7

Biometric Data Mining Applied to On-line Recognition Systems

José Alberto Hernández-Aguilar[1], Crispin Zavala[1], Ocotlán Díaz[1], Gennadiy Burlak[2], Alberto Ochoa[3] and Julio César Ponce[4]
[1]*FCAeI-UAEM,*
[2]*CIICAp. Universidad Autónoma del Estado de Morelos*
[3]*Universidad Autónoma de Ciudad Juárez*
[4]*Universidad Autónoma de Aguascalientes*
[1,2,3,4]*México*

1. Introduction

Data mining has become an increasingly popular activity in all areas of research, from business to science, biometrics being no exception. **Data mining** is the computer-intensive activity of exploring large data sets with the purpose of discovering, within a subset of data, some relationship of patterns or hypothesis that may be worthy of further study (Hernández-Aguilar et al., 2008; Amaratunga & Cabrera, 2004). According to a widely accepted definition, knowledge discovery in databases (KDD), more widely known as data mining, is a non-trivial process of identifying valid, novel, potentially useful and understandable patterns in data (Fayyad et al., 1996).

1.1 Basic definitions

But, what is biometric data mining? What does it study? First of all, let us clarify the meaning of biometric. According to Dunstone & Yager, 2009, there is a considerable amount of inconsistency among the terminology used within the biometric research and industrial communities. The best effort to date is ISO/IEC 17975-1, Information technology – Biometric performance testing and reporting. The following definitions of biometric and biometrics are consistent with this document:

Biometrics is the automatic identification of an individual based on his or her physiological or behavioural characteristics.

Biometric: A measure of a biological or behavioural characteristic used for recognition. There are four requirements for a biometric attribute: every person must have it, it should be sufficiently different for every person, it should remain constant over time, and it must measurable quantitatively.

Let us now define **Biometric data mining (BDM)**. BDM is the application of knowledge discovery techniques to biometric information with the purpose to identify underlying patterns. A principal objective of many data mining problems in biometrics research is to uncover characteristics of subsets of cases that are substantially different from the rest of the cases (Amaratunga & Cabrera, 2004). Consider the following:

Case study 1. Structure activity databases in the pharmaceutical industry are datasets prepared with the objective of studying the relationship between the biological activities of a series of compounds and their chemical properties. The primary goal is to identify ranges of values of x=(x1,...,xn) associated with higher likelihood of in vivo activity.
Case study 2. An epidemiological database of several women of Indian Heritage, collected by the US National Institute of Diabetes and Digestive Kidney Diseases, studied the relationship between the incidence of diabetes among them using several predictors such age, plasma concentration level, serum insulin level, diastolic blood pressure and body mass index (Blake and Merz, 1998). The objective was to identify the characteristics of the subjects associated with high incidence of diabetes.
Case Study 3. The selection of a small panel of proteins from the thousands of m/z points in mass-spectra, by means of selected variables, and their respective roles and interactions with the purpose to generate a classifier, that makes biological sense, for cancer diagnosis (Hilario et al., 2004).
Case Study 4. Several sports use anthropometric (biometric) analysis based on Data Mining such as Water-polo, Fencing, Synchronized Diving, Tumbling, Synchronized Gymnastics, Badminton, Archery and Curling (Ochoa et al., 2009). A study on people of Asian-Canadian descent has been performed in Quebec regarding their practice of Curling. The results have shown that the particular anthropometric characteristics of equilibrium, stamina and speed of this ethnic group minimize its possibility to play Ice Hockey, but are of high value to Curling.
Case Study 5. In (Ochoa et al., 2009 the biometric characteristics that Professional Wrestler Idols must have are discussed. Study reveals that age, height and weight always play a very significant role in wrestling, which should of course not be surprising. Besides, hidden patterns found by intelligent agents are related to the size of circuit, match records and cultural distances (ethnicity), as well as the selection of a good wrestler with specific attributes. A Wrestler with features similar to Scott Steel was selected as the most popular by the majority of societies because the attributes he has are adequate for other wrestlers.
In the aforementioned cases the relationship between a response variable *Y* and a set of predictor variables is studied. As shown in figure 1, the heart of the process is the learning step or automatic construction of a predictive model by generalizing from the training data. Here is where data mining finds applicability.

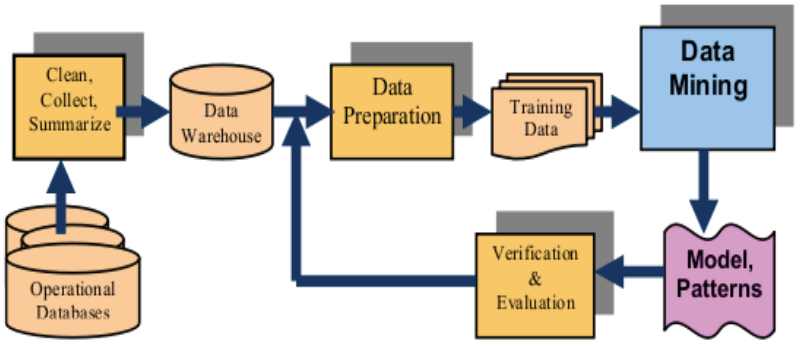

Fig. 1. The KDD (Knowledge Discovery Process) and data mining process (Han & Kamber, 2002)

Data mining is the process of searching through a large volume 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 these techniques, applied to biometrics, is that it can automatically uncover hidden trends within a system, allowing researchers and system integrators to identify, diagnose and correct problems (Dunstone & Yager, 2009).
For example: Identifying a negative correlation between age and template quality would indicate that elderly people are more likely to have poor quality enrolments than young people. Another example is using a decision tree to classify specific behaviour of users during the enrolment process, see Figure 2. The resulting tree defines two goat populations (i.e. having difficulty matching against their own enrolments): children and adults who wore glasses.

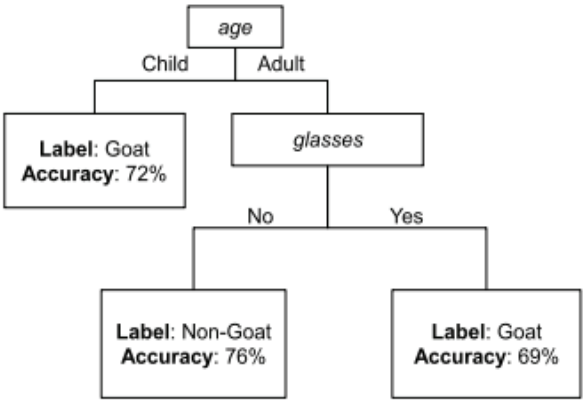

Fig. 2. In this case, 72% of children and 69% of adults who wear glasses exhibit goat-like behaviour (Dunstone & Yager, 2009).

According to (Van Der Ploeg, 2005), the information of the body answers the questions: who you are, how you are, and how you are going to be treated in various situations. - All this is information that is going to be processed through the networks (i.e. the Internet), databases, and algorithms of the information society. This author suggests that once translations of bodily characteristics are transformed into electronic data, they can become amenable to forms of analysis and categorization in ways not possible before. She suggests three levels of categorization:

- Authentication (1:1 comparisons). Classification of people as legitimate and truthful or illegitimate or fraudulent.
- Identification (1:n). Classifies and categorizes people according to the type and purpose of the database against which the biometric signal is checked. People may be identified as: someone with a criminal record, an illegal immigrant, a bad credit risk subject, or an employee with legitimate access to a high security facility.
- Aggregate level. Bringing together biometric information with other types of information, to generate, through a process of cross matching and data mining, more or less detailed and specific profiles that will be used to predict behaviour, assess dangerousness, or label as a risk.

Level three is used for preventing terrorist attacks, whereby suspicious electronic patterns of behaviour are tracked and recorded. According to (Neil, 2008) in the U.S., computer-driven searches look through an individual's web history, credit card transactions and personal background, allowing authorities to flag suspect behaviour. This author cites the example of a Muslim chemistry graduate who takes an ill-paid job at a farm-supplies store. What does this signify? Is he merely earning some cash, or using the job as cover to get his hands on a supply of potassium nitrate (used in fertiliser, and explosives)? What if his credit-card records show purchases of timing devices? Data mining allows analysts access to this information, but it is left to their judgement to decide whether or not it constitutes the beginning of a criminal plot, or just some innocent individual's "eccentric but legal behaviour".

1.2 Advantages and disadvantages of BDM

Main **advantages** of BDM include the following: most of the current authentication systems include a type of biometric measure like photo id, fingerprints and even iris recognition systems. Biometric characteristics are unique for each person. The cost of computers and biometric devices is decreasing and computer power is growing exponentially. The cost of building and operation of large data bases is decreasing. Every year more and more devices and systems are designed to support biometric characteristics. It is now possible to use several biometric characteristics (multi biometry) to support authentication. Most modern computer systems allow biometric identification upon logging in.

Disadvantages. A major disadvantage we can name is definitely the privacy concern for inappropriate use of biometric information in remote authentication settings, where biometric measurements are collected at remote sites and some weak points certainly exist. Dishonest entities such as servers that impersonate a user or perform data mining to gather information could be the source of successful attacks. Furthermore, the communication channel could also be compromised and biometric data could be intercepted or modified (Vetro et al., 2009).

To perform proper BDM, large data sets are mandatory. According to (Turban et al., 2004) to perform proper data mining a clear understanding of the problem is required. Identifying trends and patterns inside large volumes of data is a time consuming activity. Technical knowledge about biometric devices and systems is necessary. Performance of biometric devices can be reduced by dust or unintentionally by user misuse.

2. Application of biometric data mining to on-line recognition systems

Humans are believed to have several unique characteristics such as fingerprints, hand geometry, and iris (Tapiador & Singuenza, 2005), and nowadays almost all computerized systems involve an identity authentication process before a user can access requested services (Iglesias et al., 2008); for instance : secure access to buildings, logging into a computer system, laptop or cell phone, or logging in with the purpose of making e-commerce transactions on ATM machines, or to use on-line assessment systems (Hernandez-Aguilar et al., 2010). The applications of biometric identification range from forensics and law enforcements to novel biometrics-based access to personal information that protects user privacy and mitigates fraud. Biometric systems recognize users based on two types of characteristics: 1) behavioural (i.e. voice, signature, keystroke dynamics or haptics) or 2) physiological characteristics (i.e. fingerprints, iris pattern, face image or hand geometry). We will concentrate our analysis on

those technologies that make applying biometric data mining to on-line recognition systems possible. In Table 1 a summary of these techniques and their reported overall accuracy is shown.

Technology	Overall accuracy	Reported on
Behavioural		
Keystroke patterns/dynamics	95 %	(Revett et al. 2005)
Haptics and virtual reality	95 – 97%	(Iglesias et al. 2008)
Mouse movements	95.00%	(Kaklauskas et al. 2008)
Stylometry	82.00% or less	(Shalhoub et al. 2010)
On-line behaviour	56.0%-98.0%	(Yan & Padmanabhan, 2010)
Physiological		
Fingerprint	98.00 - 100%	(Umamaheswari et al. 2007)
Face	90.00%	(Lovell & Chen, 2008)

Table 1. State of the art Biometric data mining technologies for on-line recognition systems

Converting physiological characteristics of users into techniques for biometric identification has been an active research for several years, and combined with data mining techniques, produces biometric data mining, a hot topic of discussion in the scientific community.

2.1 Keystroke patterns/keystroke dynamics

Research focused on Keystroke patterns, in terms of Keyboard Duration and keyboard latency. Evidence from (Revett et al., 2005) indicated that when two individuals entered the same login details, their typing patterns would be sufficiently unique as to provide a characteristic signature that could be used to differentiate one from another. Keystroke dynamics is a cost effective means of enhancing computer access security, and has been successfully employed as a means of identifying legitimate/illegitimate login attempts based on typing style of the login entry.

2.2 Biometric behavior through haptics and virtual reality

Haptics refers to the science of sensing and manipulating through touch in real and virtual environments. Haptics technology allows users to interact via sense of touch by applying forces, vibrations and/or motions to users. Examples are vibrating phones, gaming controllers, force-feedback control knobs in cars and the wiimote controller. Data directly generated by the user that interacts with the system is recorded and used for authentication purposes. Therefore, Haptics can be seen as a mechanism to extract behavioural features that characterize a biometric profile for authentication. (Iglesias et al., 2008) applied non-linear transformations to the original feature space to produce Euclidean 3D spaces preserving the similarity structure of the samples, which were represented with Virtual Reality (VR) techniques. These new spaces were analysed using visual data mining to know how certain features (i.e. position, pressure and torque) contain more meaningful information that can characterize a biometric profile when signing in.

2.3 Mouse movements

User identification using mouse movement parameters has been discussed by different researchers (Eusebi et al., 2008; Brodley & Pusara, 2004; Weiss et al., 2007). In (Kaklauskas et al., 2008) a Web-based biometric Mouse decision support system for user's emotional and labour productivity is discussed, and reported as able to analyse data from a biometric mouse – designed for the same authors- and e-self reports. They mixed different biometric parameters, including physiological (skin conductance, amplitude of hand trembles, and skin temperature), physiological (self-reports) and behavioural (mouse pressure, speed of mouse pointer movement, acceleration of mouse, etc.) and made a correlation between the user's emotional state and labour productivity. The possibilities of the biometric mouse are remarkable; it is able to measure the temperature and humidity of a user's palm and his/her intensity of pressing. These parameters could be used to identify suspicious behaviour and single out impostors.

2.4 Stylometry

Stylometry is a discipline that determines authorship of literary works through the use of statistical analysis and machine learning. When someone authors a literary work, document, or email they leave behind certain attributes to their writing style that can be analysed and used to determine other works by the same author. The rise of the Internet has opened new uses for stylometry in the area of e-mail, social networking and various types of digital content. In (Shalhoub et al., 2010) a study to identify if some stylometric tools (i.e. C# tool testing) can correctly assign authorship of electronic mail to its original author is presented, and has produced results which indicate moderate accuracy - suggesting that none of the tools evaluated is capable of correct author identification.

2.5 On-line behavior

Research in biometrics suggests that the time period a specific trait is monitored over (i.e. observing speech or hand writing long enough) is useful for identification, most notably the research in Web usage mining (Adomavicius & Tuzhilin, 2001; Pirolli, 2007) which suggests that user behaviour is not random and there is often a purpose that translates into revealed on-line behaviour. In (Yang & Padmanabhan, 2010) a data mining analysis of the effect of observation time period on user identification based on on-line behaviour is presented, and the identification of unique behavioural characteristics that can possibly serve as identifiers is discussed. The quality of data can be measured by the features created from his/her behaviour. For user identification from on-line behaviour, quantity is a measure of how much user data is observed. In this research the authors use aggregation to describe the processes of observing and collecting data over long periods of time. Specifically they use aggregation over multiple web sessions. Results suggest that at the user-centric level it is possible to build reasonably accurate models identifying the user by observing enough data, at least for some users.

2.6 Fingerprint

Fingerprinting is the first biometric science used worldwide for the validation and verification of an entry into specific tasks and is one of the most popular techniques to perform biometric recognition. However, Fingerprint classification and recognition is still an open and very challenging problem in real world applications. In (Umamaheswari et al.,

2007) fingerprint classification and recognition using data mining techniques is discussed, and the proposed method involves various stages like image enhancement, line detector based feature extraction and neuronal network classification using Learning Vector Quantization and Kohonen networks. Optimization of neuronal parameters and recognition of images is done by a genetic algorithm K nearest Neighbour. The exact image is recognized from the classified database using a crisp and fuzzy K Nearest Neighbour algorithm. The resulting system is one of the most reliable methods of personal verification (98 to 100%), and can be used for authority access verification, ATM verification and other civilian applications.

2.7 Face recognition

According to (Lovell & Chen, 2010), there are several applications of data mining for face recognition: 1) Person recognition and location services on a planetary wide sensor net, 2) Recognizing faces in a crowd from video surveillance, 3) Searching for video or images of selected persons in multimedia databases, 4) Forensic examination of multiple video streams to detect movements of certain persons, 5) Automatic annotation and labelling of video streams to provide added value for digital interactive television. All of these applications are subject of intensive research around the world and require on-line processing.

2.8 Architecture of BDM systems

The underlying architecture for a Biometric Data Mining System consists of a client-server application (Burlak et al., 2005). On the server side resides the biometric information derived from in site or on-line enrolments, as well as all of the information and algorithms required to perform the data mining process. What's worth noting is that this information and algorithms can be requested as a web service, but if this approach is used then computational cost will increase. On the client side there may be important practical applications that might mitigate on-line fraud and identity theft, along with client-side software from a trusted third party that will track client-side activities to build users identification models (Yang & Padmanabhan, 2010). Such models may be used to provide behavioural authentication services on behalf of the user. For instance, when the user makes a large on-line brokerage transaction, the financial institution may, in real time, request the client-side software for a user score, and if the user is identified as who he claims to be, the firm may proceed with the transaction. The challenge consists on designing a system with accurate user identification models. This requires a deeper understanding of the factors that can result in better or worse identification accuracy.

3. Privacy concern

Some research on Internet privacy has examined various aspects of privacy regulation and user privacy concerns. Nowadays, the Internet has heightened a variety of users' concerns regarding privacy. The concept of privacy is "the claim for individuals to determine when, how and to what extent information about themselves can be communicated to others" (Westin, 1967). Other concepts refer to the degree to which a website is safe and user information is protected. This dimension holds an important position. Users perceive significant risks in the virtual environment of e-services and e-commerce stemming from the possibility of improper use of their financial and personal data. There is good reason

for the Anglo-American public to be resistant to national identification cards. Yet the British and American governments seem increasingly willing to neglect privacy in pursuit of personal data (Hansen, 2009). According to (Lovell & Chen, 2008) Privacy concerns that have hindered public acceptance of these technologies in the past are now yielding to society's need for increased security while maintaining a free society.

3.1 Perception of on-line users

Information security is increasingly recognized as a vital element for ensuring wide participation in Internet use. The successful use of the Internet in the Society depends on trust and confidence in our information infrastructures. Within this context, the real effect of the security problems is the inhibition of the development of the Internet use and of e-commerce as a whole.

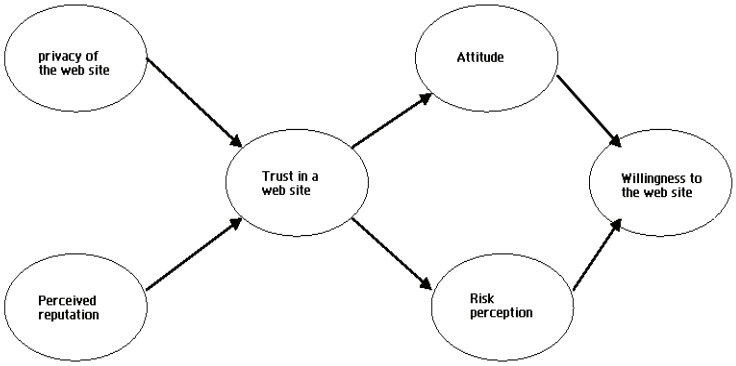

Fig. 3. Model of trust in a web site by the users

There are three variables that influence people's perception of Internet safety:

- The involvement level with respect to the category of products/services sold by the company.
- Demographics (age, gender, education level).
- Use of the Internet (frequency, familiarity).

Reputation is defined as the belief in that the web site is honest and concerned about its users. Users of Internet can favour sites that represent merchants or organizations with which they are familiar from traditional channels. This is due to the fact that this familiarity increases any positive affect as well as any positive cognition on the part of the user (Jarvenpaa et al., 2000).

Specifically, the user's preoccupation or concern with privacy includes the proliferation of databases, the great volumes of personal data being collected, and the possibility of privacy violations and loss of control in the process of collecting, accessing, and utilizing this information (Hiller & Cohen, 2001). Generally speaking, the more standardized the biometric technology, the more interoperability between different systems and databases is attained, and the more ubiquitous and pervasive the categorization of people can become (Van der Ploeg, 2005).

In fact, the rapid progress in the development of communication technologies, biometrics, sensor technologies and data storage and analysis capabilities are perceived as causing

constant pressure on the fundamental right to privacy for both economic and security reasons (Pavone & Pereira, 2009).
The rapid technology changes, accelerated general acceptance of the Internet, Social Networks, E-Commerce, and the development of more sophisticated methods of collecting, analysing, and using personal information have made privacy a major socio-political issue in a lot of countries.
In many social networking sites, users are responsible for deciding what information to disclose and whether or not to protect any of that information with privacy settings. From the time they join the community, users are challenged to create a mental model of their on-line audience and desired levels of privacy, and then determine how to best match the disclosures and accessibility of their personal information (Strater & Lipford, 2008). Some times the users have personal experience of privacy intrusions, usually in the form of unwanted contact from an unknown person.
Security as well as privacy is a need of society and its members. Designing security technologies without keeping privacy requirements in mind may result in systems which create additional risks to society and where side-effects are difficult to control.
Privacy issues have increasingly attracted the attention of the media, politicians, government agencies, businesses, etc. In addition, the public has become increasingly sensitised to the protection of their personal information.
The perception may also be due, at least in part, to the level of familiarity associated with this type of sites, and technology use. Only when the users have noticeable and disturbing events, such as a privacy intrusion, do users modify the privacy level. Based on the perceived intrusiveness many users use the privacy controls of that feature. One challenge is that users learn what to disclose and what to protect over time, both through the social norms of the community and through their own experiences.

4. Prototype BDM system applied for on-line assessments

Virtual proctoring involves using biometric technology to monitor students at remote locations. For virtual proctoring, using a layered approach depending on critical maturity of the test is recommended. With high stakes tests, video monitoring and a biometric measure such as iris scanning may be used. For medium stakes tests, a single biometrics measure may be acceptable (BSU, 2006). Despite most on-line assessments being located in the middle of both definitions, we consider the fact of high levels of cheating in remote assessments. On one hand, fingerprint recognition is a single biometric measure, the cheapest, fastest, most convenient and most reliable way to identify someone. And the tendency, due to scale, easiness and the existing foundation, is that the use of fingerprint recognition will only increase. Cars, cell phones, PDAs, personal computers and dozens of products and devices are using fingerprint recognition more and more (Tapiador & Singuenza, 2005). One current trend is to incorporate fingerprint scanners into personal computers, laptops, and mice. In addition, computer networks and large databases can be secured using fingerprint technology. This is a hot topic of discussion since the phenomenon of the Internet and the development of Intra nets has spawned new digital technologies such as E-commerce and online services. Besides, users are more willing to use fingerprint recognition than iris recognition, as they believe it is safer, health-wise. Unfortunately, fingerprint recognition is used merely for authentication, and then what? The student is free to use any media to cheat on the exam. To avoid that situation we considered the

possibility of using web cams. Web cams are inexpensive and most of the students are used to dealing with them, they are part of their common tools for work and chat. Naturally, some students reject the possibility to be monitored, and the percentage varies from country to country, but it is our intention to measure this figure as a part of our research. Based on the aforementioned, we propose the mixed use of video monitoring, by means of web cams, and fingerprint recognition to provide a secure on-line assessment environment.

4.1 Technical requirements

The Server Side

- Keep information of biometrics measures (fingerprints) and associated student information in data base.
- Scanning of finger prints (enrolment of students).
- Provide a recognition tool to determine validity of fingerprint and grant authorization to on-line assessment.
- Monitor remote students by means of web cams located in remote locations.
- Support the on-line assessments process.
- Provide security mechanisms to ensure confidentiality and validity of data: Encryption of data transmitted and received and log files.

The Client Side

- Scanning of finger prints.
- Enrolment of students (optional).
- Avoid unauthorized access to on-line assessment.
- Show the diagnosis of security.
- Provide capacity of students' monitoring using web cams during assessment process.
- Provide mechanisms for client set-up, students authentication (using fingerprint), and evaluation preferences.
- Support the evaluation process and show final results of evaluation.

4.2 Performance schema (n-Tier C-S system)

We separated the application in two main modules: the first one is in charge of the conduction the on-line assessment, and the second one in charge of fingerprint recognition and web cam monitoring in real time.

The server must be in listening mode waiting for Clients that require a service. In order to use fingerprint recognition, the first step is to enrol students –top, right side in Figure 4-, the student's fingerprint is saved and indexed in the Features Database. We highly recommend separating this from the Assessment System Database, using even separated servers, to improve overall system performance A Student Personnel ID is assigned in the features database which is used to link the students' personnel information with the fingerprint image.

If the Server is in listening mode and the student has been enrolled, the assessment process can start. The student enters the on-line assessment application, and when the system requires the username and password, the student uses the Mouse Id –superior right side of Fig. 4- to scan his/her fingerprint.

The fingerprint is verified in the Features Database, and if it is recognized as valid, the Server authorizes access to the on-line assessment application. If the print is invalid, an error

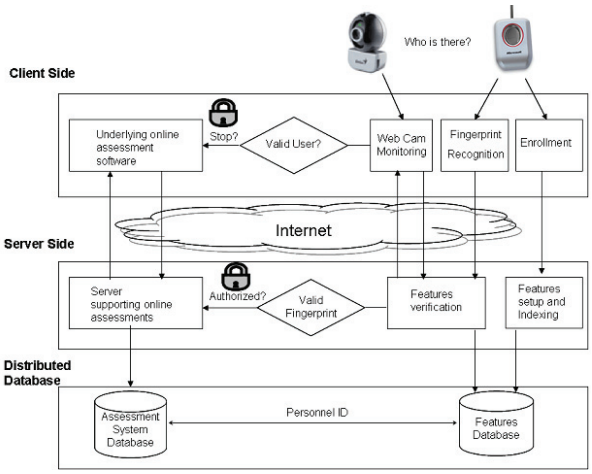

Fig. 4. Fingerprint recognition in real time in online assessments. The Client side (at top) and the Server side (at bottom)

message is sent to the Client instructing it to try again. If, on the other hand, if the student's fingerprint is valid, the user is authenticated into the system, the evaluation process starts and web cam transmission is initialized at the Client Side to conduct real time monitoring by means of multitasking. If someone else tries to get the control of the computer during the on-line assessment, the evaluation process is finished prematurely, and the results are sent to server side to be processed as they are. If the evaluation process is finished successfully, the assessment is processed at the Server Side, and the final results of evaluation and security status are shown at the Client Side. For a detailed review of the performance schema refer to (Hernández-Aguilar et al., 2008).

4.3 Methodology

For our experiment, we selected a random sample of students (n=102) from the José María Morelos y Pavón High School, located in Temixco, Morelos, México. We carried out two evaluations, a control evaluation (paper and pencil), and a second evaluation with our on-line assessment system with biometric recognition.

- **Tests design.** Tests were designed by the professors on August 5th and 6th 2007, of which one was implemented for the on-line assessment using our authoring tool. The tests consisted of 30 questions with similar level of complexity; we evaluated arithmetic, algebra, geometrics, and trigonometric subjects.
- **Setting up.** Computers were prepared and our on-line assessment client software and biometric devices installed, network connectivity was tested.
- **The traditional test**. The paper and pencil test was conducted on August 14th 2007.
- **Enrolment.** Students were enrolled into the system by taking their left-hand index fingerprint on August 15th 2007. We made sure the students were identified by the system after their enrolment.
- **The on-line assessment with biometric recognition test.** The test was conducted in the Computers Network Laboratory located at the High School facilities from August 16th to August 17th 2007. Each computer used in the experiment had a Microsoft Fingerprint

Reader attached, a web cam, and a broad band connection to our server as well as our proprietary client system. First of all, the students were instructed in the usage of the system, and were explained that a web cam was monitoring their activities. At that point, they authenticated by means of their fingerprint into our Server System and the computerized assessment started. The use of calculators and cellular phones was forbidden.

- **The Survey.** At the end of the exam we performed a survey to determine the students' profile and perceptions about our system's operation.
- **Statistical Analysis.** Data was processed using descriptive analysis, using relative numbers and percentages using Ccount gnu free software.

4.4 Preliminary results and discussion

In this test with an obtained FAR of 99.99% and FRR of 97.09%, only one female student could not be recognized despite several trials. Her fingerprint template cannot be understood by the system due to her fingerprints having stain-like shape. Similar cases are registered in (Michigan Org, 2007). We worked around this problem by providing her with a username and a strong password. The average grade in the paper and pencil test was 3.8 while the on-line average was 3.5. This difference can be explained with the fact that a small percentage of students must improve their computer skills. We noticed that video games and chat could improve students' skills and performance in on-line assessments. In general, students perceived our system as faster, easy to use and secure, fingerprint recognition playing an important role in this last point. However, they dislike time limited answers, and 13% dislike web cam monitoring. They felt under pressure, get nervous and dislike being monitored or watched. 20% noticed a way to cheat using a system like ours. We made an in-depth analysis and discovered that students with poor performance (low grades) are more willing and likely to cheat.

4.5 Future work

This work is by no means complete, and could be improved by a number of future research directions. First, we want to compare the results obtained with Weka as a DM engine with results obtained using other software tools like Matlab and SPSS. Secondly, we want to improve the human-computer interface and assessment methodology by analysing students' comments and users' feedback. Third, regarding biometric recognition, we want to improve facial recognition since at this point of our research we can only detect a student's presence or absence, while comparing face patterns automatically by means of photo Ids stored in our features databases is our final intention. Four, we want to include behavioural characteristics in our BDM analysis and verify its feasibility to identify remote users. Finally, we want to test the newest fingerprint scanners and bio-sensors included in new mice, keyboards and in some laptops and try to incorporate them to work within our system.

5. Conclusions

We consider the use of BDM to identify student cheating and impersonation in on-line assessments very important. We believe that Data mining can be successfully used to find out whether the student cheats in on-line tests as long as enough information is provided. Best results can be obtained from information received from remote places and during different intervals of time.

To perform a better DM analysis about student's cheating in on-line assessments we have to pay attention not only to the student's perceptions and behaviour, but to the professor's teaching style as well. The information received from such an analysis can result in very important findings in discussing and improving the testing environments as well as the institution's general philosophy. To study this process behavioural data mining can be performed.

Biometric data mining is a promising technology proposing to deal with the problem of "who is there?" in not only on-line assessments, but on-line recognition in general. This technology is generally very well accepted but must be improved to be perceived as unobtrusive.

6. Future of biometric data mining applied to on-line recognition systems.

Many techniques analyse Mouse Movement, Stylometry, and Keystroke Capture data sets using data mining techniques. Numerous algorithms and methodologies have been used during the last five years. In the future, data sets will be using PredictiveApriori (Ibk), while the Stylometry data set will additionally be analysed using simpleKmeans. All of these machine-learning methods have differences in applicability, meaning there is no one best method; rather, there are only optimal methods, depending on the particular data set. It is important to note that most of the algorithms in reality do not produce results which are 100% accurate. An observed trend is to mix several physiological and behavioural characteristics to perform a better identification of remote users, and consequently high quality data mining, and we will find more creative ways to mix those parameters and techniques in the near future.

For example, although highly accurate results were obtained using sophisticated learning methods on many datasets, some approaches were more successful than others. The most successful approaches have been shown in detail - and future researchers may find that they can improve the results found in the literature using similar techniques. Future researchers may be particularly interested in trying different approaches for their authentication experiments. In these experiments, a community of subjects could be authenticated against another community of subjects. However, for the use of the biometric information to identify an individual, it would be more efficient to attempt authentication based solely on the subject in question. Experiments that would lead to an adequate system for identifying individuals would require splitting the data set into separate data sets that hold only the within and between class records pertaining to each of the subjects.

In many applications of On-line Social Networking as Facebook, when the user moves the location, for example another country with government restrictions as Belarus, the user's authenticity is proven by answering questions about the person's contact photos, and analysing the time required to answer each question. Such an approach enables detection of an impostor. Once there are people recognized within one's social networking contacts, the user is given 7 pictures of which he must identify at least five, and is given a 15 seconds time frame to do so for each.

Authentication and identification tests will be run in order to produce a comparative study on Mouse Movement, Stylometry, and Keystroke Biometric data. In most cases, the identification tests will result in higher accuracies than the authentication tests. Reactive recommendations will be made for legitimately increasing the classification accuracy of the authentication experiments.

The k-nearest neighbour approach will be used with cross validation and 80% splits on the Stylometry and Keystroke Capture data sets, which showed high accuracy results for non-training data before. Decision rule and k-means clustering will be used on the Stylometry data set and make for interesting experiments, in that the decision rules may be useful to improve future researches and in that the clustering algorithm will show an average accuracy that is similar to the accuracies achieved using the k-nearest neighbour approach with cross validation on the modified Stylometry data. Many researches have extended previous studies by running additional experiments on the Mouse Movement, Stylometry, and Keystroke Biometric data, new and previously obtained, using data mining tools like Orange, Weka or SAS. The data mining algorithms with which the experiments will be conducted are widely used and provide an entry point for future researchers into the use of data mining with biometric data sets.
Biometric data mining will be used to identify emotions and will identify suspicious biometrics to fight against crime and terrorism - to see what is being done now refer to (Security Focus, 2006) and related web sites. Biometric technologies are coming of age due to the need to address heightened security concerns in the 21st century. Consequently, the performance and robustness of systems will increase significantly but more research effort is necessary. In the future, the use and reconstruction of 3D images, as well as the use of virtual reality will increase, and biometric data mining will analyse this data as well as physiological and behavioural characteristics simultaneously.

7. References

Adomavicius, G. & Tuzhilin, A. (2001). Using data mining methods to build customer profiles. *IEEE computer,* Vol. 34 (2), 74-82

Amaratunga, D. & Cabrera, C. (2004). Mining data to find subsets of high activity. *Journal of statistical planning and inference,* Journal of Statistical Planning and Inference. 122, 23-41, ISSN 0378-3758

Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases. http://archive.ics.uci.edu/ml/

Burlak, G.; Hernandez, J.A. & Zamudio-Lara, A. (2005). The Application Of Online Testing For Educational Process In Client-Server System, *Proceedings of CONGRESS: IADIS International Conference*, pp. 389-392, ISBN 972-8924-04-6 , October 2005, Lisboan, Portugal

Brodley, C. & Pusara, E. (2004). User Re-Authentication via Mouse Movements, *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pp. 1-8, ISBN 1-58113-974-8, Washington, D.C., ACM, NY, USA

BSU Ball State University (2006) Technology assessment reports on The use of Biometrics in education. http://web.bsu.edu/dpmuller/biometrics/

Dunstone, T. & Yager, N. (2009). *Biometric System and Data Analysis. Design Evaluation and Datamining*. Springer Science+Business Media, LLC. ISBN-13: 978-0-387-77625-5. USA

Eusebi, C.; Gliga, C.; John, D. & Maisonave, A. (2008). A Data Mining Study of Mouse Movement, Stylometry, and Keystroke Biometric Data, *Proceedings of Student-Faculty Research Day*, pp. B1.1-B1.6, Pace University, May 2008, NY, USA.

Fayyad, U. et al. (1996). From Data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data mining. MIT Press. 1-34.

Han, J. & Kamber, M. (2002). *Data mining Concepts and Techniques,* Morgan Kaufman Publishers, ISBN 1-55860-489-8, CA, USA.

Hansen, M. (2009). Concepts of Privacy-Enhancing Identity Management for Privacy-Enhancing Security Technologies, *Proceedings of PRISE Conference: Towards privacy enhancing security technologies – the next steps*, pp. 91-93, April 2008, Vienna.

Hernández-Aguilar, J.A.; Burlak, G. & Lara, B. (2010). Design and Implementation of an Advanced Security Remote Assessment System for Universities Using Data Mining. Resumen de Tesis Doctoral. *Revista Computación y Sistemas,* Vol.13, No.4, 463-473, ISSN 1405-5546

Hernández-Aguilar, J.A.; Ochoa, A.; Andaverde, J. & Burlak, G. (2008). Biometrics in online assessments: A Study Case in High School Students, *Proceedings of 18th International Conference on Electronics, Communications and Computers Conielecomp 2008*, pp. 111-116, ISBN 0-7695-3120-2 & ISBN 978-0-7695-3120, Cholula, February 2008, IEEE-Explore, Puebla, México

Hilario, M.; Kalousis, A.; Prados, J. and Alain, P. (2004). Data mining for mass spectra-based cancer diagnosis and biomarker discovery. *Biosilico,* Vol.2, No.5, 214-222, ISSN 1741-8364.

Hiller, J. & Cohen, R. (2001). *Internet Law and Policy, Prentice-Hall*, ISBN 9780455222639, Upper Saddle River, NJ.

Iglesias, R.; Orozco, M. & Valdes, J. (2008). Characterizing Biometric Behaviour Through Haptics and Virual Reality, *Proceedings of ICCST*, pp. 174-179, ISBN 978-1-4244-1817-6, Lillehammer, April 2008, IEEE Computer Society, Norway

Jarvenpaa, S.; Tractinsky, N, and Vitale, M., (2000). Consumer trust in an Internet store. *Information Technology and Management,* Vol.1, No. 1-2, 45-71, ISSN 1741-5179.

Kaklauskas, A.; Zavadskas, E.K.; Seniut, M. et al. (2008). Web-based biometric mouse decision support system for user's emotional and labour productivity analysis, *Proceedings of The 25th International Symposium on Automation and Robotics in Construction*, pp. 69-75, Lithuania, June 2008, Institute of Internet and Intelligent Technologies, Vilnius

Lovell, B. C. & Chen, S. (2008). *Robust face recognition for data mining. In John Wang, editor, Encyclopedia of Data Warehousing and Mining, volume II ,deaGroup Reference, ISBN 1-59140-559-9, Hershey, PA.*

Michigan Org. (2007) Fingerprint Classification. http://www.reachoutmichigan.org/funexperiments/agesubject/lessons/prints_ext.html

Neil, R.L. (2008) Privacy, Biometrics and the war on terror. http://www.acus.org/new_atlanticist/biometrics-civil-liberties-and-war-against-terror

Ochoa, A.; Hernández, A.; Sánchez, J.; García, Y. et al. (2009). Identify an Adequate Antropometry to Water Polo Using Social Data Mining, *Proceedings of International Conference on Electrical, Communications, and Computers, 2009*, pp. 144-147, ISBN ISBN: 978-0-7695-3587-6, Cholula, February 2009, IEEE Xplore, Puebla, México

Pavone, V. and Pereira M. (2009). The privacy Vs security dilemma in a risk society. *Proceedings of PRISE Conference: "Towards privacy enhancing security technologies – the next steps"*, pp. 109-127, Vienna.

Pirolli, P.L. (2007). Information Foraging Theory: Adaptive Interaction with Information, Oxford University Press, Cambridge, UK

Revett, K.; De Magalhaes, S.T. & Santos, H. (2005). Data Mining a Keystroke Dynamics Based Biometrics Database Using Rough Sets, *Proceedings of Portuguese conference on Artificial intelligence, 2005. epia 2005* , pp. 188-191, ISBN 0-7803-9366-X, Harrow Sch. of Comput. Sci., Westminster Univ., Dec 2005, IEEE Xplore, London

Security Focus. (2006). Biometric polygraph next for airport security? Biometric polygraph next for airport security? http://www.securityfocus.com/brief/281

Shalhoub, G.; Simon, R.; Iyer, R. et al. (2010). Stylometry System – Use Cases and Feasibility Study, *Proceedings of Student-Faculty Research Day, CSIS*, pp. A3.1-A3.8, Pace University, May 2010, NY, USA.

Strater, K. & Lipford, H. (2008). Strategies and Struggles with Privacy in an Online Social Networking Community, *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1*, pp. 111-119,ISBN 978-1-906124-04-5, Liverpool, UK , 2007, *British Computer Society*, Swinton, UK

Tapiador, M. & Singuenza, J.A. (2005). *Tecnologías biométricas aplicadas a la seguridad,* AlfaOmega Grupo Editor, S.A. de C.V., México, D.F.

Turban, E.; Aronson, J.; Liang, T. & Sharda, R. (2004). *Decision Support and Business Inteligence Systems,* Pearson Education, Inc., ISBN 0-13-198660-0, New Jersey, USA

Umamaheswari, K.; Sumathi,S.; Sivanandam, S.N. & Anburajan K.K. (2007). Neuro Genetic-Nearest Neighbor Based Data Mining Techniques for Fingerprint Classification and Recognition System. *ICGST-GVIP Journal,* Vol.7, No.3, November 2007, 39-50, ISSN 1687-3998

Van der Ploeg, I. (2005). The Politics of Biometric Identification, In: *Biometric Technology & Ethics,* BITE, Policy Paper no. 2, 1-16, www.biteproject.org/documents/politics_of_biometric_identity%20.pdf

Vetro, A.; Drapper, S.; Rane, S. & Yedidia, J. (2009). *Securing Biometric Data,* MERL, Technical Report TR-2009-002, Massachusetts, USA.

Weiss, A.; Ramapanikar, A.; Shah, P.; Nobel, S. & Immohr, L. (2007). Mouse Movements Biometric Identification: A feasibility Study, *Proceedings of Student/Faculty Research Day*, Pace University, May 2007, NY, USA

Westin, A. (1967). Privacy and Freedom, *Atheneum*, New York

Yang, Y. & Padmanabhan, B. (2010). Toward user patterns for on-line security: Observation time and online user identification. *Decision Support Systems,* Vol. 48, 548-558, ISSN 0167-9236