International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:11, No:8, 2017

Identity Verification Using k-NN Classifiers and Autistic Genetic Data

Fuad M. Alkoot

Abstract**—**DNA data have been used in forensics for decades. However, current research looks at using the DNA as a biometric identity verification modality. The goal is to improve the speed of identification. We aim at using gene data that was initially used for autism detection to find if and how accurate is this data for identification applications. Mainly our goal is to find if our data preprocessing technique yields data useful as a biometric identification tool. We experiment with using the nearest neighbor classifier to identify subjects. Results show that optimal classification rate is achieved when the test set is corrupted by normally distributed noise with zero mean and standard deviation of 1. The classification rate is close to optimal at higher noise standard deviation reaching 3. This shows that the data can be used for identity verification with high accuracy using a simple classifier such as the k-nearest neighbor (k-NN).

Keywords**—**Biometrics, identity verification, genetic data, k-nearest neighbor.

I. Introduction

IDENTITY verification and identification using machine learning techniques has become popular in the past two decades. Many identity verification methods exist using the person's physiological characteristics (known as biometrics). Although biometrics do not solve all security problems and are not appropriate for all applications, they represent powerful means of identity verification in some applications when correctly implemented. Some biometrics such as fingerprint or palm-print require the active participation of the subject, while other techniques, such as face recognition, do not require any participation or knowledge of the subject.

Fingerprint, voice and face are the earliest biometrics used and implemented as a machine learning based identification technique. However, advances in the field have expanded the options to include other biometrics such as iris, retina, ear, vein, gait, smell and more. Each biometric is suitable for certain but not all circumstances. DNA, which has been used in the criminal justice system, has lately gained interest for biometric identification. The slow speed of the process from taking the sample of the subject to identification using a machine learning based system has delayed its use commercially. There is an ongoing research that focuses on rapid DNA testing. Authors [1] demonstrate a rapid cycling protocol that amplifies 15 STR loci and the sex-typing marker amelogenin from either Identifier or PowerPlex 16 STR typing kits in less than 36 minutes. This is in contrast to existing methods for PCR amplification portion only that take approximately three hours. Current forensic DNA typing is conducted in approximately eight to 10 hours with steps including DNA extraction, quantitation, polymerase chain reaction (PCR) amplification of multiple short tandem repeat (STR) loci, capillary electrophoresis separation with fluorescence detection and data analysis, and DNA profile interpretation, [1]. According to [2], the likelihood that any two individuals (except identical twins) will have the same 13-loci DNA profile can be as high as 1 in 1 billion or greater.

Previously there were attempts at using chromosome data in machine learning for medical diagnosis [4]-[7]. Gene data has been used for autism detection [8], the data used here is from [3]. We experiment using partial DNA data from a single chromosome which was intended for automated identification of the autism disorder [8] as a tool for person identity verification. We show that it is possible to quickly and accurately identify the subjects based on this partial data, using a nearest neighbor classifier. Investigations on this data may be continued furthermore to find, for example, the minimum number of features that yield maximum identification rate. Also, one may calculate the distance to the closest false identity sample. This will also provide us with an estimate of the robustness of the identification system.

In the next section, we present the data followed by a section on experimental methodology. This is followed by a section on results and the paper is brought to conclusion in the last section.

II. The Genome Data

Autism is correlated with DNA copy number variations (DCV). In this paper we use the data preprocessed by [3]. The following description of the data and preprocessing method is extracted from [3], where they used a statistical-based approach to analyze 75Mb of DNA samples of 71 autistic children and 71 typically developing children to prove the existence of significant differences between the two groups in their copy number burden. Even though the DNA copy numbers variations occur frequently in the genome of normal people, especially in the segmental duplication regions SDs, it has been demonstrated that some variations are associated with behavioral and developmental abnormalities such as cognitive impairment, autism, mental retardation, and possibly psychiatric diseases. Different studies tested the whole genome and detected autism-related abnormalities in 5 SD-rich intervals. Our study is confined to analyze and detect the recurrent variations across these five intervals which have a total length of 75Mb using family-tiled oligonucleotide arrays. Table 1 shows the genomic locations of each interval. The

Fuad M. Alkoot is with the HITN, PAAET, Kuwait (e-mail: fm.alkoot@paaet.edu.kw).

data, made available by [3], is initially preprocessed by to enhance the class separation. The preprocessing techniques used by [3] is explained as follows:

For a given aCGH profile, the data can be modeled as piecewise function corrupted by noise (e.g. AWGN). Formally,

$$y[n] = f[n] + w[n] \; where \; n = 0,1,2,\dots N-1 \qquad (1)$$

where $y[n]$ is the observed DCN data, $w[n]$is AWGN and $f[n]$ is the true DCN signal to be estimated with M segments defined as:

$$f[n] = \sum_{i=1}^{M} A_i \left[u[n_{i-1}] - u[n_i]\right]. \qquad (2)$$

where $n_0=0<n_1<n_2< \dots <n_{M-1}<n_{M=N}$ and $u[n]$ is the unit step function. Here A_i and ni are the unknown parameters representing the intensity level and the breakpoint, respectively. N is the length of DCN data. Moreover, each variant region is assumed to be statistically independent of all other regions. Hence, the PDF of the entire data record can be written as:

$$p(\mathrm{y};\mathrm{A},\mathrm{n}) = \prod_{i=1}^{M} p_i(y[n_{i-1} : n_i - 1]; A_i) \; = \qquad (3)$$

$$\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} \left[\sum_{n=n_{i-1}}^{n_i - 1} \left(y[n] - A_i \right)^2 \right] \right].$$

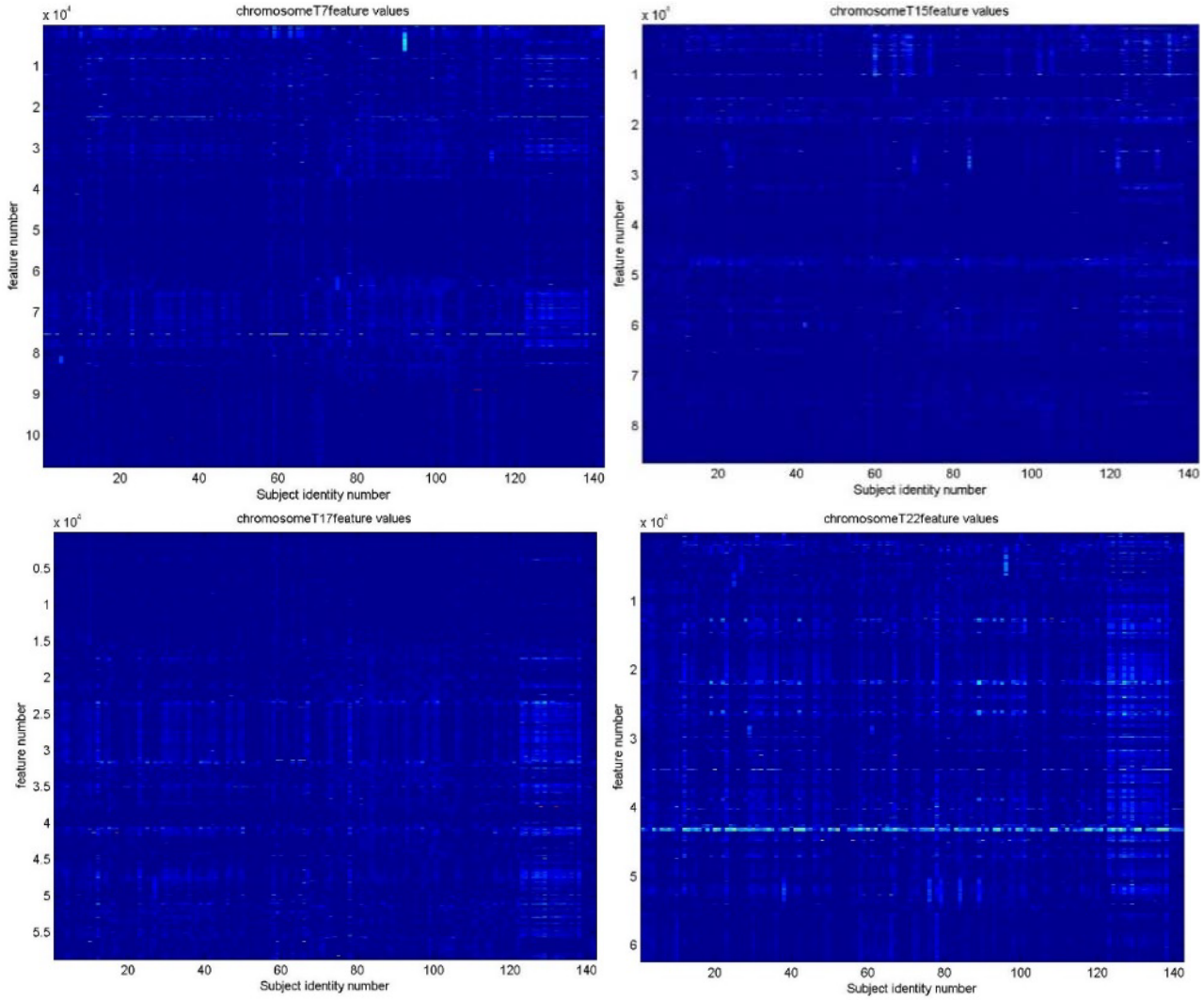

Fig. 1 Maps of values of four chromosomes for all identities using the full feature set

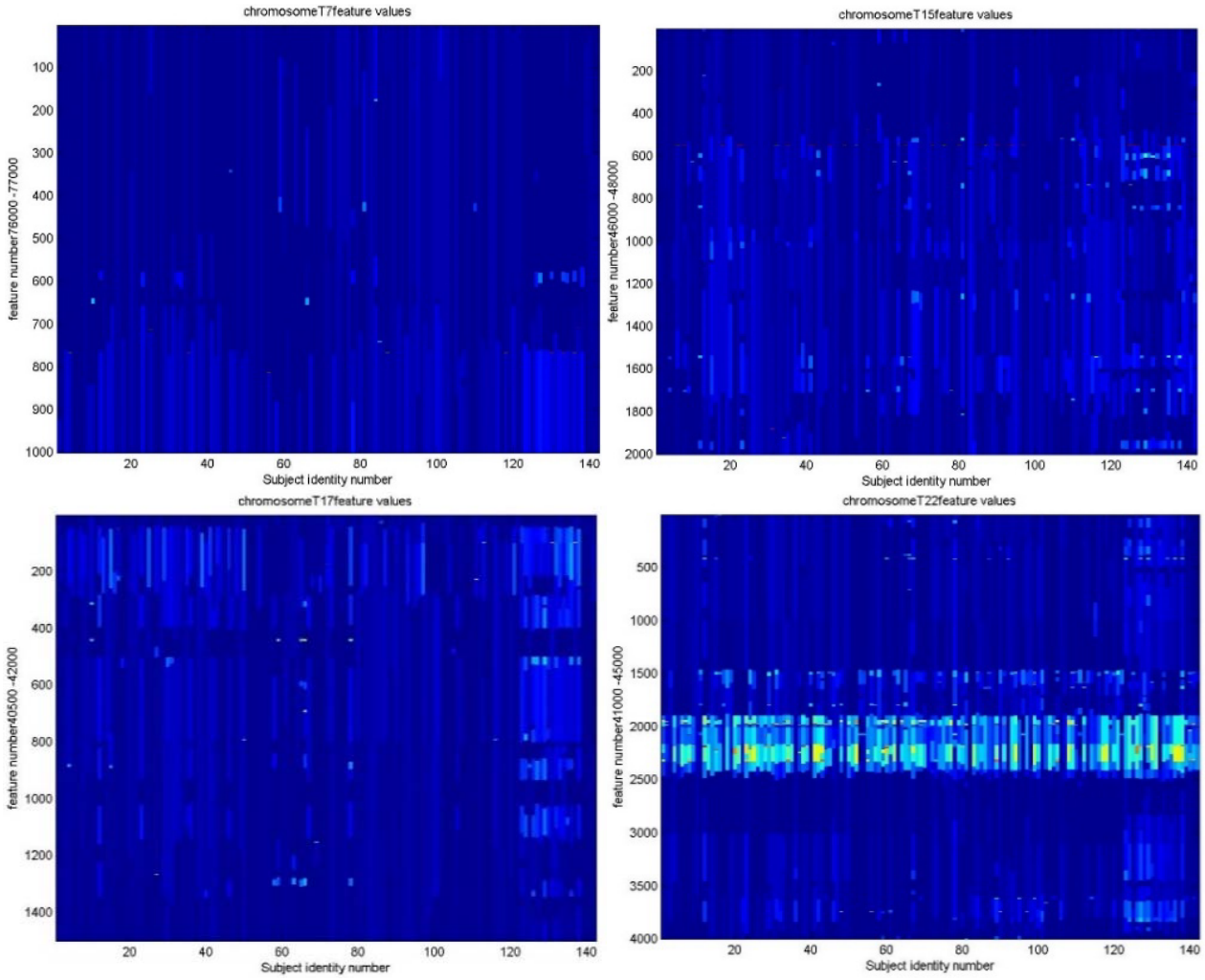

Fig. 2 Maps of values of four chromosomes for all identities using a subset of the feature set

Before further analysis, we apply our recently developed method, Maximum Likelihood Estimator (MLE) [1], to identify and detect the variant regions by discretizing the normalized aCGH datasets. The method can be summarized as follows.

1. Estimate the number of variant segments using minimum description length (MDL) algorithm.

$$MDL(k) = -\ln \prod_{i=1}^{k} p_i\left(y[\hat{n}_{i-1}], \ldots, y[\hat{n}_i - 1]; \hat{A}_i\right) + \frac{m_k}{2} \ln N. \quad (4)$$

where mk is the number of estimated parameters or equivalently the dimensionality of the unknown parameters Ai's and ni's with k breakpoints.

2. Estimate the values of the breakpoints ni's of the variant segments maximizing the likelihood ratio test (LRT) or minimizing the least square errors.

$$J(A, n) = \sum_{i=1}^{M} \left[\sum_{n=n_{i-1}}^{n_i - 1} \left(y[n] - \hat{A}_i\right)^2 \right]. \quad (5)$$

3. Evaluate the predicted segments values using the sample mean for the points within the segment boundaries.

$$\hat{A}_i = \frac{1}{n_i - n_{i-1}} \sum_{n=n_{i-1}}^{n_i - 1} y[n]. \qquad for\ i = 1 : M \quad (6)$$

More details about the algorithm step can be found in [3].

TABLE I
STUDIED INTERVALS

chromosome	Start	end	length
7	61058424	81999980	20,941,556
10	77000071	91999901	14,999,830
15	18260026	34999924	16,739,898
17	12000112	22187009	10,186,897
22	14430001	25999992	11,569,991

Looking at the data after the preprocessing method of [3], we find that the number of features is very large reaching tens of thousands with mixed class distributions. Therefore, a second stage of feature selection is required for the data to become useful for further machine learning based classification process. However, no further feature selection method is required for use in a biometric identification task, as will be discussed in the results section.

III. Experiment Methodology

As described in the previous section the data consists of 142 samples where each sample belongs to a different identity. Therefore, in order to perform an identification task all the data must be used in the design of the classifier. The test set can be created from the training set by adding random noise to the samples. Due to the nature of the genetic data, any reproduction of genetic data will yield identical data for each subject. Therefore, it is reasonable to use the same sample per subject that was used in the training set, as a test sample. However, to simulate the case when the data produced for an identity is different at each production trial we add a random value to the test set. The added random noise has a normal distribution with zero mean. We vary the standard deviation to simulate increasing degrees of error in the data production process. When the standard deviation is 1 we obtain 100% accurate classification. As the standard deviation is increased the features deviate further from their original value and matching identities becomes more difficult.

The nearest neighbor classifier is used to classify or identify subjects. The Euclidean metric is used as a distance measure in the nearest neighbor classifier.

IV. Results

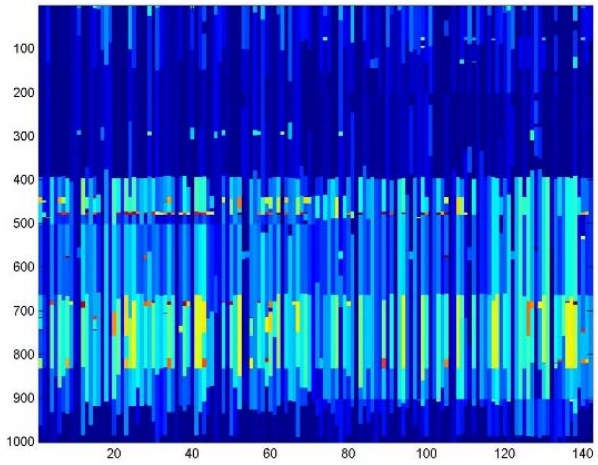

Fig. 3 Maps of values of chromosomes N22 for all identities zoomed at 1000 features

Using the nearest neighbor classifier we achieve 100 percent classification when the test set is corrupted by normal noise with a standard deviation of 1. As the standard deviation is increased the classification rate decreases to the levels shown in Table II.

TABLE II
Classification Rate of the Nearest Neighbor Classifier for the Different Noise Standard Deviation Values at Each of the Five Chromosomes

Standard deviation	Chromo. N7	Chromo. N10	Chromo. N15	Chromo. N17	Chromo. N22
5	85.21	61.27	97.88	71.83	91.55
4	95.07	80.99	99.29	81.69	97.89
3	98.59	93.66	100.	96.48	100.
2	100.	100.	100.	98.59	100.

V. Conclusion

We experiment with DNA data for use in an identity verification application. Using a nearest neighbor classifier we found that our data that was originally prepared for autism detection can be used to identify subjects. The classification rate of 100 percent was obtained when the test data was corrupted by normally distributed noise with a standard deviation of 2. At standard deviation of 3, high classification rate close to optimal was also obtained.

References

[1] Peter M. Vallone; Carolyn R. Hill; John M. Butler, “Demonstration of Rapid Multiplex PCR Amplification Involving 16 Genetic Loci”. Forensic Science International, Vol. 3, 1, pp42:45. 2008.

[2] John Ashcroft, Deborah J. Daniels Sarah V. Hart. Using DNA to Solve Cold Cases, U.S. Department of Justice Office of Justice Programs National Institute of Justice special report, July 2nd 2002.

[3] Abdullah Alqallaf and Ahmed Tewfik, “Maximum Likelihood Principle for DNA Copy Number Analysis,” IEEE Int’l Conference on Acoustics, Speech, and Signal Processing, IEEE/ICASSP, Taipei, Taiwan, April, 2009.

[4] Alexey Tsymbal, Padraig Cunningham, Mycola Pechenizkiy and seppo Puuronen, Search strategies for ensemble feature selection in medical diagnosis, 16th IEEE symposium on computer based medical systems, 2003, June 2003, 124-129.

[5] M. P. Sampat, et. al., Supervised parametric and non parametric classification of chromosome images, Pattern Recognition 38(2005) 1209-1223.

[6] Hyunseok kook et. al., Multi-stimuli multi-channel data and decision fusion strategies for dyslexia prediction using neonatal ERPS, Pattern Recognition vol. 38, no 11, 2005, 2174-2184.

[7] Hojin Moon, Hongshik Ahn, Ralph L Kodell, Chien-Ju Lin, Songjoon Baek, and James J Chen, Classification methods for the development of genomic signatures from high-dimensional data, Genome Biol. 2006; 7(12): R121.

[8] Fuad M. Alkoot, Abdullah K. Alqallaf. Investigating machine learning techniques for the detection of autism. International Journal of Data Mining and Bioinformatics, 2016; 16 (2): 141 DOI: 10.1504/IJDMB.2016.10000981.

Fuad M. Alkoot obtained his PhD in Electronic Engineering from CVSSP - University of Surrey, UK in 2001. He obtained his M.Sc. And B.Sc. in Electrical Engineering from Rochester Institute of Technology, N.Y. and Fairleigh Dickenson University, N.J, USA, respectively. He is currently affiliated with HITN-PAAET, Kuwait, as an academic teaching staff. His main research interests are fusion methods, multiple classifier systems, biometric authentication methods, combiner methods and bioinformatics.