SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING
Available online at www.scjournal.com.ba

Denver Groups Classification of Human Chromosomes Using Fuzzy C-Means Clustering

Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences
Hrasnicka Cesta 15, Ilidza 71200, Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba

Abstract— Unbanded human chromosome can be classified into seven Denver Groups (A-G) based their lengths and the ratio of the length of the shorter arm to the whole length of the chromosome, which is called the centromere index (CI). In this article, the fuzzy c-means method will be used to perform the Denver Group classification of a given set of human chromosomes. The objective in clustering is to partition a given human chromosome set into homogeneous clusters; by homogeneous we mean that all points in the same cluster share similar attributes and they do not share similar attributes with points in other clusters. However, the separation of clusters and the meaning of similarity are fuzzy notions and can be described as such. It is found that the clusters iterations converge, highly depend on the initial partition matrix, $\widetilde{U}^{(0)}$.

Keywords— human chromosome, Denver Group, fuzzy c-means clustering, unbanded chromosomes, centromere index.

1. INTRODUCTION

In 1956 Tjio and Levan using the improved cell culturing and staining technique discovered that the number of human chromosomes is 46 (Tjio, and Levan 1956). From this time on, the research on chromosomal abnormalities, as a cause of diseases, became one of the main branches of the molecular biology.

Disorder in human chromosomes is a powerful indicator in diagnosis of leukemia, skin and breast cancers, and other genetic diseases. Clinical laboratories routinely performed researches to identify chromosome abnormalities, and provide medical doctors the diagnostic results and help them decide therapeutic treatments for patients.

The most prominent difficulty in chromosome analysis is the absence of clear microscopic chromosome images. The variation of cell culturing conditions, chromosome staining, and microscope illumination make finding analyzable chromosomes in a genetics clinical laboratories very difficult. For human experts, identification and classification of chromosomes is a tedious and time-consuming task. The human error also introduces variation and affects the accuracy of the diagnostics made by physicians.

The development of computer-assisted metaphase finding and karyotyping systems, slowed down by the noisy cell images.

2. HUMAN CHROMOSOMES

Since Waldeyer in 1898 (Verma, and Babu 1995) coined the term chromosome, it is known that chromosomes resides within a cell's nucleus, and contains the person's deoxyribonucleic acid (DNA). Each chromosome is made up a single extremely long DNA molecule. Using cells cultured from fetal lung tissue, Tjio and Levan, demonstrated that human cells contain 46 chromosomes as they appear during cell division or mitosis. A healthy human cell nucleus includes 44 autosomes and 2 sex chromosomes: X and Y.

The test cells used for chromosome imaging and analysis are taken mostly from blood sample, amniotic fluid, and bone marrow. These test samples are cultured overnight in a mitotic arresting agent. Then cells are processed with hypotonic solutions to increase cell volume. This procedure spreads the chromosomes apart. The methanol-acetic acid is used to fix them for analyses. The fixed cells are dropped onto a standard glass microscope slide and allowed to dry.

If karyotyping and classification are going to be performed using banded chromosomes, the slide is then subjected to a staining process. Staining makes clear the distinctive reproducible patterns of bands along chromosomes. These bands permit accurate identification of chromosomes and recognition of abnormalities.

2.1 Classification of Banded Chromosomes

In order to improve the performance of automated chromosome classification including recognition of disordered chromosomes, artificial intelligence and machine learning methods have been widely used in the computer-assisted chromosome detection and classification systems (Gagula-Palalic, and Can 2012). Among them, ANN is the most popular tool owing to its capability of modeling the human brain decision making process to recognize objects based on incomplete or partial information, as well as its simple topographic structure and easier training process (Mitchell, 1997).

Early studies also indicated that ANN performance could achieve comparable results compared with that obtained by simpler statistical methods (Sweeney, 1993). A large number of different feature based and pixel value distribution based ANN have been tested and evaluated in classification of banded chromosomes, which include supervised multi-layer neural networks (Delshadpour, 2003, Wu et. Al., 1990, Can, and Gagula-Palalic 2012), Hopfield network (Ruan, 2000), and unsupervised architecture of self organizing nonlinear maps (Lerner et. Al., 1996), SOFM (Kyan et. Al. 1999) and mutual information maximization based training method (Mousavi et. Al., 1999).

However, the study found that performance of unsupervised nonlinear learning methods was lower than a supervised nonlinear paradigm (Lerner et. Al., 1996). Although ANN is a powerful machine learning tool in pattern recognition and classification, its relatively poor robustness in detection and classification of abnormalities depicted on the complicated chromosome images and its 'black box' type of optimization approach are its major disadvantages.

To provide researchers and clinicians with a better understanding of the logic or reasoning in automated classification of chromosomes, a variety of knowledge-based 'expert' systems were developed and evaluated (Gagula-Palalic, and Can 2012). Since clinical technicians are trained to recognize the chromosomes under non-ideal conditions, many researchers tried to record and apply or mimic the rules of manual karyotyping and diagnosis of chromosome irregularity into a knowledge-based automated classification system in an attempt to minimize the classification errors.

Hence, researchers worked with clinicians, observed their diagnostic process, summarized and quantify the diagnostic rules, and then converted these rules into the computer classification systems (Wu et. Al., 1989, Lu, and Ya 1989, Ramstein et. Al., 1992). The systems would then be trained on a bank of chromosome images, refining the rules as needed until the recognition rate was maximized. A major problem with such knowledge-based approach is the difficulty of converting karyotyping guidelines and intuitive notions (empirically diagnostic rules) into concrete rules that can be effectively programmed and applied in a computer-assisted scheme. Owing to this difficulty, the most popular knowledge-based classification system is a fuzzy logic rule-based system, which offers great promise for improving the recognition rate (Keller et. Al., 1995). One blind test involving a dataset of 180 chromosomes distributed in three classes demonstrated 88% classification accuracy using an automated system involving six phases of fuzzy logic rules (Sjahputera, and Keller, 1999).

2.2 Classification of Unbanded Chromosomes

When the chromosomes are not banded, they can be classified into seven Denver Groups (A-G) (H. C. S. Group, 1960) as seen in Table1. Denver Group classification is mainly based on:
(1) the length or size of each chromosome and
(2) the ratio of the length of the shorter arm to the whole length of the chromosome, which is called the centromere index (CI).

Table 1: The classification of chromosomes based on Denver Group classification

Chromosome Class	Denver Group
#1-#3	Group A
#4-#5	Group B
#6-#12,X	Group C
#13-#15	Group D
#16-#18	Group E
#19-#20	Group F
#21-#22,Y	Group G

In this article, the fuzzy c-means method will be used to perform the Denver Group classification of a given set of human chromosomes.

3. FUZZY c-MEANS (FCM)

The concept of a fuzzy set first arose in the study of problems related to pattern classification (Bellman et al., 1966). Since the recognition and classification of patterns is integral to human perception, and since these perceptions are fuzzy, this study seems a likely beginning (Zadeh, 1971). This section presents a simple idea in the area of classification and has dealt in depth with a particular form of classification using a popular clustering method: FCM.

The objective in clustering is to partition a given data set into homogeneous clusters; by homogeneous we mean that all points in the same cluster share similar attributes and they do not share similar attributes with points in other clusters. However, the separation of clusters and the meaning of similarity are fuzzy notions and can be described as such. One of the first introductions to the clustering of data was in the area of fuzzy partitions (Ruspini, 1969, 1970, 1973a), where similarity was measured using membership values. In this case, the classification metric was a function involving a distance measure that was minimized.

Ruspini (1973b) points out that a definite benefit of fuzzy clustering is that stray points (outliers) or points isolated between clusters (Figure 1) may be classified this way; they will have low membership values in the clusters from which they are isolated. In crisp classification methods, these stray points need to belong to at least one of the clusters, and their membership in the cluster to which they are assigned is unity; their distance, or the extent of their isolation, cannot be measured by their membership. These notions of fuzzy classification described in this section provide for a point of departure in the recognition of known patterns.

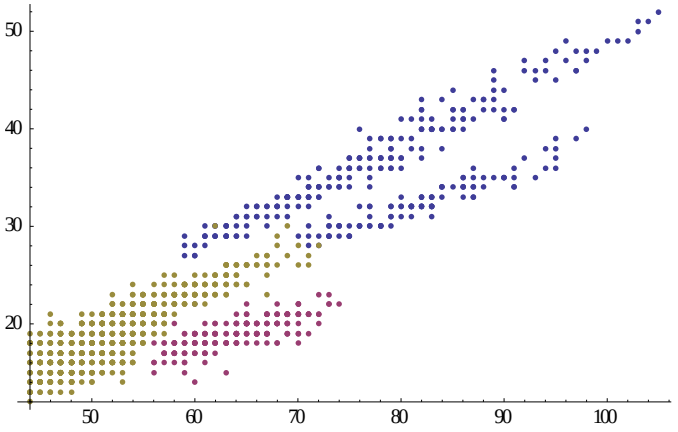

Figure 1. In fuzzy clustering outliers or points isolated between clusters will have low membership values in the clusters from which they are isolated.

To develop fuzzy methods in classification, we define a family of fuzzy sets $\{\tilde{A}_i = 1,2,\ldots,c\}$ as a fuzzy c-partition on a universe of data points, X. Because fuzzy sets allow for degrees of membership, we can assign membership to the various data points in each fuzzy set. Hence, a single point can have partial membership in more than one class. It will be useful to describe the membership value that the kth data point has in the ith class with the following notation:

$$\mu_{ik} = \mu_{\tilde{A}_i}(x_k) \in [0,1],$$

with the restriction that the sum of all membership values for a single data point in all of the classes has to be unity:

$$\sum_{i=1}^{C} \mu_{ik}{}' = 1, \text{ for all k } = 1,2,\ldots,\text{n.} \qquad (1)$$

There can be no empty classes and there can be no class that contains all the data points. This qualification is depicted by the following expression:

$$0 < \sum_{i=1}^{C} \mu_{ik}{}' < n. \qquad (2)$$

Because each data point can have partial membership in more than one class, one has,

$$\mu_{ik} \wedge \mu_{jk} \neq 0. \qquad (3)$$

We can now define fuzzy c-partitions $\tilde{U}_{c\times n}[\mu_{ik}]$.

Fuzzy c-Means Algorithm

To describe a method to determine the fuzzy c-partition matrix $\tilde{U}$ for grouping a collection of n data sets into c classes, we define an objective function J_m for a fuzzy c-partition:

$$J_m(\tilde{U}, v) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\mu_{ij})^{m'}(d_{ij})^2, 1 \le m \le \infty \qquad (4)$$

where $\tilde{U}$ is the partition matrix, v_i are cluster centers, d_{ij} are Euclidean distance measures in *m*-dimensional feature space, between the j^{th} data sample x_j and the i^{th} cluster center v_i, and μ_{ij} is the membership of j^{th} data point to the i^{th} class.

Partition matrix $\tilde{U}$ is used for grouping a collection of *n* data sets into *c* classes, and as such each entry in the partition matrix is represented by the membership function μ_{ij}. The Euclidean distance and cluster centers are given by equations (5) and (6).

$$d_{ij} = \left[\sum_{k=1}^{m}(x_{jk} - v_{ik})^2\right]^{1/2} \qquad (5)$$

$$v_{ik} = \frac{\sum_{j=1}^{n} \mu_{ij}^{\alpha} x_{ji}}{\sum_{j=1}^{n} \mu_{ij}^{\alpha}} \qquad (6)$$

The fuzzy C means is trying to tune the partition matrix, centers and distances, so that the objective function J_m is minimized (Ross 2004).

A new parameter is introduced in Equation (10.28) called a weighting parameter, m (Bezdek, 1981). This value has a range $\alpha \in [1, \infty)$. This parameter controls the amount of fuzziness in the classification process.

As with many optimization processes, the minimized objective function J_m cannot be guaranteed to be a global optimum. What we seek is the best solution available within a prespecified level of accuracy. An effective algorithm for fuzzy classification, called iterative optimization, was proposed by Bezdek (1981). The steps in this algorithm are as follows:

1. Fix c ($2 \leq c < n$) and select a value for parameter α. Initialize the partition matrix, $\tilde{U}^{(0)}$. Each step in this algorithm will be labeled r, where r =0, 1, 2,...

2. Calculate the c centers $\{v_i^{(r)}\}$ for each step.

3. Update the partition matrix $\tilde{U}^{(r)}$ as follows:

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(d_{ik}^{(r)} / d_{jk}^{(r)}\right)^{2/(\alpha-1)}\right]^{-1} \qquad (7)$$

4.If $\left\|\tilde{U}^{(r+1)} - \tilde{U}^{(r)}\right\| \leq \epsilon_L$, stop; otherwise set $r = r + 1$ and return to step 2.

In step 4, we compare a matrix norm $\| \ \|$ of two successive fuzzy partitions to a prescribed level of accuracy, ϵ_L, to determine whether the solution is good enough. In step 3, when the variable $d_{jk}^{(r)}$ is zero, since this variable is in the denominator of a fraction, the operation is undefined mathematically, and computer calculations are abruptly halted. So when some of the distance measures $d_{jk}^{(r)}$ are zero, or extremely small in a computational sense, it is replaced by a small positive real number.

4. DATA DESCRIPTION

The data used in this work is taken from Copenhagen data base. We omitted gray level features, and only keep (1) the length of each chromosome and (2) the ratio of the length of the shorter arm to the whole length of the chromosome, which is called the centromere index (CI).

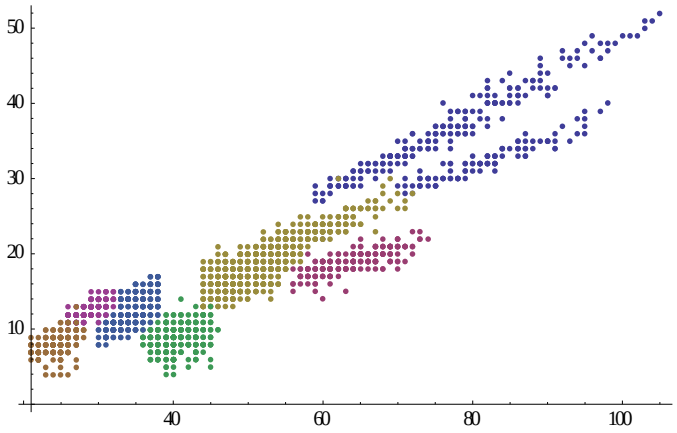

Figure 2. The distribution of 2200 human chromosomes into seven Denver Group classes from A, to G.

5. CLASSIFICATION USING FUZZY c-MEANS (FCM)

Using Fuzzy c-Means Algorithm described in Section 3., it is found that the clusters iterations converge, highly depend on the initial partition matrix, $\tilde{U}^{(0)}$.

Denver Group classification from A, to G are distributed to clusters C1 to C7 as in Table 2. below.

Table 2. Distribution of Denver Group classes A, to G into clusters C1 to C7

clusters	C1	C2	C3	C4	C5	C6	C7
A	69	31	0	0	0	0	0
B	0	67	33	0	0	0	0
C	0	9	75	16	0	0	0
D	0	0	0	91	9	0	0
E	0	0	0	0	78	22	0
F	0	0	0	0	0	100	0
G	0	0	0	0	0	7	93

Correct classification rate of the FCM clustering method is 81.86 %.

7. SUMMARY

Article presents a simple idea in the area of classification and is dealt in depth with a particular form of classification using a popular clustering method: FCM. Although the idea behind the method is very simple, it succeeds to classify given 700 human chromosomes in seven Denver Group classes A, to G with a rate of 81.86 %.

8. REFERENCES

Bellman, R., Kalaba, R., and Zadeh, L. (1966) Abstraction and pattern classification. J. Math. Anal. Appl., 13, 1–7.

Bezdek, J. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York.

Can, M., and S. Gagula-Palalic (2012) Application of Ensemble Machines of Neural Networks to Chromosome Classification, *SEJSC*, Vol 1, No. 231-35.

Delshadpour, S, (2003) Reduced size multi layer perceptron neural network for human chromosome classification, *Proceedings of the 25th Annual International Conference of the IEEE* (Engineering in Medicine and Biology Society).

Gagula-Palalic S.G, and M. Can (2012) Automatic Segmentation of Human Chromosomes, *SEJSC*, Vol 1, No. 2, pp. 80-83.

Gagula-Palalic S, and M. Can (2012) Extracting Gray Level Profiles of Human Chromosomes by Curve Fitting, 66-71

H. C. S. Group (1960) A proposed standard of nomenclature of human mitotic chromosomes, *Cerebral Palsy Bulletin*, vol. 2, pp. 159-162.

Keller, J. M., P. Gader, O. Sjahputera, and C. W. Caldwell (1995) A fuzzy logic rule based system for chromosome recognition, *presented at Proceedings of the Eighth IEEE Symposium on Computer-Based Medical Systems*.

Kyan, M. J., L. Guan, M. R. Amison, and C. J. Cogswell, (1999)Feature extraction of chromosomes from 3D confocal microscope images, presented at 1999 International Conference on Image Processing.

Lerner, B., H. Guterman, M. Aladjem, and I. Dinstein (1996) Feature extraction by neural network nonlinear mapping for pattern classification," presented at Proceedings of the 13th International Conference on Pattern Recognition.

Lu, Y., and Y. Ya (1989) An expert system for banded chromosomes recognition, *presented at Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*

Mitchell, T. M. (1997) *Machine Learning*. Boston MA: WCB McGraw-Hill.

Mousavi, P., P. K. Ward, and P. M. Lansdorp (1999) Feature analysis and classification of chromosome 16 homologs using fluorescence microscopy image," presented at IEEE Can J Elec. & Comp Eng..

Ramstein, G., M. Bernadet, A. Kangoud, and D. Barba (1992) A rule-based image analysis system for chromosome classification, *presented at Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Ross, T. *Fuzzy Logic with Engineering Applications*. John Wiley & Sons, Aug 16, 2004.

Ruan, X. (2000) A classifier with the fuzzy Hopfield network for human chromosomes, Intelligent Control and Automation," presented at Proceedings of the 3rd World Congress on Intelligent Control and Automation.

Ruspini, E. (1969) A new approach to clustering. Inf. Control, 15, 22–32.

Ruspini, E. (1970) Numerical methods for fuzzy clustering. Inf. Sci., 2, 319–350.

Ruspini, E. (1973a) New experimental results in fuzzy clustering. Inf. Sci., 6, 273–284.

Ruspini, E. (1973b) A Fast Method for Probabilistic and Fuzzy Cluster Analysis using Association Measures. Proceedings of the 6th International Conference on System Sciences, Hawaii, pp. 56–58.

Sjahputera O., and J. M. Keller, (1999) Evolution of a fuzzy rule-based system for automatic chromosome recognition, *presented at 1999 IEEE International Fuzzy Systems Conference Proceedings.*

Sweeney W. P., and M. T. Musavi, (1993) Application of neural networks for chromosome classification," presented at Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,

Tjio J. H., and A. Levan (1956) The chromosome number in man, *Hereditas*, vol. 42, pp. 1-6.

Verma R. S., and A. Babu (1995) *Human Chromosomes, Principles and Techniques*, 2 ed. New York: McGraw-Hill.

Wu, Q, P. Suetens, and A. Oosterlinck, (1990) Chromosome classification using a multi-layer perceptron neural net, presented at Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Wu, Q., P. Suetens, and A. Oosterlinck (1989) On knowledge-based improvement of biomedical pattern recognition-a case study, *presented at Proceedings of 5th conference on Artificial Intelligence for Applications.*

Zadeh, L. (1971) Similarity relations and fuzzy orderings. Inf. Sci., 3, 177–200.