Denver Groups Classification of Human Chromosomes Using CANN Teams Supplemented by a Nearest Neighbor Technique CANNT-S

SadinaGagula-Palalic, Mehmet Can

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina

Article Info

Abstract

Human chromosomes can be classified into seven Denver groups (A – G) based on the position of the centromere. This classification is an important stage of human chromosome classification, as its output influence the second stage of classification, the correct classification of 24 classes of human chromosomes based on the banding pattern.In this article, the novel artificial neural network committee machines technique (CANNT) developed earlier is supplemented by a nearest neighbor technique, CANNT-S, and the correct classification rate in Denver Groups Classification of Human Chromosomes rose from 96% to a level of 98%.

1. INTRODUCTION

Human chromosome analysis plays an important role in medical diagnostics since any disorder or abnormality in the chromosomes of the cell may be a powerful indicator in diagnosis of leukemia, skin and breast cancers, Down's syndrome and other genetic diseases. For the chromosomes to be analyzable, a sample of blood smear or amniotic fluid must be cultured and fixed in the stage of metaphase. Then, by standard, cytologist must find twenty cells in the sample and make their karyograms which is the image of the correctly classified chromosomes.

Healthy human cell consists of 46 chromosomes, out of which are 22 pair of autosomes and two sex chromosomes, which come in pair as XX or XY. Thus there are 24 classes of chromosomes in total (22+X+Y). Fig. 1 shows the healthy female cell and its karyogram.

If performed manually, human chromosome karyotyping is very tedious and time consuming job. Computer aided system in great helps the speed of the karyotyping and classification. Since 1980s, automated chromosome classification has been extensively researched area. While automatic metaphase finding, automatic segmentation and feature extraction are more challenging areas, most of the researchers were dealing with the classification stage of the chromosome analysis only. As there are publicly available databases containing chromosome feature sets (among them the most famous ones are Copenhagen, Edinburg, Philadelphia datasets), many articles proposed different algorithms and classifiers. Among them, Artificial Neural Networks (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 (Mitchel, 1997; Haykin, 2009).
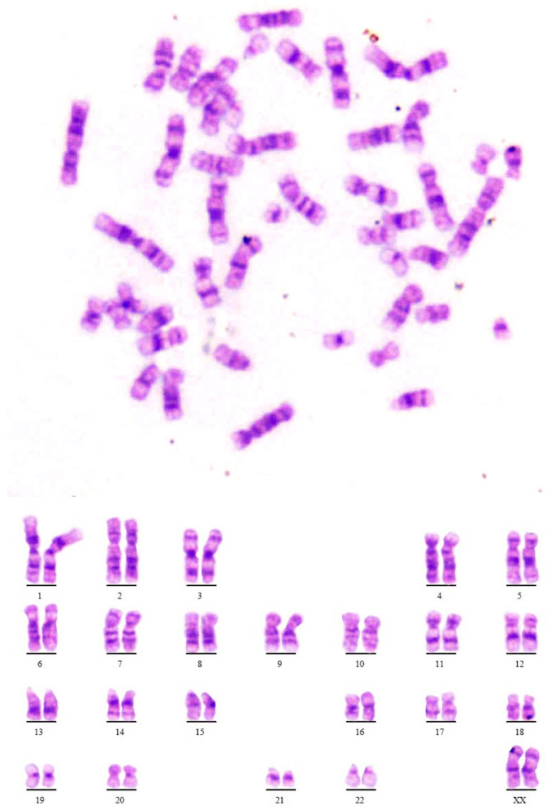
Figure 1.Human Metaphase Cell and its Karyogram.

A large number of different ANNs have been tested in classification of human chromosomes, which include supervised neural network architecture. Multi-layer neural networks are studied in (Lu and Ya, 1989; Wu et al., 1989; Erington and Graham 1993; El Emary, 2006; Wang et al., 2009; Can and Palalic, 2012) and Hopfield network (Ruan, 2000); fuzzy neural techniques (Ruspini, 1973ab; Ramstein et al., 1992; Keller et al., 1995; Sjahputera et al. 1999); and unsupervised architecture of nonlinear maps (Sarosa et al, 2005), self-organizing feature maps (Kyan et al. 1999) and mutual information maximization based training method (Mousavi et al., 1999).

In chromosome classification and pairing, back propagation training method is used to train ANNs. In multi-layer feed-forward ANNs, the number of output neurons is equal to the number of human chromosome types. The number of input neurons is equal to the dimension of the input data, which is the number of features used for classification. The number of hidden layers, number of hidden neurons, steepness of the activation function, learning rate, and momentum factor, number of learning iterations and upper bound of training error are chosen by the user experimentally. While the proper choice of these parameters is important for the performance and robustness of an ANN used in

chromosome classification (Cho, 2000), studies indicated that ANN performance was slightly lower than that obtained using simpler statistical methods (Granum and Thomason, 1990; Sweeney and Mousavi, 1993; Conroy et al., 2000). Unnecessary complexity of the ANN architecture and overtraining of ANNs dramatically reduce the robustness of the ANN in chromosome classification. One study (Mitchell, 1997) using multi-layer perception based ANN obtained 0% error rate in the training data set but 24.2% error rate in the testing data set. To increase ANN performance, another study showed that by reducing the complexity of an ANN, its testing accuracy can be increased from 75.8% to 88.3% [22]. Recently one study showed that by using two-layer ANN, correct classification rate of 93.7% can be achieved. In the first layer of the proposed method, a single ANN was employed to classify 24 chromosomes into seven classes. In the second layer, seven ANNs were adaptively optimized (using training-testing-validation) for seven classes to identify individual chromosomes (Delshadpour, 2003).

One of the other more sophisticated neural networks proposed and tested in this area is a fuzzy Hopfield neural network. It holds fuzzy clustering capability and learning mechanism of acquiring knowledge about the human chromosomes from noisy inputs. In a test involving 100 human chromosomes Ruan (Ruan, 2000) succeeded to achieve a very high identification rate of 96.67%.

Recently Palalic and Can, (Gagula-Palalic and Can, 2013) developed a novel committee of neural network machines, competing artificial neural network teams technique (CANNT) which over scores almost all previous human chromosome classifiers.
The rest of the article is organized as follows. In the second part, dataset used in the experiment is described. Third part gives the classification results when nearest neighbor (NN) applied. In the fourth section a new method called competing ANNs is applied to the data set and the correct classification rate in the training, validation and testing stages are shown. Part 5 describes the improvement of the results obtained in the part 4 when mixed signals are used. The article finishes with conclusion and discussion on the results.

2. CHROMOSOMES DATA SET

According to the Denver system of chromosome classification, chromosomes can be classified into seven 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). Based on these two features we differentiate metacentric, submetacentric and acrocentric chromosomes (see Table 1). The classification of chromosomes into

these 7 groups is the first stage of the classification process and its performance merely influences the correct chromosome classification into 22 classes, which is the second stage of the classification process. In this article, only the first stage, classification into seven Denver classes is performed.

Table 1: The classification of chromosomes based on Denver System

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

The data used in this work is taken from the Copenhagen data base. We omitted gray level features, and only kept (1) the length of each chromosome and (2) the centromere index (CI).
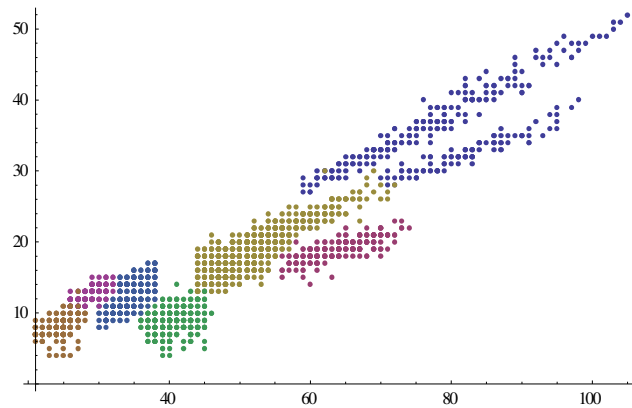

Fig 2.The distribution of 2200 human chromosomes into seven Denver Group classes from A, to G.

3. NEAREST NEIGHBOR CLASSIFIER

Input data is two dimensional and they are in two dimensional clusters as seen in Figure 2. A natural classifier for such a set is nearest neighbor technique.

From each of seven Denver class, 100 chromosome chosen as training set, 100 chromosome chosen as testing set. Test data is classified according to the Euclidean distance to the seven training clusters. The correct classification rates of seven test clusters are as in the Table 2.

Table 2: The classification of Denver Groups by nearest neighbor technique.

Denver Class	Correct %
1	100
2	100
3	97
4	97
5	90
6	99
7	97
Average	96.14

4.COMPETING ARTIFICIAL NEURAL NETWORK (CANNT)

Architecture of ANN

We represent the network consisting of 2 inputs x[i], i=1, 2, 12 neurons in the hidden layer and one neuron in the output layer as shown in the Fig 1. A special organized committee of 42 simple perceptrons is used to improve the rate of correct classification of 7 types of unbanded human chromosomes. Each of these simple perceptrons is trained to distinguish between two types of chromosomes. These multilayer perceptrons use Back-Propagation algorithm.
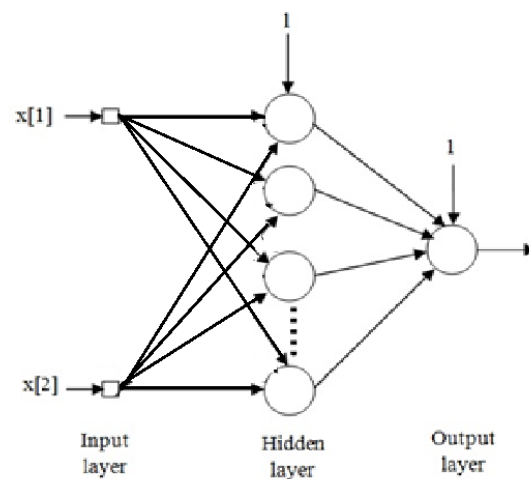

Fig 2: Neural network architecture for a simple multilayer perceptron

Assembling votes

Let $P(i, j)$ be the simple perceptron which is trained to distinguish chromosomes of type i, and of type j, and let

$$T(s) = \{P(i, s), P(s, j) \mid i = 1 \ldots 7, j = 1 \ldots 7\}, s = 1 \ldots 7$$

be the seven teams of perceptrons.

When a new data of type x enters the network, the perceptrons

$$P(i, x), \qquad i = 1 \ldots 7$$

of the team $T(x)$ creates mostly an output 1, while the perceptrons

$$P(x, j), \qquad j = 1 \ldots 7$$

of the same team $T(x)$ creates mostly an output -1. The perceptrons of other teams also creates outputs either 1, or -1. But since the other teams are not trained to distinguish chromosomes of type x from other chromosome types, their consensus will be weaker than the consensus of team $T(x)$. So we expect that the team $T(x)$ will be the winner of the competition.

For completeness, the dummy perceptrons $P(j, j), j = 1, 2, \ldots, 7$ which always give output 0 are added. When 7×7 perceptrons are arranged as a 7×7 grid, the votes of teams appear in crosses:

Fig. 3: Example of the decision matrix. The team $T(3)$ is the winner of the competition. The nearest competitor is $T(4)$.

The score of each team is its distance to its consensus. In Figure 3, the score of the team $T(3)$ is zero, while the score of nearest competitor $T(4)$ is four. The team with smallest score is the winner of the competition, and the new chromosome data entered, belongs to the chromosome type of winners label.

Another representation of the winner team can be visualized attaching gray levels to the team members proportional to their scores as seen in Figure 4:
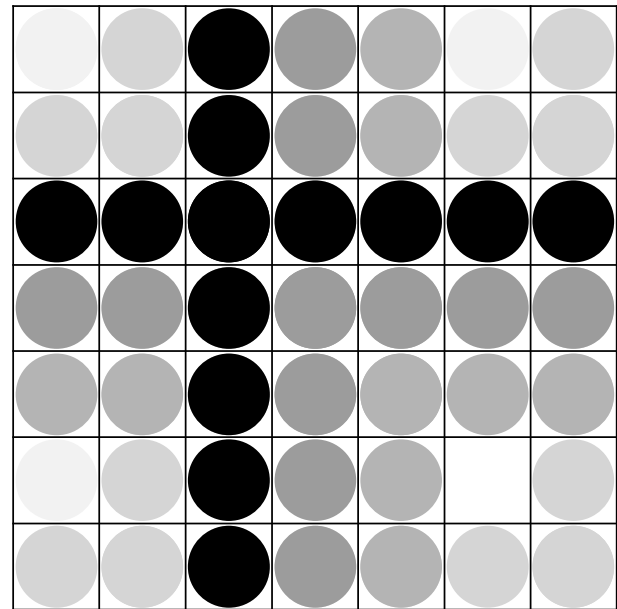

Fig 4: Competing teams. The darkest cross is the one which consist of 3rd row and 3rd column that wins the competition. The nearest competitor to team 3 is team 4.

5. RESULTS

During the training of 462 simple multilayer perceptrons, it is possible to complete training with zero error. But this leads to overtraining that causes lower rates in testing. From each chromosome type 50 random samples are chosen for training. The same numbers of random samples are also chosen for validation and testing. We have seen that it is possible to go over 97% correct classification rates with this special committee of perceptrons.

Table 1. Correct classification rates during training and testing. Using a validation data set, the overtraining is prevented.

Denver group	Correct Classification Rate (%)		
	Training	Validating	Testing
A	98	100	100
B	100	100	98
C	96	98	92
D	100	98	100
E	100	96	94
F	92	96	100
G	94	98	88
Average	97.14	98.00	96.29

6. CONCLUSION

In this study we presented a special organized committee of 42 simple perceptrons used to improve the rate of correct classification of 7 Denver types of unbanded human chromosomes. Each of these simple perceptrons is trained to distinguish between two types of chromosomes. When a new data is entered, the votes of these 42 simple perceptrons and additional 7 dummy perceptrons create a decision matrix of the size 7×7. By a special assembling of these votes we get a higher rate of correct classification of 7 Denver types of human chromosomes, with an average of 98.00% correct classification when tested on Copenhagen Chromosome Dataset.

REFERENCES

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

Cho J.M. (2000) Chromosome Classification using back-propagation neural networks, IEEE Engineering Medicine and Biology, January/February 2000, pp. 28-34

Conroy J.M., Kolda T.G., O'Leary D.P., O'Leary T.J. (2000) Chromosome Identification using Hidden Markov Models: Comparison with Neural Networks, Singular Value Decomposition, Principal Components Analysis and Fisher Discriminant Analysis, Laboratory Investigation, Vol. 80, No. 11, pp. 1629-1641

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).

El Emary I. M. (2006) On the application of artificial neural networks in analyzing and classifying the human chromosomes, Journal of Computer Science Vol. 2, No.1, pp. 72-75

Erington P.A., Graham J. (1993) Classification of Chromosomes using a combination of neural networks, IEEE International Conference on Neural Networks, 28 Mar – 1 Apr, San Francisco, Vol. 3, p.p. 1236-1241

Gagula-Palalic S, and M. Can (2013) An Organized Committee of Artificial Neural Networks in the Classification of Human Chromosomes International Journal of Computer Applications, Vol. 80, No. 8, pp. 38-41

Granum E., Thomason M. G. (1990) Automatically inferred Markov Network Models for Classification of Chromosomal Band Pattern Structure, Cytometry, Vol. 11, pp. 26-39

Haykin S. (2009) Neural Networks and Machine Learning, Pearson, Third edition, pp. 153-205

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.

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.

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. (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.

Sarosa M., Ahmad A.S., Riyanto B., Noer A.S. (2007) Optimization of Neuro-Fuzzy System using genetic algorithm for chromosome classification, ITB J. ICT, Vol 1 C, No. 1, pp. 56-69

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

Wang X., Zheng B., Li S., Mulvihill J.J., Wood M.C., Liu H. (2009) Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme, Journal of Biomedical Informatics 42, pp. 22–31

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.