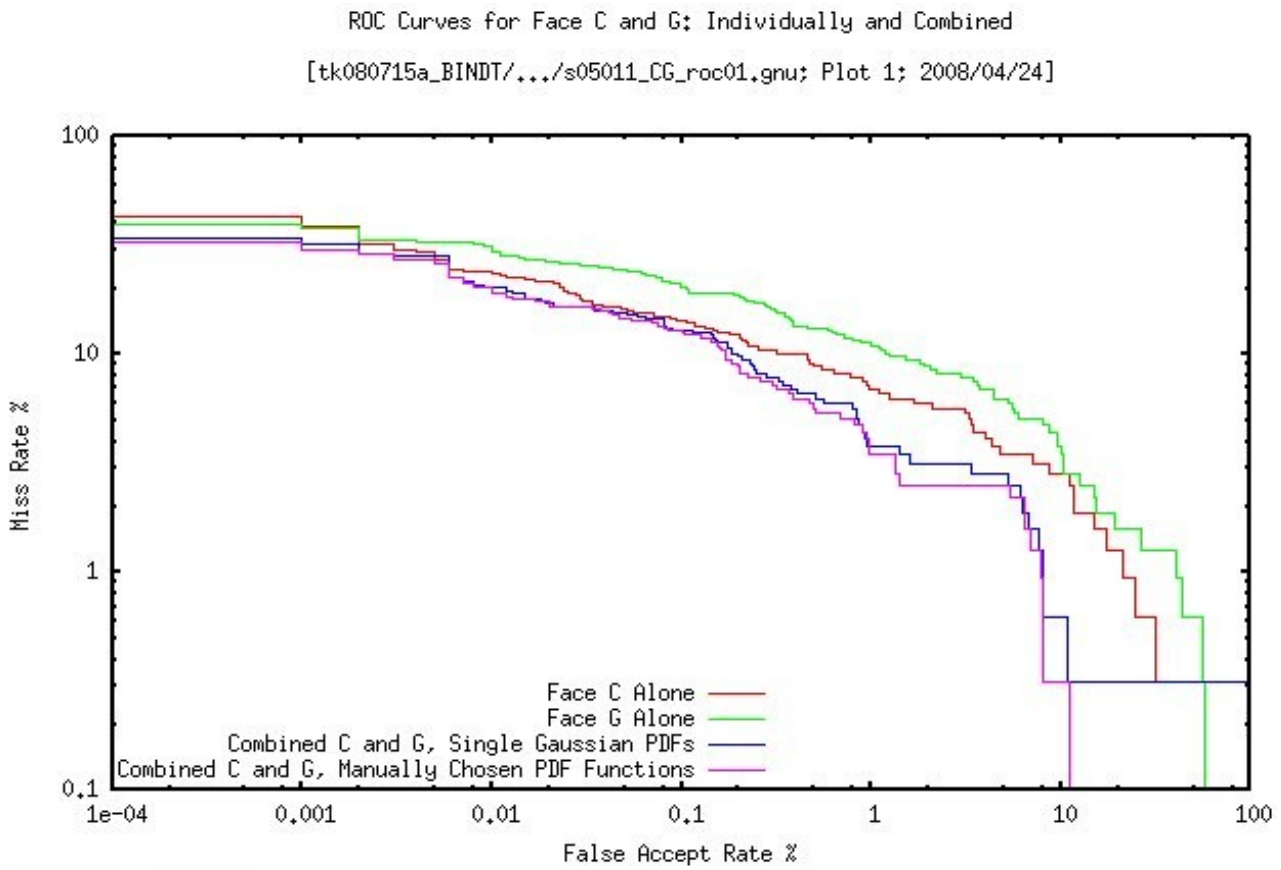
1. **Implementation is underway, of GUI-based software for automatic optimisation of parameters and to facilitate manual assistance in PDF selection, modelling and correction.**
2. **Evaluation on more extensive experimental datasets, including many others from the University of California at Irvine (UCI).**
3. **Evaluation of the approach to arbitrary numbers of features, including automatic building of binary trees defining good orders for fusion of features.**
4. **Systematic extension of the approach to arbitrary numbers of pattern classes, ensuring loops of preferences do not form in ways that cannot be handled automatically.**
5. **Integration of DFFTs with syntactic approaches, as necessary for Automatic Speech Recognition, etc.**

Dataset Source Acknowledgements

1. **Experimental data publicly available through the UCI Machine Learning Repository (MLR):** ***http://archive.ics.uci.edu/ml/***
2. **Stefan Aeberhard et al, for the wine dataset** (available from the UCI MLR). Main reference: S. Aeberhard, D. Coomans and O. de Vel, ***Comparison of Classifiers in High Dimensional Settings***, Tech. Rep. no. 92-02, (1992), Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.
3. **Professor John Daugman of the Cambridge University Computer Laboratory, for provision of biometric performance data on Iris Recognition.** Main reference: John Daugman, ***How Iris Recognition Works***, IEEE Trans on Circuits and Systems for Video Technology, CVST 14(1), January 2004, http://www.cl.cam.ac.uk/users/jgd1000/csvt.pdf
4. **National Institute for Science and Technology (NIST) of the USA for provision of performance data from two face biometric algorithms (2004, BSSR1).** Previously described at URL: ***http://www.itl.nist.gov/iad/894.03/biometricscores/***
5. **University of Michigan (USA) for provision of performance data on hand and face biometrics.** Main reference: Arun Ross and Anil Jain, ***Information Fusion in Biometrics***, Pattern Recognition Letters, 24 (2003). Also private communications.

Wallpaper and Inspiration

ROC Curves Before and After Multi-Algorithmic Biometric Fusion



ROC Curves for Multi-Algorithmic Fusion of Face Biometrics (algorithms C and G)