# Chapter 9
# Group Evaluation: Data Mining for Biometrics

Chapter 7 presented methods for the holistic analysis of biometric systems, while Chap. 8 illustrated how the performance of individual users within a single system can vary. In a similar manner, certain subsets of the user population may be consistently having difficulty with the system, while others may be performing very well. For example, assume that a particular system has a significant goat population (recall that a goat is a user who has trouble matching against their own enrollments). On one hand, it is possible that each of these people has a unique reason for their poor performance. However, it is more likely that there are a few common underlying causes that affect whole groups of people. Discovering these factors, and the groups they have the greatest effect on, is an important part of the analysis of biometric systems that is often neglected.

The following are hypothetical systems that have groups of problem users:

- All fingerprints can be classified based on their overall pattern of ridges and valleys. The main classes are: left loop, right loop, whorl, arch, and tented arch. An automated fingerprint identification system (AFIS) may apply a classification algorithm to an input print, and only match it against enrolled prints belonging to the same class. This reduces the number of one-to-one comparisons that need to be conducted, reducing system load and potentially avoiding false matches. However, fingerprint classification is a difficult problem in itself, with challenges distinct from those of fingerprint recognition. Consider a system that uses a fingerprint classification algorithm for pre-selection, and further assume that the algorithm often misclassifies whorl inputs as arches. In this case, the "whorl" sub-population may consistently receive low scores, for both genuine and impostor matches, leading to a group of phantoms. In this case, it is features inherent in the physiology of the subgroup that are related to their poor system performance.
- Covert surveillance systems capture images of people without their knowledge. Therefore, unlike many biometric systems, there is very limited control over the behavior of the subjects who pass through the system. Consider a group of users who wear large sunglasses that obscure a significant portion of their face. This will hamper the ability of the face recognition algorithm to correctly identify the

individual. In this case, it is a behavioral aspect of the subgroup that leads to poor recognition performance.

- Consider a face recognition identification system that is installed at several detention centers throughout a country. At each site, detainees are enrolled in the system with a new photo, which is matched against the existing database to ensure they have not been previously enrolled under a different name. Since there are a variety of different capture locations throughout the country, the conditions at each site will vary; some variations favorable to face recognition systems, others unfavorable. For example, imagine that one site has a problem with backlighting, resulting in a disproportionate number of false accepts. In this case, the lamb population is due to environmental factors.

As these examples illustrate, there are many potential reasons why a particular group may perform poorly. Large, integrated, full-scale production systems are complex and have many sources of data. Each of these sources introduce new factors that potentially relate to system performance.

The subject of this chapter is detecting problem groups. In general, it is assumed that the biometric data is available a priori, either from an evaluation, or from a live system. Typically, the system-wide performance has already been established, and further analysis is being conducted to determine if any groups are causing a disproportionate number of system errors. Section 9.1 outlines the data relevant for this mode of analysis.

Very little research has been published about evaluating the performance of user groups for biometric systems. Traditionally, this type of analysis has been a largely manual process. The performance of common subgroups (e.g. gender and age) is established by filtering the system-level results, and computing performance measures for each group individually. This approach is explained in Sect. 9.2.

Data mining and machine learning algorithms can be used to discover patterns and trends in biometric data. This has the advantage that the process is largely automated, so in theory subtle trends may be uncovered that would otherwise be hidden among volumes of score logs and metadata. This approach is discussed in Sect. 9.3. Section 9.4 contains a discussion of approaches for dealing with problem groups once they have been identified, and Sect. 9.5 presents the limitations of group-level analysis.

After reading this chapter, you should know:

- The types of metadata information that is most relevant to group analysis (Sect. 9.1).
- How to evaluate the performance of common groups, such as age, ethnicity, and gender, by partitioning the test data (Sect. 9.2).
- How data mining techniques can be used to automatically detect groups of problem users in your data (Sect. 9.3).
- What action can be taken to deal with known problem groups (Sect 9.4).

## 9.1 Group Metadata

The introduction to this chapter gave several examples of user groups with poor performance for hypothetical biometric systems. The sources of the problems varied widely, from physical characteristics of the people themselves, to their behavior and to their environment. This section provides a framework for the types of factors that can impact group performance. The list is not meant to be exhaustive, as the information that is relevant depends heavily on the system under evaluation. Instead, the goal of this section is to categorize the various sources of data, and illustrate how they can impact group performance. Annex C of ISO 19795.1 contains a more comprehensive listing of performance factors relating to biometric systems [4].

It is important to note that for the purposes of this chapter, a "group" is not necessarily restricted to a group of users. A group can be defined at the template level (e.g. all enrollments from a particular location) or at the match level (e.g. all verifications that occurred in the afternoon). However, as groups of users are most commonly considered for discussion, for this chapter a "group" should be assumed to mean a group of users unless otherwise specified.

### *9.1.1 User Level*

At the user level is information about an individual that is relatively constant over time. In other words, data that is unlikely to change between enrollment in a biometric system and subsequent verification or identification transactions. These are some examples of user level data:

- **Sex:** In some biometric systems, one gender may perform better than the other. This may be due to either physiological or behavioral reasons. For example, the vocal range of men and women is different, and may influence their performance within a speaker verification system. An example of behavioral factors are the use of fashion accessories or makeup by women that inhibit facial recognition by obscuring or altering part of the input signal.
- **Ethnicity:** Some recognition algorithms are tuned using a training set, that in essence "teaches" the algorithm to distinguish between people. However, if a particular ethnic group is over-represented in the training set, the algorithm may be biased towards them, and struggle to distinguish other groups. This is particularly relevant for face recognition systems where the effects of race are most visible, however it has been demonstrated to be a factor in other biometrics as well.
- **Occupation:** A person's occupation may, over time, alter their biometric. For example, the fingerprints of people who work extensively with their hands, such as bricklayers, are known to fade over time. This can have a negative impact on the person's performance within a system.

- **Accent:** A person's accent may be relevant to their performance within a speaker verification system.

## *9.1.2 Template Level*

Information regarding templates and samples is the largest category because it embodies all the information relevant to the presentation of a biometric at a particular instance in time. In general, this is data that is likely to change between subsequent presentations of the same biometric characteristic. This category can be further divided into sub-categories for user, environment and system. Examples from each category are found below.

### 9.1.2.1 User

- **Age:** The age of the user depends on the date that the biometric was captured. In some systems, certain age ranges may pose more difficulties than others. For example, children present a unique challenge due to the fact that some biometrics (such as face and voice) can range relatively rapidly during adolescence. Other biometric systems are known to struggle with the elderly.
- **Behavior:** Certain behaviors can influence interaction with a biometric device. For example, familiarity with the system and user motivation can affect the quality of capture. However, in many circumstances behavior can be difficult to measure, quantify, or classify.
- **Mood:** Some biometrics, especially behavioral biometrics, are influenced by the mood of the subject. For example, anger can alter the way one speaks, causing a problem for speaker verification.
- **Physiology:** This category depends heavily on the biometric being used. In general, there are many physiological aspects that can change between interactions with a biometric system. For example, with face recognition an important factor is facial hair, such as beards and mustaches which can impair performance. Annex C of ISO 19795.1 contains an extensive list of physiological factors that can influence the different biometric modalities [4].
- **Clothing and accessories:** As with physiology, the relevance of clothing depends on the biometric being used. For example, large, bulky overcoats may hamper the recognition performance of gait systems, and contact lenses are known to impact the performance of iris recognition systems. On the other hand, voice and fingerprints are unlikely to be affected by clothing.

#### 9.1.2.2 Environment

- **Time of day:** There are several reasons why the performance of a biometric system may vary throughout the day. User behavior, clothing and physiology can actually be time dependent. For example, the voice and behavior of a person who has just woken up may differ from their voice later in the day. Environmental changes, such as lighting and temperature, can also change considerably.
- **Lighting:** The lighting at the capture location can impact the quality of a biometric image. Lighting conditions are particularly relevant to face recognition systems, as they generally perform best with frontal, uniform lighting.
- **Weather:** Temperature and humidity can impact the performance of biometric systems, in particular fingerprint-based systems. Extremely hot and humid conditions can lead to sweaty finger tips, which add noise and smudging to the prints. On the other hand, cold and dry climates can lead to dry, cracked skin. Both of these situations have been known to adversely affect fingerprint-based recognition systems.

#### 9.1.2.3 System

- **Location:** Many large-scale biometric implementations include a number of different locations where people are enrolled, verified, or identified. Each of these sites use different hardware and staff, and has unique environmental conditions. It is not uncommon to find variation between performance rates for different locations.
- **Equipment:** The equipment used to capture a biometric template or sample can influence matching performance. Some systems are designed to work with the output from a specific manufacturer. For example, an iris system may be optimized to work with images of a specific resolution, and using another camera may result in sub-optimal performance. Furthermore, faulty or dirty equipment can lead to poor quality enrollments and samples.
- **Operator:** Some systems require an operator to help users enroll, verify, or identify themselves. Variation between the operators can impact system performance. For example, a highly motivated operator may help a user achieve better results. On the other hand, an operator who neglects to clean the equipment (such as a fingerprint sensor) between presentations can be associated with poor quality enrollments.

### *9.1.3 Match Level*

A match consists of a comparison between a sample and a template. Therefore, information at this level includes all the metadata from the user(s), template(s) and sample(s). In addition to this, it contains *relational* information. For example, consider a face recognition application in which a user enrolled while wearing glasses, but occasionally wears contacts. In this case, verification performance may depend on what they are wearing when the sample is acquired. In other words, poor performance may be caused by the *difference* between the template and sample. In theory, a relationship between any attribute pair at the template/sample level may impact performance.

The most important match level metadata concerns the length of time between the enrollment template capture and sample capture. This is known as *template aging*, and affects all types of biometric systems. Biometrics do not remain perfectly stable as one ages. Changes due to aging are particularly apparent for face recognition, but can impact any form of biometric identification. When there has been a long period of time (typically several years) between two presentations of a biometric, the recognition task is considerably more difficult. It is important to quantify this performance degradation for systems that are intended for long-term use.

### 9.1.4 Attribute Notation

The following notation will be used for the remained of the chapter. Assume a user population $\mathscr{P}$, a set of enrollment templates $\mathscr{T}$, a set of samples $\mathscr{S}$ and a set of matches $\mathscr{M}$. A match $m(s,t) \in \mathscr{M}$ consists of a sample $s \in \mathscr{S}$, belonging to the user $\texttt{person}(s) \in \mathscr{P}$, matched against an enrollment template $t \in \mathscr{T}$, belonging to a user $\texttt{person}(t) \in \mathscr{P}$. Attributes of people, templates, and matches will be represented as appropriately named mapping functions. These will not all be defined formally, but rather their meaning can be inferred from their names. Here are some examples:

- $\texttt{sex}(p)$ gives the sex of person $p$
- $\texttt{age}(\texttt{person}(t))$ gives the age of the person contained in template $t$ at the time of capture
- $\texttt{score}(m)$ is the similarity score achieved by match $m$
- $\texttt{quality}(s)$ is the quality score for the sample $s$

## 9.2 System Analysis Approach

The most common approach for discovering problem groups within a system is to search for them directly. This is done by segmenting the test results into subsets representing each group of interest, and analyzing each set individually. In a sense this is a *top-down* approach, as the groups are defined at a high-level (e.g. men and women), and collective system statistics are computed for each group. The analysis conducted on each subset consists of the system level evaluation techniques presented in Chap. 7.

There are several advantages to this approach. Firstly, it is intuitive and relatively straightforward to conduct. The process is clearly defined, and the results are easy to interpret. Secondly, no special data, knowledge, or software is required. All that is required is access to the original test results, and the ability to compute performance statistics.

### *9.2.1 Splitting the Data*

The first step in the system analysis approach is to select an attribute to split on. Any of the user, template or match information from Sect. 9.1 can be used. However, the most common properties are sex, age, and ethnicity, as these represent the major subgroups for most biometric systems.

Recall from Sect. 9.1.4 that $\mathscr{M}$ designates the set of matches. The goal is to divide $\mathscr{M}$ into mutually exclusive subsets $\mathscr{M}_1, \mathscr{M}_2 ... \mathscr{M}_N$ such that $\mathscr{M}_1 \cup \mathscr{M}_2 ... \cup \mathscr{M}_N \subseteq \mathscr{M}$. Performance results are computed for each of $\mathscr{M}_1, \mathscr{M}_2 ... \mathscr{M}_N$ individually, and the results are compared to determine relative performance of the subgroups.

Assume we are interested in comparing the performance of men and women. There are three options for partitioning the test results. The samples can be filtered:

$\mathscr{M}_{m1} = \{m(s,\cdot) \in \mathscr{M} | \texttt{sex}(\texttt{person}(s)\text{=Male}\}$

the templates:

$\mathscr{M}_{m2} = \{m(\cdot,t) \in \mathscr{M} | \texttt{sex}(\texttt{person}(t)\text{=Male}\}$

or both the samples and templates can be filtered:

$\mathscr{M}_{m3} = \{m(s,t) \in \mathscr{M} | \texttt{sex}(\texttt{person}(s))\text{=Male,}\texttt{sex}(\texttt{person}(t))\text{=Male}\}$

$\mathscr{M}_{m1}$ includes matches for men against everyone, $\mathscr{M}_{m2}$ includes matches for everyone against men, and $\mathscr{M}_{m3}$ only contains matches of men against men. This raises an important question: when filtering match scores based on person or template attributes, should the filtering condition be applied to the samples, the templates, or both? Each approach will result in a different subset of results, and the most appropriate method depends on the goal of the test. In general, one should select the method that reflects how the system is intended to be used in a real world setting. Consider the following two scenarios:

1. A male criminal has fraudulently obtained hundreds of bank cards and their associated PINs. Assume that the ATMs for withdrawing cash are enabled with face recognition technology to verify that the person operating the machine is the rightful owner of the card. Assume the criminal tries each card with its PIN in the hope that he will generate a false accept and be permitted to conduct a transaction. Since the fraudster intends to try every card, his face will be matched against the true card owner's enrollment, regardless of their sex. In this case, $\mathscr{M}_{m1}$ is the correct test set as it contains the impostor distribution for "men against everyone", which reflects the scenario under consideration.
2. A passport issuing authority uses face recognition to ensure that an applicant does not already have a passport under a different name. The photo of the appli-

cant is only matched against other people of the same sex so that obvious false matches are not considered. In this case $\mathscr{M}_{m3}$ is the appropriate test set for determining male performance because only matches between men and men will be conducted.

In general, the question is an important one, and the answer will depend on the nature of the system being evaluated.

### *9.2.2 Comparing Results*

After the filtering has been completed, performance statistics are computed for each set, and the results are compared. The method of comparison depends on the type of system being evaluated. For a verification system, the most common method of comparing subgroups is to plot ROC (or DET) curves for each group on the same graph. Section 7.1.3.3 contains information about the interpretation of ROC curves, which is especially useful for comparison. For closed-set identification, CMC curves can be plotted on the same graph. In this case, the superior performance is indicated by a higher identification rate at a given rank. Rank 1 graphs are another useful method for comparing subgroups. Subgroups within open-set identification systems are typically compared using alarm curves (see Sect. 7.2.2.4).

As outlined in Sect. 7.3.3, it is important to generate confidence intervals when computing results that will be compared to each other. The reason for this is to ensure that perceived performance differences actually reflect real trends, and are not due to sampling error (i.e. random chance).

## 9.3 Data Mining

*Data mining* is the process of searching through large volumes of data in an effort to discover patterns, trends, and relationships. Data mining is an umbrella term, and refers to a wide variety of processes and algorithms for knowledge discovery. The potential value of this in the context of biometrics is obvious. In theory, these techniques can automatically uncover hidden trends within a system, allowing researchers and system integrators to identify, diagnose and correct problems.

Data mining is a broad area, and there has been little work published on its use for biometric data. Two techniques for extracting knowledge will be discussed in this section. The first is a simple statistical technique that looks for relationships between attributes and performance measures (Sect. 9.3.1). The second approach is machine learning, which automatically finds patterns and relationships in the data (Sects. 9.3.2-9.3.4).

### *9.3.1 Correlation Analysis*

Biometric systems are probabilistic by nature, due to the inherent variability of biometric samples. In other words, no two presentations of a biometric will ever be identical, so 100% certainty about a particular match is theoretically impossible. Even for powerful matching algorithms, there will be signal noise when taking digital measurements of the physical environment, leading to some uncertainty in a result. However, the key idea of this chapter is that there are some sources of variation that are intimately related to performance, and can be observed and controlled. An example of this is a *relationship* between user age and enrollment quality, where elderly people tend to have poor quality templates.

A simple approach to finding relationships between attributes and performance measures is by computing their correlation coefficient. Correlation measures the strength of the linear relationship between two variables. In other words, it measures the tendency of an attribute (often known as a predictor) to vary in the same direction as the measurement of interest. If the correlation is positive, an increase in one variable indicates a likely increase in the other variable. A negative correlation indicates the two are inversely related. For the example mentioned above, a negative correlation between age and template quality would indicate that elderly people are more likely to have poor quality enrollments than young people.

The most common method for computing the correlation of two random variables is the Pearson product-moment correlation coefficient. The input is [attribute, performance measure] pairs, and the output is the correlation coefficient, which is the strength of the linear relationship. The attribute can be any available metadata (see Sect. 9.1). In the case of categorical data (e.g. sex), each category is assigned a number (e.g. Male = 0, Female = 1). The two most common performance measures for biometrics are:

- Template quality: Many feature extraction algorithms output a quality value as a result of enrollment or acquisition. For example, an image of a smudged fingerprint would likely lead to a low quality score, while a clean image with well defined ridges would result in a high quality score. In this case, correlation analysis is used to find relationships between metadata and data capture problems.
- Match scores: A correlation with genuine (impostor) match scores may help identify groups having trouble with false rejects (accepts).

The correlation coefficient ranges from -1.0 to 1.0, with -1.0 and 1.0 indicating a perfect negative and positive linear relationship respectively (i.e. all the points lie on a straight line). A coefficient of 0.0 indicates that there is no linear relationship between the variables. Generally speaking, an absolute value below 0.3 is considered to be a small degree of correlation, and an absolute value above 0.5 is large.

**Statistically Significant Correlations**

When computing the correlation coefficient, one can also compute a *p-value*, which is a measure of statistical significance of the result. This is important because the correlation coefficient itself can be misleading. For example, if there are only two inputs there is guaranteed to be a perfect linear relationship between them (because there is a line that connects any two points). However, this does not prove that there is a real linear relationship. In this case, the p-value will be high, and the result is not considered statistically significant. However, a major drawback of the standard Pearson product-moment p-value is that it assumes both variables are normally distributed. This is not always the case. For example, when a categorical attribute such as sex is used, the p-value can be unfounded. Therefore, the recommended approach to verifying the significance of trends discovered by correlation analysis is by using the methods outlined in Sect. 9.2.2.

### *9.3.2 Machine Learning*

Correlation analysis examines individual attributes, so cannot be directly used to find trends involving two or more factors. On the other hand, the top-down approach presented in Sect. 9.2 was able to test the performance of groups with multiple attributes (e.g. sex and age). A disadvantage of both approaches is that one must conduct a separate test for each potential problem group. Therefore, one must know in advance which groups are likely to be having difficulties. This is fine when the problem group is a common demographic, such as “men” or “children”. However, consider a system where Asian females between the age of 25 and 45 are having trouble authenticating. In order to discover this knowledge using the system analysis approach, one would need to test many different permutations of ethnicity, sex, and age. Assume that the population is categorized into 2 genders, 5 ethnic groups, and 3 age ranges. In this case, there are $2 \times 5 \times 3 = 30$ demographics. With this number of groups, a test of each is feasible given sufficient time. However, Sect. 9.1 lists many factors that may have an influence on group performance. As more factors are considered in the analysis, there is a combinatorial explosion of the number of possible groups. For example, with 12 attributes that are divided into 3 categories, there are over $3^{12} > 500,000$ groups (although most will have few, if any, members). Obviously, if one wishes to discover trends and patterns in groups characterized by more than 2 or 3 attributes, the approach of testing each possible group directly is not practical.

This combinatorial explosion is a classic problem of artificial intelligence (AI). The *feature space* (all the possible combinations of the input attributes) of the problem is prohibitively large for an exhaustive search, so "intelligent" techniques must be used to search progressively smaller sub-spaces that are likely to contain a good, although not necessarily optimal, solution. A central focus of AI, under many different guises, is developing efficient search techniques for new problem domains. One approach is known as *machine learning*, which is concerned with the development of algorithms that allow computers to dynamically "learn" patterns from previously unseen input data.

The group analysis approach of Sect. 9.2 was referred to as *top-down*. The reasoning for this was that the groups were defined in advance at a high level (e.g. men and women), and the set of all match results was partitioned accordingly. However, the machine learning approach is significantly different, and can be viewed as *bottom-up*. Each record is associated with metadata and performance measures, and knowledge is built upon this foundation by building classifiers that model the data.

In general, there are two basic approaches to machine learning algorithms of interest to biometric applications:

- **Supervised learning**: For supervised learning each input has metadata and an associated label, and the goal is to generate a function that maps the input data to the label. For biometrics, the input would be metadata for users (e.g. sex and ethnicity), templates (e.g. capture location) or matches (e.g. time of day), and each input would be assigned a performance label. The performance label is supplied by a domain expert, and is the concept that is being modeled. For example, a person may be a "lamb", a template may be a "failure to enroll" and a verification transaction could be a "false accept". Alternatively, the label can be quantitative, such as a person's average genuine match score. An example for the output of supervised learning is a function that embodies a rule along the lines of "fingerprint verifications conducted in hot, humid conditions" $\mapsto$ "potential false accept". For this application, the goal of the process is not to develop a classification algorithm to predict the performance of unseen data, but rather to use the model that has been developed to label user groups according to performance.
- **Unsupervised learning**: Unlike supervised learning, which uses a test set of labeled samples, the input to the unsupervised learning problem is unlabeled. Therefore, the goal is not only to develop a model to distinguish the groups, but also to define the number and nature of the groups themselves. Due to its unrestricted nature unsupervised learning is more difficult than supervised learning. The most common approach is clustering algorithms, such as k-Means, which automatically discover homogeneous subgroups in the population, such as groups of people with similar properties. In the context of biometric data, the input would be all of the metadata and performance labels associated with a people, templates, or matches. The output would be groups of people defined by a set of common attributes.

Both supervised and unsupervised learning techniques can be applied to biometric data, and are treated individually in the following sections.

### *9.3.3 Supervised Learning*

Some common approaches to supervised learning are [3]:

- **Artificial neural networks:** Neural networks are connected groups of artificial neurons that were originally motivated by the computational framework of biological brains. The weights and connections between neurons are updated dynamically to adjust the relationship between the input and output data [5]. Neural networks have been successfully applied to a wide variety of problems.
- **Decision trees:** Decision tree algorithms create rooted trees that are used for classification. Each node of the tree (including the root) contains a branching rule concerning a specific attribute, and based on the outcome of this rule, one of the sub-nodes is chosen. This process continues until a leaf is reached, which will contain a label for the instance being classified [5]. For example, the root node might contain "sex", and it would lead to separate branches for "men" and "women". A leaf contains a label such as "wolf" or "failure to enroll".
- **Naive Bayes classifier:** The Naive Bayes classifier uses probability models based on Bayes' theorem [2]. The posterior probability that an input record belongs to a given label is calculated based on the conditional probabilities that its attribute values could be obtained for that label.[1] The classification rule is defined by selecting the label with the highest posterior probabilities.
- **Support Vector Machines:** Support Vector Machines (SVMs) are based on statistical learning theory. SVMs are a binary classifier that work by finding a hyperplane in the feature space that maximizes the margin between the plane and the instances of the two classes. By mapping from the original feature space to one with high dimensionality it is able to find discriminating functions even for complex data patterns [1].

These are some of the most common algorithms used for supervised learning. Other algorithms include nearest neighbor methods, genetic algorithms, and rule induction. In theory, any of these techniques can be applied to biometric data mining. However, decision trees tend to dominate data mining applications as they have several advantages over the other techniques. However, it should be kept in mind that decision trees represent one of many learning algorithms available, and are not necessarily the optimal choice for all situations.

#### 9.3.3.1 Decision Trees

The name "decision tree" reflects the graphical representation of the classification model, with a root, branches, and leaves. Classifications are made by following a path from the root to a leaf, making a new decision at every internal node. Figure 9.1 contains an example of a decision tree. For this example, the object of classification

[1] The classifier is called "naive" because it assumes that the input data attributes are conditionally independent.

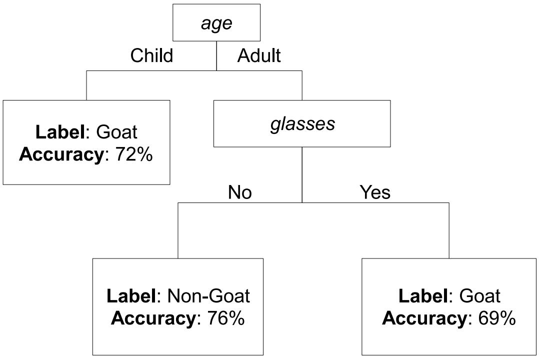

**Fig. 9.1** An example of a decision tree. In this case, 72% of children and 69% of adults who wear glasses exhibit goat-like behavior (i.e. have difficulty matching against their own enrollments).

is the user's genuine performance, and they are categorized by both demographic (age) and behavioral (glasses) attributes. The resulting tree defines two goat populations: children, and adults who wear glasses. The accuracy values on the leaves are applied to all of the people who fall into that category. For example, assume there are 200 people in the example data used to build tree who are adults and do not wear glasses. 76% of these people (152) are non-goats, while the other 48 people are goats.

There are a variety of ways to build a decision tree. The most common approach is to use an information theoretic framework, which uses the concept of entropy (or information gain) in an attempt to minimize the expected number of internal nodes necessary to classify the training set [6]. The general concept is that simple trees (those with few levels) are preferred to complex trees (those with many levels) as they are more likely to reflect true patterns in the data, and less likely to be due to overfitting the data.[2] Overfitting occurs when a model has a very low error rate for training data, but is unable to generalize to unseen data.

A significant drawback of many supervised learning algorithms is their "black box" nature. In other words, there is no intuitive interpretation for the models generated for classification, which is a common requirement for data mining applications. Many methods are able to create classification rules (e.g. person $X$ is likely to be a "worm"), however, they do not state *why* the decision was made. Decision trees are unique in that the classification decisions are easily interpreted. Other advantages of decision trees include [3]:

- Mixed data types: The input can be a mix of categorical and numerical metadata.
- Missing values: The algorithm can still generate models if some data is missing (e.g. the sex of some users is unknown).
- Robust to outliers: The presence of a few outliers will not greatly affect the tree.

[2] This principle is known in philosophical circles as Occam's Razor.

- Computationally efficient: Trees can be built quickly for large amounts of data.
- Irrelevant input: The trees are resistant to the presence of metadata that is not related to performance. For example, assume a fingerprint system in which ethnicity is not a relevant predictor. In this case, a tree will be built that does not include any "ethnicity" nodes. However, it should be kept in mind that the trees are resistant, not immune, to irrelevant input.

The primary disadvantage of decision trees is that they are not as powerful as some methods, such as neural networks and SVMs. However, the requirement of being able the interpret the models is important, so they remain a strong option for biometric data mining.

There is a large body of publications that deal with building decision trees. A thorough review is beyond the scope of this text, however there are two key concepts that are worth mentioning. Both concepts are related to *overfitting* the data, which occurs when trees are built that reflect idiosyncrasies of the training data, rather than general patterns within the population. These trees usually perform very well on the testing data, but do not generalize to the population as a whole. Firstly, as mentioned above, simple trees are preferable to complex trees. Large trees with many levels are usually the result of overfitting. Therefore, it is important to keep the tree relatively shallow. A technique known as *pruning* is sometimes used to "clip" branches to restrict the depth of the trees. Secondly, another technique to avoid overfitting is to randomly partition the input examples into a number of different sets, and build a tree for each set independently. These trees are then merged using an averaging technique known as *bagging* [3].

Many data mining toolkits have (such as Weka [7]) come with algorithms for pruning, bagging, and other techniques to avoid overfitting. It is strongly recommended that these techniques are used when building decision trees for biometric data mining.

### 9.3.3.2 Problem Formulation

There are four basic steps for the application of decision trees to biometric data. The first three of the steps are concerned solely with deciding what data will be used for the input and output of the learning task. Obviously, this is restricted by the data that is available. However, it also depends heavily on the goal of the test. The final step is running the algorithm and examining the results.

In general, it is important to have a well defined objective for a data mining task. Blindly including all information available and hoping the algorithm will sort it out is not a good strategy. When used properly learning algorithms can be very powerful, yet they are useless when used naively. One should always keep in mind the adage "garbage in, garbage out".

*Step 1 - Subject:* The first step is to decide the subject of classification (the first column of Table 9.1). The most common subjects will be the users of the biometric systems. In this case, the goal of data mining is to detect groups of users with

| Subject of classification | Examples of input metadata | Possible labels |
|---|---|---|
| Templates | Glasses, contact lenses, etc. | False accept, false reject |
| Users | Sex, age, ethnicity | Lamb, goat, wolf |
| Location | Demographics, environment, etc. | Frequency of failure to enroll (FTE) and failure to acquire (FTA) |

**Table 9.1** Examples for the input data for the supervised learning problem. The metadata is information that may be relevant to the subject performance, and the labels are examples of categories that are applicable to the groups. For example, by using the information in the first row, one may be able to detect the properties of templates that are leading to verification errors (e.g. enrollments with glasses may lead to false accepts).

a property in common. However, learning can also be applied to other entities, depending on the system under evaluation. In general, the more complex the system, the more data available, and the greater the need to apply intelligent techniques to untangle patterns in the data. For example, individual templates can be used as the subject for classification. This would enable the discovery of trends specific to an instant in time, such as enrollment while wearing glasses. At a higher level, physical locations may be the subject of classification. A large-scale biometric system may include dozens of sites worldwide where enrollments are captured, and one may wish to know what role location plays in performance.

*Step 2 - Attributes:* The second step is to decide on the properties that will be used to describe the input (the second column of Table 9.1). These properties are referred to as metadata, attributes, or predictors. Obviously, the choice of attributes depends on the subject of classification chosen in the first step. The two guiding principles are to select a) the properties relevant to the subject of classification, and b) only the properties that are likely to impact performance. For example, environmental variables such as lighting may be important considerations for a face recognition location, but will have little impact on performance within a fingerprint system.

*Step 3 - Labels:* The third step is to define and assign the labels for the input data (the third column of Table 9.1). In essence, this step defines the overall goal of classification, and depends on the nature of the subjects. For users of biometric systems, the most common labels will be related to performance. For example, all animals from the biometric menagerie (see Chap. 8) are potential labels. These animals embody concepts like “trouble matching against their own enrollments” (goats), or “high match scores against everyone” (chameleons). The same labels can be applied to other subjects of classifications as well. For instance, animal names can be applied to locations, where a “goat” is a location where there is an unusual number of false rejections. Similarly, the animals are applicable at the template level. A “lamb” template may be an enrollment that often ranks highly against others in an identification system.

Another performance measure that may be of interest is the likelihood of enrollment and acquisition errors. Once again, these are relevant at all levels: templates, users, and locations.

The best strategy for deciding on which labels to use is to find and label two groups that have opposing meanings. For example, assume we are interested in finding people who exhibit lamb-like behavior for a surveillance system. In this case, two groups would be found: those whose average impostor rank is high (lambs) and those with low ranks (non-lambs). Samples from each group would be selected and labeled, and used as the input to the learning algorithm. The output of the algorithm would be a model that characterizes the user groups. In general, it is best to have two equal sized groups.

*Step 4 - Learning:* The final step is to apply the learning algorithm to generate the classification model. If a decision tree algorithm is used, each path from the root to a leaf defines a group, and associates it with a label. For each group found, one must verify that it is statistically significant (see Sect. 9.2.2).

### *9.3.4 Unsupervised Learning*

For unsupervised learning, the input data is unlabeled, so there is no direct target for classification. Therefore, the learning process is open-ended, and there is little control over the nature of the groups found. The most common approach to unsupervised learning is based on *clustering*, which seeks to find groups that are "close" in the feature space (i.e. have a lot of properties in common). For example, applying clustering to raw biometric metadata may uncover a group of young Asian males. This indicates that they make up a distinct subgrounp of the user population, but makes no implication about the performance of the group.

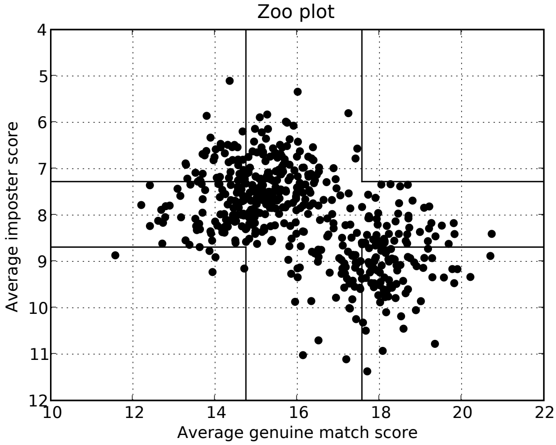

**Fig. 9.2** A zoo plot with two clusters: the upper left group has low genuine and impostor scores (phantoms), and the lower right has high genuine and impostor scores (chameleons).

The problem must be formulated in a manner that emphasizes the role of performance. Instead of clustering based on metadata, clustering can be performed on match scores. For example, consider the zoo plot of Fig. 9.2. Recall that a zoo plot (see Sect. 8.2.3) plots each user based on their average genuine and impostor match score. The y-axis is reversed, so people in the upper right are performing well (high genuine and low impostor scores), and people in the bottom left are performing poorly (low genuine and high impostor scores). There appears to be two distinct clusters in Fig. 9.2, one with phantom properties (upper left) and one with chameleon properties (lower right). In theory, a clustering algorithm may be able to automatically detect these two clusters. Having found the groups, one would examine their metadata to see what properties the groups have in common. For example, it may be observed that the majority of the group in the upper left are males.

There are limitations to the clustering approach. First of all, clustering techniques work best when it is known a priori how many clusters are expected. However, in most cases this information is unknown. Secondly, groups that are nicely behaved (e.g. well separated, symmetric, and normally distributed) are easiest to detect, but real data is rarely so cooperative. Finally, in two dimensions no existing automated clustering algorithms are able to compete with the performance of the human visual system. In most cases, a quick glance at a zoo plot is more fruitful than the extensive application of unsupervised learning algorithms.

## 9.4 Dealing with Problem Groups

The techniques outlined in the previous sections of this chapter can be used to find problem groups within a biometric system. Typically, these will be groups of users, but can also be groups of templates (e.g. those enrolled while wearing glasses) or locations (e.g. enrollment stations *X* and *Y*). For many evaluations, the next step will be to find ways to minimize the impact of known problems.

In order to address a problem, one must first find the underlying cause. This is largely a manual process, and involves examining the data from different viewpoints. For example, assume that it has been determined that women are performing poorly in a system. In general, there are two potential underlying causes: physiological differences between men and women, and behavioral differences between men and women. In the first case, it is an inherent difference between the sexes that is causing the performance discrepancy. For example, consider the case of a speaker verification system that is better at distinguishing low pitch voices than those with a high pitch. In this case, men will tend to perform better. In the second case, a behavioral difference between men and women is impacting the ability of the system to perform a correct match. For example, consider a face recognition system that is based on skin texture analysis. In this case, makeup can cause performance problems for the recognition algorithm, impacting women more than men.

### *9.4.1 Physiological Problems*

Problems due to between-group physiological differences are the most difficult to deal with, as they reflect weaknesses in the underlying biometric algorithm. For algorithm developers, this information is useful as it highlights areas in need of further research, and will ultimately lead to stronger algorithms. However, in many cases, the person conducting an evaluation has no access to the source code of the verification algorithms. Therefore, it may not be possible to address the root cause of the problem. The following are some possible approaches for mitigation:

- In some cases, the algorithmic adjustments will be relatively straight-forward. For example, the recognition model may be tuned using learning algorithms that can be re-trained using more appropriate test data. In some cases vendors may provide this service to their clients.
- A system policy that requires users to re-enroll periodically can be beneficial. For example, consider a system that struggles with the physiological problem of template aging. In this case, re-enrollments will reduce the problem significantly.
- Group-specific thresholds can be used to address some problems. For example, assume the children users of a system are found to be consistently receiving low genuine match scores. In this case, a lower match threshold for people younger than a certain age can be defined, reducing the risk of false rejects. However, one must be careful when implementing group-specific thresholds. In general, one must be aware of the trade-off between false rejects and false accepts, and be careful not to introduce new problems when attempting to fix old ones.
- Pre-selection algorithms can be used to avoid troublesome matches altogether. For example, if there is a high occurrence of false matches between left-loop and right-loop fingerprints, the problem can be addressed by only matching prints of the same class. Similarly, user data such as sex and age can be used to filter matches in identification systems (these are sometimes known as *soft biometrics*). While it is rarely possible to avoid the problem completely, quite often the impact can be reduced significantly.
- In some cases, a physiologically based problem may be so fundamental that new technology will be required. This may include: upgrading to the latest version of an existing engine, switching vendors, switching biometrics (e.g. from face to iris) or combining multiple biometrics. Combining multiple biometrics is known as *multimodal biometrics*, and is a powerful approach when faced with fundamental limitations of a single biometric modality (see Chap. 4).

### *9.4.2 Behavioral Problems*

In general, problems arising from behavioral factors are easier to deal with than problems due to physiology. In most biometric systems there is some degree of control over the behavior of users, and careful procedures can be established to address the major issues found. For example, assume it has been determined that templates enrolled while the user is wearing glasses are vulnerable to false accepts by other people wearing similar glasses. A system policy that requires users to remove glasses during enrollment will eradicate the problem. In general, providing training to the users and operators of biometric systems will help lead to high quality enrollments and acquisitions, considerably reducing system errors.

One case in which there is little direct control over user behavior is covert surveillance systems (see Chap. 11). In this case, the user is not aware, and not meant to be aware, of their participation. However, even in this case there are often subtle techniques that can be used to alter behavior without the user's knowledge. For example, consider a covert face recognition system. Face recognition generally requires frontal images of subjects. However, when people are walking in everyday situations, their gaze typically wanders unpredictably around the environment. An "attractor" can be used to draw the attention of people walking through the capture zone. This is a device that has been designed to attract attention, while still looking like it naturally belongs to the environment. An example would be a brightly lit sign containing a strongly worded warning message. In this case, the surveillance cameras can be hidden behind one-way mirrors adjacent to the sign, increasing the chance of a frontal capture.

### *9.4.3 Environmental and Equipment Problems*

There is often considerable control over the physical environment where biometrics are acquired. For example, a common problem for face recognition is lighting, and re-configuring the lighting in a room to ensure uniform, frontal lighting is a relatively straightforward task. Furthermore, the equipment used for capturing biometrics is often tightly linked to performance, and regular maintenance or replacement will lead to performance improvements.

As with behavioral problems, there are some situations in which there is limited control over environmental and equipment. For example, consider a situation where a phone-line is taped by investigators, and speaker verification is being used to identify the person talking. In this case, there is no control over the type of phone being used, and the environment in which the conversation is taking place. Similarly, consider biometric identification used at crime scenes for forensic purposes. In these situations, one must accept the limitations of the biometric data available, and rely on the development of robust techniques for pre-processing and feature extraction for accurate matching.

## 9.5 Limitations of Group-Level Analysis

In general, there is little research published in the scientific literature on group-level analysis for biometrics. There are a few reasons for this, some practical, and others theoretical. The researchers most likely to publish their results and findings in this area are academic. However, these groups tend to have limited access to the biometric data necessary for the task. Gathering the data from volunteer participants is time-consuming and costly. Institutions that already have access to large amounts of labeled biometric data (e.g. driver's license authorities) are unlikely to make the data available to researchers due to privacy concerns. Biometric data is inherently sensitive as it contains an indelible link to true identities.

A related factor is not just the difficulty involved in obtaining the data, but the large volume that is necessary for the analysis. The techniques of Sect. 9.2 can be used to find problems among common subgroups, such as those based on sex or ethnicity. On the other hand, the data mining techniques of Sect. 9.3 have the potential to find more subtle and complex trends among the data. However, the more subtle the trend, the more training data that is necessary to find it.

Another issue concerns the nature of the metadata attributes (see Sect. 9.1). Often the relevant attributes are difficult to measure. For example, a user's mood can be closely linked to their performance. For instance, simply being "in a hurry" can have a negative impact on performance. In other cases, such as criminal identification, users may be deliberately and actively uncooperative. However, factors such as these are difficult to quantify. They are inherently subjective, and obtaining appropriately labeled data poses significant challenges in its own right.

With all data mining techniques, there is always a risk of "over-fitting" the data. Supervised learning algorithms will almost always output *something*, even when applied to random data. One must not fall into the trap of "data fishing", where non-significant patterns are uncovered and reported. Any problem groups discovered must be verified using statistical techniques, such has hypothesis testing.

Finally, there is a fundamental limitation in that learning algorithms can only discover trends that are apparent in the metadata. However, the underlying causes for problems may not be reflected in the attributes available. In many cases, the reason why an individual performs poorly is too abstract to be predicted using the attributes such as "Caucasian" and "male". For example, consider a face recognition algorithm that performs poorly for people with a crooked nose - information about nose morphology is rarely embodied in the metadata. Ethnic labels may be useful to some degree, but certainly do not capture all inter-group variability. Therefore, underlying causes such as this are beyond the grasp of automated group level analysis, and one must rely on a visual inspections and analysis to identify sets of problem users.

## 9.6 Conclusion

A central theme of this book is that biometric data is complex, and performance can be evaluated at a number of different levels. This has motivated a hierarchical approach to analysis. At the highest level is system analysis, which has traditionally received the most attention, and thus reached the highest level of maturity. At the lowest level are the individual users of the system. Individual users have varying performance, and the biometric menagerie of the previous chapter is becoming recognized as an important tool for finding and characterizing these users. However, the middle layer of the pyramid, the fuzzy region between users and systems, still receives little attention. This is where groups of people, images, or locations follow patterns and trends. Unfortunately, the patterns and trends are often hidden, and require some work to identify. This chapter introduced the concept of using data mining techniques for biometric knowledge discovery. This mode of analysis is still young, and requires more research to determine the most appropriate techniques. However, due to the large volumes of data generated by biometric systems, and the importance of thorough analysis, automated and intelligent techniques will undoubtedly play an important role in the future of biometric data analysis.

## *References*

[1] Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**, 121–167 (1998)

[2] Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley & Sons, Inc. (2001)

[3] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)

[4] ISO: Information technology – biometric performance testing and reporting – part 1: Principles and framework (ISO/IEC 19795-1:2006) (2006)

[5] Mitchell, T.: Machine Learning. McGraw-Hill Companies Inc. (1997)

[6] Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (1980)

[7] WEKA: Weka data mining software, The University of Waikato. `http://www.cs.waikato.ac.nz/ml/weka/` (2008)