Handwriting analysis for reduction of Juvenile Crime using Machine Learning

[1]Mrs.K.Valli Madhavi, Associate Professor,
[2] Mrs.R.Tamilkodi, Associate Professor
[3] Ms.Neha Nair, Assistant Professor
[1]Department of Computer Science
[1]Godavari Institute of Engineering & Technology, Rajahmundry, India.

Abstract: **Handwriting is often called mind indicting or encephalon inscribing. All that is a component of the mind is reflected by an individual in many ways, inscribing being one. Handwriting reveals the true personality including emotional outlay, fears, veracity, bulwarks and many others. Data such as the dynamically captured direction, stroke, distance, size, pressure and shape of an individual's handwriting enable to be a reliable designator of an individual's identity. This paper exposes the reduction of juvenile malefaction by analyzing the handwriting. The objective of this paper is to discuss methodology to prognosticate the demeanor of a person from the baseline and suggest him/her for desideratum of counseling. The digital handwritings accumulated are processed utilizing segmentation and trained utilizing back propagation algorithm and conclusively machine learning algorithm Decision Tree i.e. Classification and Regression Tree(CART) is applied to presage the desideratum of counseling to reduce the juvenile crime.**

Index Terms: **Handwriting, juvenile crime, segmentation, back propagation, machine learning, counseling, Decision Tree, Classification, Regression**

I. INTRODUCTION

Graphology is the study of handwriting. It is a scientific method of evaluating, and withal understanding a person's personality by identifying the strokes and patterns revealed by his handwriting. Handwriting is apperceived as being unique to each individual. Irrespective of the fact whether the person has inscribed with his hand, foot or mouth his handwriting will be identically tantamount and unique. The handwriting is done by the encephalon and not by the hand or by the feet. Hence handwriting is withal kenned as "brain writing". Some scientists in the neuromuscular field of research state that some diminutive neuromuscular forms of kineticism are associated to the person's personality. Each trait of personality is shown by a neurological encephalon pattern. A unique neuromuscular kineticism is engendered by each neurological encephalon pattern which is kindred for every person who has that personality trait. These minuscule forms of kineticism occur insensately while inscribing. Each stroke or indicted kineticism reveals a categorical personality trait. Graphology is the discipline of identifying these strokes as they appear in handwriting and describe the corresponding personality trait.

Handwriting can reveal a number of elements of the person's deportment, character or personality. This additionally gives us some glimpse about the person's astuteness, his emotional responsiveness and energy, his bulwarks and fears, his motivation, integrity and imaginative power and his aptitude. In this paper, a method has been proposed to predict the deportment of a person from the features extracted from his handwriting and suggest him for counseling. Few parameters, baseline, slant, pen pressure etc. are input to the Neural Network which outputs the personality trait of the essayist. Classification and Regression Tree is applied by identifying the independent and dependent.

II. SEGMENTATION

Image segmentation is a process in which regions or features sharing kindred characteristics are identified and grouped together. Image segmentation may use statistical classification, thresholding, edge detection, region detection, or any amalgamation of these techniques. The output of the segmentation step is conventionally a set of classified elements. Edge-predicated techniques rely on discontinuities in image values between distinct regions, and the goal of the segmentation algorithm is to accurately demarcate the boundary disuniting these regions. Segmentation is a process of extracting and representing information from an image is to group pixels together into regions of associated attribute. In this paper we move with the fusion of Edge Detection and Region Detection. Extracted edge information is used within region segmentation algorithm.

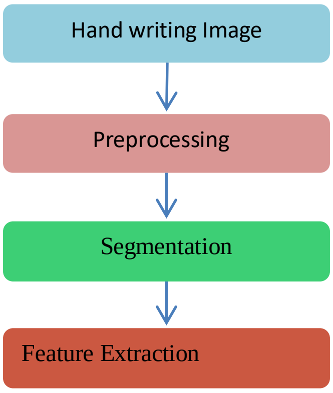


Fig. 1 Flow Chart for Feature Extraction

2.1 Edge Information can be used in two ways

Control of decision criterion –Including the edge information in the definition of decision criterion controls the growth of the region.

Seed placement guidance - edge information helps to decide which is the most appropriate position to place the seed of the region in region growing process.

2.2 Architecture

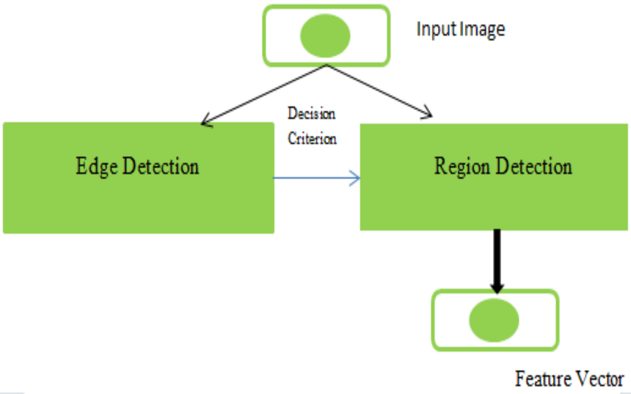


Fig. 2 Feature Extraction Architecture

Algorithm

- Select a starting pixel
- Taking into consideration homogeneity criterion add neighboring pixels that are similar
- Existence of pixel in growing region is determined based on criterion
- Region growing ends if an edge is encountered
- Merge if no edge is detected

2.3 Edge Based Segmentation method

A connected pixel that is found on the boundary of the region is called an edge. So these pixels on an edge are kenned as edge points. Edge can be calculated by finding the derivative of an image function. Some edges are very simplistic to find. These are: Ramp edge, Step edge, Roof edge, Spike edge. Step edge is an abrupt transmutation in intensity level. Ramp edge is a gradual vicissitude in intensity. Spike edge is an expeditious transmutation in intensity and after that returns immediately to a pristine intensity. Roof edge is not instantaneous over a short distance. Edge predicated image segmentation method falls under structural techniques.

2.4 Region Based Segmentation Method

This method is predicated on segmented an image on the substructure of kindred characteristics of the pixels.Region growing methods: The region growing predicated segmentation methods are the methods that segments the image into sundry regions predicated on the growing of seeds (initial pixels). These seeds can be culled manually (predicated on prior erudition) or automatically (predicated on particular application). Then the growing of seeds is controlled by connectivity between pixels and with the avail of the prior cognizance of quandary, this can be ceased. The rudimental algorithm predicated on 8-connectivity) steps for region growing method are:

If (x,y) is the original image that is to be segmented and s(x,y) is the binary image where the seeds are located. Let ‘T’ beany predicate which is to be tested for each (x,y) location.

• First of all, all the connected components of ‘s’ are eroded.

• Compute a binary image P_T. Where $P_T(x, y) = 1$, if $T(x, y) = \text{True}$.

• Compute a binary image ‘q’, where $q(x, y) = 1$, if $P_T(x, y) = 1$ and (x, y) is 8-connected to seed in ‘s’.

These connected components in ‘q’ are segmented regions.

III BACK PROPAGATION

The back propagation algorithm is a method for training the weights in a multilayer aliment-forward neural network. As such, it requires a network structure to be defined of one or more layers where one layer is entirely connected to the next layer. A standard network structure is one input layer, one hidden layer, and one output layer.

In relegation quandaries, best results are achieved when the network has one neuron in the output layer for each class value. For example, a 2-class or binary relegation quandary with the class values of A and B. These expected outputs would have to be transformed into binary vectors with one column for each class value. Such as [1, 0] and [0, 1] for A and B respectively. This is called a one hot encoding.

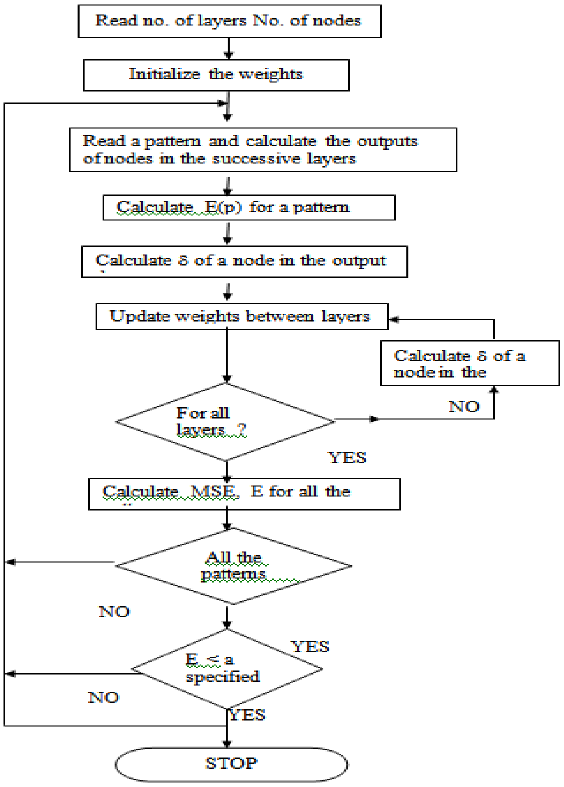

Read no. of layers No. of nodes
Initialize the weights
Read a pattern and calculate the outputs of nodes in the successive layers
Calculate E(p) for a pattern
Calculate δ of a node in the output
Update weights between layers
Calculate δ of a node in the
For all layers ?
NO
YES
Calculate MSE E for all the
All the patterns
NO
YES
E < a specified
NO
YES
STOP

Fig. 3 Flowchart for Back propagation

IV MACHINE LEARNING

Machine learning is utilized to edify machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the pattern or extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in elevates. Many industries from medicine to military apply machine learning to extract pertinent information. The significance of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves. Many mathematicians and programmers apply several approaches to find the solution of this problem.

Kinds of Machine Learning

There are three kinds of Machine Learning Algorithms.

a. Supervised Learning
b. Unsupervised Learning
c. Reinforcement Learning

4.1 Supervised Learning

A majority of practical machine learning uses supervised learning. In supervised learning, the system tries to learn from the previous examples that are given. (On the other hand, in unsupervised learning, the system attempts to find the patterns directly from the example given.)Speaking mathematically, supervised learning is where you have both input variables (x) and output variables(Y) and can use an algorithm to derive the mapping function from the input to the output. The mapping function is expressed as Y = f(X).Supervised learning problems can be further divided into two parts, namely classification, and regression.
Classification: A classification problem is when the output variable is a categorical or a group, such as “black” or “white” or “spam” and “no spam”. It gives binary values like either true or false or “0” or “1”.
Regression: A regression problem is when the output variable is a numeric value.

4.2 Unsupervised Learning

In unsupervised learning, the algorithms are left to themselves to discover interesting structures in the data. Mathematically, unsupervised learning is when you only have input data (X) and no corresponding output variables. This is called unsupervised learning because unlike supervised learning above, there are no given correct answers and the machine itself finds the answers. Unsupervised learning problems can be further divided into association and clustering problems. Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as “people that buy X also tend to buy Y”.Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

4.3 Reinforcement Learning

A computer program will interact with a dynamic environment in which it must perform a particular goal (such as playing a game with an opponent or driving a car). The program is provided feedback in terms of rewards and penalizations as it navigates its problem space. Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it continuously trains itself using trial and error method.

V METHODOLOGY

Classification algorithms are used when the desired output is a discrete label. In other words, they're ancillary when the answer to your question about your business falls under a finite set of possible outcomes. Many use cases, such as determining whether an electronic mail is spam or not, have only two possible outcomes. This is called binary classification. Given a data of attributes together with its classes, a decision tree creates a sequence of rules that can be habituated to classify the data.

5.1 Description

Decision Tree, as its name states, makes decision with tree-like model. It splits the sample into two or more homogeneous sets (leaves) grounded on the most principal differentiators in your input variables. To cull a differentiator (prognosticator), the algorithm considers all features and does a binary split on them (for categorical data, split by feline; for perpetual, pick a cut-off threshold). It will then operate the one with the least cost (i.e. highest precision), and reiterates recursively, until it prosperously splits the data in all leaves (or reaches the maximum depth).

5.2 Procedure

Step 1: Assign a feature to root node where the selected feature (predictor variable) classifies the data set into the desired classes.

Step 2: Make relevant decisions at each internal node by traversing down from the root node, such that each internal node best classifies the data.

Step 3: Repeat step 1 until you assign a class to the input data.

Step 4: Select the model.

Step 5: Set the model hyper parameters.

Step 6: Create a feature data set as well as a target array containing the labels for the instances.

Step 7: Fit the model to the training data.

Step 8: Use the fitted model on test data.

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of over fitting.

VI RESULTS

6.1 Model Evaluation

Consider a binary class problem (i.e. has only two classes: positive and negative), the output data of a classification model are the counts of correct and incorrect instances with respect to their previously known class. These counts are plotted in the confusion matrix as shown in table 1.

6.2 Confusion Matrix

Table1 Confusion Matrix

True Class	Predicted Class		
	Positive	Negative	
Positive	TP	FN	CN
Negative	FP	TN	CP
	RN	RP	N

As shown in table 1, TP (True Positives) is the number of instances that correctly presaged as positive class. FP (Erroneous Positives) represents instances prognosticated as positive while their true class is negative. The same applies for TN (True Negatives) and FN (Mendacious Negatives). The row totals, CN and CP, represent the number of true negative and positive instances and the column totals, RN and RP, are the number of soothsaid negative and positive instances respectively. Conclusively, N is the total number of instances in the dataset.

There are many evaluation measures used to evaluate the performance of the classifier predicated on its perplexity matrix resulted from testing. We will describe in more details some of the commonly used measures to be used later in our experiment.

Relegation Precision (Acc) is the most used measure that evaluates the efficacy of a classifier by its percentage of correctly prognosticated instances.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}}$$

Recall (R) and Precision (P) are measures that are based on confusion matrix data. Recall (R) is the portion of instances that have true positive class and are predicted as positive. On the other hand, Precision (P) is the probability of that a positive prediction is correct as shown in (6).

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{CN}} \text{ and } \mathrm{P} = \frac{\mathrm{TP}}{\mathrm{RN}}$$

VII Experiment

In this paper, we are provided by sample writings collected from students of various schools located across 8 districts of AP in three consecutive years (2016, 2017 and 2018). The dataset contains about 20262 records, while each record represents an instance with 4 attributes and the class attribute with two values: Counselling required and counselling not required. The classes are distributed as 53% of the total records for “Rejected” and 47% for “Accepted” class. Table 2 shows detailed information about datasets attributes.

Table2 Attributes Considered

Attribute	**Possible Values**
Gender	Male
	Female
School Type	Government
	Missionary
	Private
	International
Grade Gained	A, mark $\geq$ 85
	B,75 $\geq$ mark > 85
	C,65 $\geq$ mark > 75
	D,50 $\geq$ mark > 65
	E,35 $\geq$ mark > 50
	FAIL,mark < 35
Area	Pin code
Writing Category	Large
	Medium
	Small
Slant of words and letters	Right
	Left
	Vertical
Baseline	Raising
	Falling
	Straight
	Erratic
Pen Pressure	Light Pen
	Heavy Pen
Spacing between words and letters	Far
	Close

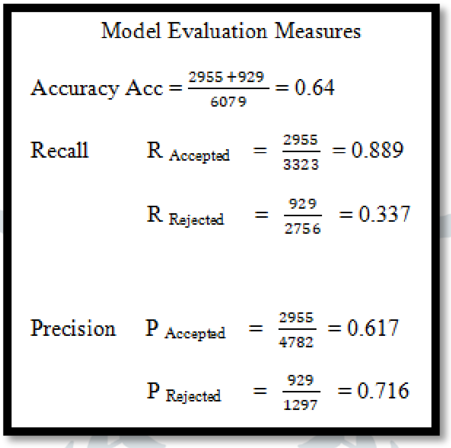
The dataset is divided into two main parts: training dataset that holds about 14183 records (about 70%) and testing dataset that contains about 6079 records (about 30%). The decision tree classifier is learnt using a training dataset and its performance is measured on not-seen-before testing datasets. Seriously

The decision tree model is generated over training dataset records using Orange data mining tool [10]. The generated decision tree is a binary tree with “One value against others” option. The confusion matrix values are shown in table 3. The values of confusion matrix are generated by applying a decision tree on testing datasets.

7.1 Testing Confusion Matrix

Table3 Confusion Matrix of Test Data

True Class	Predicated Class		
	Accepted	Rejected	
Accepted	2955	368	3323
Rejected	1827	929	2756
	4782	1297	6079



Model Evaluation Measures

Accuracy $Acc = \frac{2955+929}{6079} = 0.64$

Recall $R_{Accepted} = \frac{2955}{3323} = 0.889$

$R_{Rejected} = \frac{929}{2756} = 0.337$

Precision $P_{Accepted} = \frac{2955}{4782} = 0.617$

$P_{Rejected} = \frac{929}{1297} = 0.716$

Fig. 4 Evaluation Measures

VIII CONCLUSION

A relatively simpler method has been proposed to anticipate the requirement of counseling to minor boys and girls by exploring various handwriting features. The system considers five discriminating features such as writing category, pen pressure, baseline, slant of words and letters and spacing between words and letters. The proposed system can be used as a tool by school management to improve the accuracy and anticipate the requirement of counseling to minor boys and girls.

REFERENCES

[1]. Abdul Rahiman M, Diana Varghese, Manoj Kumar G, “HABIT- Handwriting Analysis Based Individualistic Traits Prediction”.

[2]. M. Welling, “A First Encounter with Machine Learning” M. Bowles, “Machine Learning in Python: Essential Techniques for Predictive Analytics”, John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2

[3]. S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268

[4] J. R. Quinlan, (1986), “Introduction of Decision Tree”, Machine Learning, vol. 1, pp. 86-106.

[5] K. Kira and L.A. Rendell (1992), “A practical approach to feature selection”, In D.Sleeman and P.Edwards, editors, Proceedings of International Conference on Machine Learning, pp. 249-256, Morgan Kaufmann.

[6] A.P. Bradley, (1997), “The use of the area under the roc curve in the evaluation of machine learning algorithms”, Pattern Recognition, vol. 30, pp. 1145-1159.

[7] K. Valli Madhavi, R. Tamilkodi, K. Jaya Sudha. "An Innovative Method for Retrieving Relevant Images by Getting the Top-ranked Images First Using Interactive Genetic Algorithm", Procedia Computer Science, 2016

[8] Sampurna Mandal, Supratim Bhattacharya, Jayanta Poray. "Towards a decision support system by the study of cell malfunctions for breast cancer", 2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2016