Revisiting Classification of Handwriting Deficiencies

Shirel Azogy and Ariella Richardson
Department of Industrial Engineering & Management, Lev Academic Center

Sara Rosenblum
The laboratory of Complex Human Activity and Participation (CHAP), Department of Occupational Therapy, University of Haifa

Introduction

Many people suffer from handwriting deficiencies. These deficiencies can be of various origins being reflected mainly by inferior writing speed, legibility or both [1]. The diagnosis of such problems is usually performed by trained occupational therapists using a set of Handwriting Evaluation tests [2]. The testing is subjective and limited to characteristics of the writing observable by humans [3]. An application that provides diagnosis would lower the costs of evaluation, provide support to inexperienced therapists, enable re-evaluation throughout therapy to test for improvement and is therefore an important contribution.

Disturbances in handwriting is one of the criteria for diagnosis of Developmental Coordination Disorder (DCD) according to the DSM5 [4]. DCD is a motor impairment that affects a subject's ability to perform the skilled movements necessary for daily living and among other things affects handwriting proficiency.

Problem Description

We used a tablet with special sensors to record handwriting features for multiple subjects. The data is recorded at set intervals, and at each time point several features are measured [5] [6]. Thus, the data is essentially a set of time-series with multiple attributes. This creates a matrix of data for each subject. In each matrix, a column represents one feature and a row represents one time point. Examples of features recorded are tilt, pressure and azimuth. Each matrix was labeled as belonging to a 'poor' or 'proficient' subject by a trained occupational therapist. There are 42 subjects in our dataset, 22 are 'poor' and 20 are labeled as 'proficient'.

Richardson et al. introduced COACH [7], and studied how to use Data Mining to classify handwriting deficiencies. They studied attributes such as the time between strokes, tilt, pressure and the azimuth of the pen on the paper as described in **Fig 1**. They used both standard methods such as decision trees, and an algorithm named COACH that they developed.

Several assumptions were made during feature selection in COACH such as feature vector length, division of the data into subsets and the use of leave one out cross validation. In this study, we re-examine some of these assumptions and try to improve classification results. In COACH, the feature vectors were composed of 200 values for the same attribute. For example, a feature vector for pressure is composed of up to 200 consecutive measurements of pressure levels in a single stroke. For strokes composed of more than 200 measurements, a new feature vector was initiated. Strokes shorter than 200 measurements were padded with zeros. Other assumptions made in COACH include the use of 10 strokes for each subject and using leave-one-out cross validation for evaluation.

Method

In this study, we expand the work in COACH by testing some new variations in the feature extraction process, the methods used and the evaluation measures reported. We will present results for variations in the following parameters: Feature vector length, the number of strokes collected from each subject and the evaluation/testing method. In this study, we used the same raw data as used in COACH.

Aside from varying the feature vector length and the number of strokes used for each subject, we made an important change to how feature vector length is defined. In COACH, strokes that had a length exceeding the selected vector length were split and treated as two strokes. This

meant that some of the feature vectors were generated from data that came from the beginning of the stroke until the middle of that stroke, and some were from the middle of the stroke until the end. Since we thought that this might hinder the classification, we decided to discard the measurements that exceed the feature vector length, thus ensuring that all vectors came from similar parts of the written strokes. These comparisons were all performed using leave-one-out cross validation as in COACH to allow comparison of these specific changes. The training was performed on data from all subjects excluding one, and then the model was tested on the one subject that was left out. This was repeated for all subjects and results were averaged over all runs.

We also decided to compare the use of leave-one-out cross validation (as described) to putting the data from all the subjects into a single pool, and using 10-fold cross validation. We repeated the experiments on feature vector length and the number of strokes used for the various validation methods. For each of the trials we performed we report various measurements, such as TP Rate, FT Rate, Precision, Recall and F-measure. In COACH *Pressure* was found to be the main contributing feature to successful classification, along with decision trees - using WEKA's J48 [8][9], therefor we focus on using decision trees and the pressure feature in this study as well.

Experimental Results

First, we present the results for using the full set of measurements in a stroke, such that strokes composed of more than 200 measurements are split into two strokes. Since most of the measurements improved by increasing the length of the feature vector, we concluded that using the *tails* (the measurements that occurred after the first 200 measurements) when we split the stroke into two features was disrupting the classification. Moreover, by increasing the number of strokes the results also increased significantly. Therefore, we proceed to run our experiments using the data without *tails* and using 50 strokes from each subject.

The results we achieved by dropping the *tails* support part of our assumptions (**Table 1**). The precision and the F-measure increase for both classes but the accuracy and the recall decreases. Therefore, when we compare the evaluation method used in COACH to 10-fold cross validation, we used the data without tails.

In this method, all of the measurements improve by increasing the number of strokes (**Table 2**). As the experiment without stroke splitting shows, increasing the number of strokes leads to better results than increasing the length of the feature vector. One explanation is that the proficient subjects have at most 200 measurements of pressure levels in a single stroke, so increasing the vector length produces zeros in the feature vectors for these subjects. As a result, increasing the vector length does not contribute significant information on proficient subjects; on the contrary, it only gives more weight to deficient subjects and hinders the classification.

Discussion and Future work

This research presents an extension of the study presented in [4]. We have shown that using Data mining on handwriting can diagnose DCD with more than a 72% success rate. These results are considered good in this domain, as labeling certainty is imperfect.

We have also shown that by making small changes in the feature vector length or the amount of strokes used we can significantly improve the results of the classification using decision trees. For example, we raise precision for the 'poor' subjects by more than 50% and the general success rate by more than 11.5%.

These results can be expanded to other deficiencies such as Alzheimer's and dysgraphia, which affect handwriting abilities; we plan to study these along with other approaches in the future.

Figures & Images

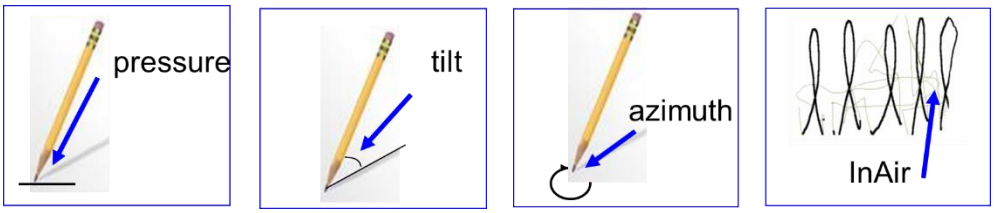

Fig 1: Handwriting's attributes that was studied at Richardson et al. research.

Tail	Record Length	Strokes	Precision	Recall	F-Measure	Accuracy
Yes	200	10	0.500	0.330	0.380	0.6406
	250	10	0.500	0.335	0.383	0.6719
	200	50	0.524	0.327	0.383	0.6756
No	200	10	0.476	0.333	0.376	0.7140
	200	50	0.524	0.331	0.381	0.6821
	350	50	0.524	0.330	0.381	0.6821

Table 1: Results for leave-one-out, "poor" class.

Record Length	Strokes	Precision	Recall	F-Measure	Accuracy
200/400	10	0.686	0.652	0.668	0.6614
200	50	0.774	0.660	0.712	0.7239
400	50	0.771	0.658	0.710	0.7219

Table 2: Results for 10-fold cross validation, "poor" class

References

1- Rosenblum, S., Weiss, P. L., & Parush, S. (2003). Product and process evaluation of handwriting difficulties: A review. Educational Psychology Review, 15, 41-81

2- Erez, N., and Parush, S. 1999. The Hebrew Handwriting Evaluation (2nded.). Israel, Jerusalem: School of Occupational Therapy. Faculty of Medicine. Hebrew University of Jerusalem.

3- Bahlmann, C. 2006. Directional features in online handwriting recognition. Pattern Recogn. 39(1):115-125.

4- American Psychiatric Association (APA) (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author.

5- Baxter, R. A.; Williams, G. J.; and He, H. 2001. Feature selection for temporal health records. Lecture Notes in Computer Science2035:198-209.

6- Rosenblum,S.;Weiss,P.L.;andParush,S.2003.Computerized temporal handwriting characteristics of proficient and non-proficient handwriters. The American Journal of Occupational Therapy 57(2):129– 138.

7- Richardson, A., Kraus, S., Weiss, P. L., & Rosenblum, S. (2008, July). COACH-Cumulative Online Algorithm for Classification of Handwriting Deficiencies. In AAAI (pp. 1725-1730).

8- WEKA Official Website, http://www.cs.waikato.ac.nz/~ml/weka/index.html , 5/2016

9- Witten, I. H., and Frank, E. 2005. Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2nd edition.