HUMANS DIFFER: SO SHOULD MODELS
Systematic Differences Call for Per-subject Modeling

Wolfgang Heidl, Stefan Thumfart and Christian Eitzinger
Profactor GmbH, Im Stadtgut A2, 4407 Steyr-Gleink, Austria

Keywords: Machine learning, Human diversity.

Abstract: While machine learning is most often learning from humans, training data is still considered to originate from a uniform black box. Under this paradigm systematic differences in training provided by multiple subjects are translated into unavoidable modeling error. When trained on a per-subject basis those differences indeed translate to systematic differences in the resulting model structure. We feel that the goal of creating human-like capabilities or behavior in artificial systems can only be achieved if the diversity of humans is adequately considered.

1 INTRODUCTION

Machine learning (ML) is often (or almost exclusively) focused on reproducing human cognitive abilities. “Learning” thus typically means “learning from a human”. Aside from a few examples, ML methods seem to ignore the fact that individuals are different and that this may also reflect in the ML structures used to reproduce their behavior.

Our hypothesis is that systematic variations in human-trained ML structures do exist and that they correlate with individual properties such as age, sex, education or cultural background. We feel that the goal of creating human-like capabilities or behavior in artificial systems can only be achieved if the diversity of humans is adequately considered.

The current approach in ML is to ignore these differences and to average over the group of individuals that provide training input. This is particularly true for industrial installations of ML systems, where training input is provided by multiple experts and machine operators. However, there is little knowledge about what is lost by “averaging” over different (groups of) individuals and how well such average models capture the behavior of individuals.

In the literature there are a few isolated studies that deal with these issues. Preliminary research on a simulated high-school task (Stevens and Soller, 2005) has shown that when self-organizing maps are used to cluster problem solving strategies, they are able to identify structural differences between genders that are not present in the outcomes, and so would not be detected by existing methods for comparing and contrasting classifiers. Also, recent analyses of human problem solving behavior (Heidl et al., 2011) report that although there is no difference in the final performance between e.g. males and females, there are significant differences in the strategies used. In (Eitzinger et al., 2009) individual behavior is compared on a visual inspection task, where it is found that four different experts only agree in about 80% of the decisions and that an improvement compared to ground truth data may be achieved by using voting procedures and other classifiers that merge the results of the single experts. Combination methods can range from simple majority voting to optimizing the prediction error of weighted combinations on novel data (Donmez et al., 2010). This way, systematic differences between individuals are reflected properly in the resulting structures. Disagreement in the predictions of multiple models can even be used as an uncertainty measure of the overall prediction.

One should clearly distinguish this type of research from approaches that try to identify individuals using ML methods (Zhao et al., 2003). This is commonly done in bio-metrics e.g. by classifying fingerprints (Jain et al., 1999), capturing the dynamics of writing (Yu et al., 2004) or typing on a keyboard (Peacock et al., 2004). These approaches use ML to identify the individual based on behavioral or bio-metric data. Instead we are looking at the structural variations of ML systems that reproduce same human ability. Another related but structurally different problem is determining the influence of gender and other de-

Heidl W., Thumfart S. and Eitzinger C..
HUMANS DIFFER: SO SHOULD MODELS - Systematic Differences Call for Per-subject Modeling.
DOI: 10.5220/0003832904130418
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 413-418
ISBN: 978-989-8425-95-9

mographic properties in mined decision models that is researched in the area of discrimination discovery (Ruggieri et al., 2010). While this work investigates the influence of demographic properties on decisions over individuals, we are interested in the influence of those properties on models trained from the individuals.

In the following section we will describe a study based on a visual inspection task that reveals significant differences in the strategies used by male and female participants when solving the task. The results of this study provide evidence that systematic differences between individuals exist and that it may be worth to further investigate this topic. Once we accept that individual differences are actually reflected in ML systems, we need analysis methods that allow us to assess and quantify their significance. Based on these analysis tools we may extract ML structures that better fit to one group or another. This will enable us, in the long term, to have ML structures that need less training data and generalize better within a certain group of people.

2 A VISUAL INSPECTION STUDY

In search for systematic variations in induced decision models we choose to conduct a visual inspection experiment, where we assume that significant differences exist and are most probably correlated with subject sex (Heidl et al., 2010). This assumption is grounded in accounts from many industrial practitioners, stating that women are better suited for visual inspection and that they perform the task with higher accuracy, and better repeatability.

In our experiment subjects had to rate a set of images according to a predefined set of rules, the so-called *inspection standard*. From the images we extract characteristic features and use them as input together with the subject responses to train ML classifier instances. We hypothesize that the resulting ML structures vary systematically and that variations are correlated with subject sex.

2.1 Stimulus Material

The stimuli consist of synthetic images inspired by die-cast parts with a machined surface (Figure 1). Two machine holes are present on the parts to provide some spatial structure and to facilitate the definition of critical zones which are located around those holes. Each image shows the same parts, however three different types of faults can be present in each image:

- Scratches: bright arcs.
- Cavities: dark, elliptic spots with scraggly edges.
- Dirt Spots: clusters of dark discs.

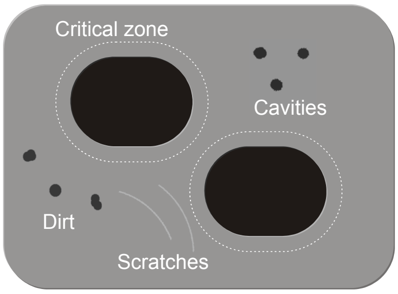

Figure 1: Stimuli images are based on stylized die-cast parts. Three types of faults may be present on the parts. The boundary of the critical zone and the labels are only given for reference and are not present in the actual stimuli.

Since we do not investigate the visual search task involved in inspection (Drury, 1978), faults are designed to be easily separable from the background. Decisions are to be made concerning fault size, fault position and fault type. The appearance of the fault type dirt and cavities is very similar to make their distinction non-trivial.

The inspection standard used in the experiments consists of seven rules. The relevant features for judging potential faults according to the inspection standard are size, position in relation to the critical zone (see Figure 1), distance to closest equal-type fault and the count of potential faults with different type. To avoid educational bias, the inspection standard was presented to subjects in a visual manner with examples.

2.2 Subjects

Fifty female and fifty male subjects were recruited through bulletins placed at adult education centers. The study was entitled “Perception Experiment” and participants have not been informed that gender differences are investigated. We have decided not to recruit people who work in visual inspection since we expect them to have a substantial preconditioning from their work experience. Graduates have also been excluded from the study to avoid gross mismatch to the typical education structure in visual inspection.

The mean age of the males was 29.7 years and the females 28.7 years. According to self-report, all subjects were in good health and free of any medications that could potentially affect cognitive performance. All subjects had normal or corrected to normal sight. Subjects have been compensated for travel costs and for their time taking part in the experiment.

2.3 Procedure

Experiments were carried out in 21 sessions with groups of four to six subjects. A video-taped briefing was used to eliminate variability due to different instructors and changes in reading speed or intonation. To ensure equal viewing conditions the video was displayed on each station screen. After a brief introduction consent forms have been distributed and signed by the subjects.

The inspection standard has been introduced to the subject by means of a six-minute slide-show with no audio. Two pages summarizing the inspection standard and providing a reference for fault sizes and distances were handed out. The subjects were asked to go over the cards and see if they have any questions. After 2 minutes the instructor answered open questions, then the experiment run started.

Within 30 minutes a total of 600 images should be inspected. An unpaced approach (Garrett et al., 2001) with slight adaption was taken. To reach the goal of 600 images a progress bar and a remaining time bar was displayed on each station screen. Subjects were encouraged to keep their progress bar in line with the remaining time bar.

The sessions ended with the completion of a questionnaire covering demographics, reflection on the computer experiment and career and gender role attitudes.

2.4 Machine Learning of Visual Inspection

In visual inspection tasks, ML classifiers are used for automating the process of finding a mapping between images and classes. This is achieved by capturing relevant image features and learning a suitable model to explain the decisions, acquired from one or several domain experts or operators during an annotation process.

During the visual inspection experiment subjects rate a set of images Ψ that are generated by taking i.i.d. samples from some distribution $\mathcal{D}_\Psi$. For each image the subjects give responses y, where

$$y = \begin{cases} -1 & \text{if shown part is accepted,} \\ 1 & \text{if shown part is rejected.} \end{cases}$$

Suppose we can characterize the features relevant to subject decisions by a d-vector $x \in \mathbb{R}^d$ and that these features can be extracted from the image ψ_i by some *extraction function* Φ such that $x_i = \Phi(\psi_i)$. If we can train a classifier f to produce predictions $\hat{y}_i = f(x_i, \theta)$ with zero expected error on new images, the identified parameters θ can be used as a perfect surrogate for the subject's decision behavior in the given task. The analysis of differences and similarities between the decision behavior of multiple subjects can then be based on the parameters θ identified for each subject (Figure 2). For most classifier types the number of parameters depends on the training data and may not allow for direct encoding of θ into vectors of equal length. To reach a fixed-length encoding and facilitate analysis and interpretation we describe the identified models by a set of *meta-features*. These meta-features should capture the relevant properties of the identified models and will be specific to the type of classifiers used.

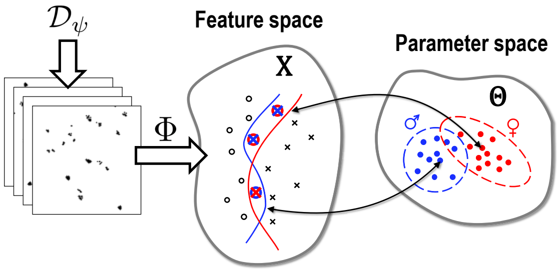

Figure 2: Classifiers trained on responses of different subjects. Each subject rates the same set of stimuli images sampled from $\mathcal{D}_\Psi$ and represented by points $x \in \mathbb{X}$ in feature space. Subject responses are indicated by × and ○ markers at those points. The decision boundary in $\mathbb{X}$ and corresponding classifier parameters $\theta \in \Theta$ represent these subjects' decision behaviors.

2.5 Classification Trees and Meta-features

In our study we used classification trees to model the subject decision behavior. These trees were induced by the CART (Breiman et al., 1993) algorithm. Such trees are *full binary trees*, where every node other than the leaves has exactly two children. Figure 3 shows a typical decision tree induced from the responses of one subject in our study.

As indicated in the previous subsection, trained classifiers may not in general be encoded into parameter vectors θ of fixed length, which makes classifier comparison a non-trivial task. This is particularly true for classification trees that can greatly vary in structure and the selection of features for the splits. Therefore, we introduce a fixed number of *meta-features* describing the tree structure.

The tree *size* N is given by the number of nodes including L leaf nodes, where $N = 2L - 1$. The *depth* d_i of a node n_i is the length of the path from the root to the node, with the maximum node depth being the tree *height* h. For trees that model decision behavior,

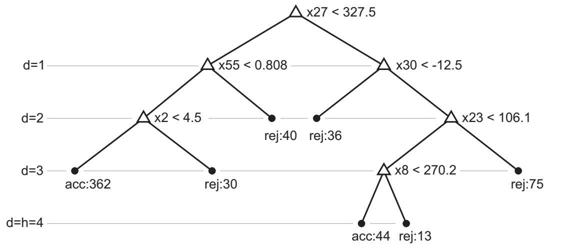

Figure 3: Decision tree induced from the responses of one subject. Triangles denote splits according to the criterion given next to it. Each leaf node is marked by a filled circle and the decision (accept/reject) associated to it. After the colon the number of samples ending up in the leaf node is given. The tree is displayed in terms of levels with equal *depth d*, with the *height* $h = 4$ being the maximum depth.

the depth at the leaf nodes can be interpreted as effort needed to come to a decision. In addition to L and h, which depend on the graph structure alone, we can also take into account how the samples traverse the tree. If we count the number l_i of training instances traversing node n_i, and denote the set of leaf nodes $\{n_i | i \in \mathcal{L}\}$, we can compute the *average depth per sample*

$$\tilde{\mu}_d = \frac{\sum_{i \in \mathcal{L}} d_i l_i}{\sum_{i \in \mathcal{L}} l_i}\,, \tag{1}$$

where $\mathcal{L}$ is the set of leaf node indices. Similarly, we define the *relative depth variability* $\tilde{\sigma}'_d$ as

$$\tilde{\sigma}'_d = \frac{\tilde{\sigma}_d}{\tilde{\mu}_d}, \quad \text{with} \quad \tilde{\sigma}_d = \sqrt{\frac{\sum_{i \in \mathcal{L}} (d_i - \tilde{\mu}_d)^2 l_i}{\sum_{i \in \mathcal{L}} l_i}}\,. \tag{2}$$

By taking the number of traversing training instances into account we define the *tree entropy*

$$H = \sum_{i \in \mathcal{L}} H_i \tag{3}$$

with the entropy contributions of each (leaf)node

$$H_i = -p_i \log_2 p_i\,, \quad p_i = \frac{l_i}{l}\,. \tag{4}$$

2.6 Results

In this section we present the analysis results of our visual inspection experiment with 50 female and 50 male subjects. As indicated before this study was primarily targeted towards identifying gender-differences. Clearly, any significant differences found correspond to correlations to a "measurable" human property and thus explain part of the variance in the measured human behavior.

Table 1 and Table 2 report mean values for all, female and male subjects along with their standard deviations in braces. Additionally the effect size (and significance level in braces) of the gender differences are given. The effect size is defined as the difference between group means μ_M and μ_F, normalized by their average standard deviation σ' (Cohen, 1988). Statistical significance is assessed by running permutation tests. In the tables significant effect sizes are written in boldface subject to a significance level of $\alpha = 0.05$.

Table 1: Overall performance of subjects.

Performance measure	**Mean value (standard deviation)**			**Effect size (p-value)**
	All	Female	Male	
Accuracy	0.741 (0.059)	0.747 (0.061)	0.736 (0.056)	-0.192 (0.340)
False alarms	0.102 (0.054)	0.113 (0.057)	0.091 (0.048)	**-0.414** **(0.041)**
Misses	0.157 (0.054)	0.140 (0.051)	0.173 (0.052)	**0.649** **(0.002)**

2.6.1 Subject Performance

In Table 1 we summarize the performance of subjects in terms of accuracy, miss rate, and false alarm rate with respect to the inspection standard. While the accuracy, i.e. the rate of correct responses shows no significant difference between female and male subjects ($p = 0.340$), we have observed significant differences in the false alarm ($p = 0.041$) and miss rates ($p = 0.002$). Male subject on average miss 35% of nonconforming parts[1] while the figure for female subjects is only 28%. Conversely, female subject falsely reject 23% of conforming parts compared to 18% for male subjects.

2.6.2 Group Response Profiles

We analyze differences in the average response behavior of subjects based on male and female response profiles. These response profiles were computed by taking majority votes on each sample from the male and female subjects, respectively. From the 600 responses 9.2% differed between the male and female profile. These differences were statistically significant ($p < 0.0007$).

2.6.3 Classifier Structure

We analyze differences in the structure of the identified subjective classification trees and the importance of input features by means of the tree meta-features

[1] A miss rate of 0.173 on all parts corresponds to 34.6% of the 50% nonconforming ones.

Table 2: Structural meta-features of subjective classification trees.

Meta feature	**Mean value** **(standard deviation)**			**Effect size** **(p-value)**
	All	Female	Male	
Leaf count	5.960 (2.964)	6.820 (3.397)	5.100 (2.169)	**-0.604** **(0.00158)**
Tree height	4.340 (1.765)	5.020 (1.813)	3.660 (1.437)	**-0.831** **(0.00009)**
Tree entropy	1.830 (0.565)	2.045 (0.511)	1.615 (0.538)	**-0.821** **(0.00007)**
Average depth per sample	3.085 (0.922)	3.389 (0.898)	2.781 (0.850)	**-0.696** **(0.00077)**
Relative depth variability	0.364 (0.132)	0.416 (0.119)	0.312 (0.125)	**-0.846** **(0.00005)**

defined in Section 2.5. Table 2 shows that all meta-features related to the tree structure show significant gender-differences.

In general, trees induced from the responses of female subjects are larger and more complex than those induced from male subjects, with a 29% ($p = 0.002$) difference in average leaf count and 31% ($p = 9 \times 10^{-5}$) in tree height with respect to their average values. The average entropy is 2.045 bits for trees induced from female subject responses versus 1.615 bits for ”male” trees. Most prominent is the difference in *relative depth variability* (see (2)) with ($d = -0.846$, $p = 5 \times 10^{-5}$).

3 CONCLUSIONS

In the previous section we have established the fact that individual differences (in this case gender) reflect in machine learning structures and that these differences are significant. It is particularly remarkable that the structural differences do not correspond to differences in performance. It is really only the problem solving strategy that differs. Such differences in cognitive approaches also exist in other tasks, e.g. in problem space navigation (Stevens and Soller, 2005) or virtual maze navigation (Moffat et al., 1998). While all those studies are focused on gender differences, we believe that correlations also exist along other social, cultural or biological dimensions. For example, in our visual inspection study significant correlations exist between induced classifier structure and subject's self assessment on their leadership qualities and intelligence.

Up to now it is not clear whether these results carry over to a wider range of machine learning problems. It should be noted that the above study relates to a comparably well-defined task, where individuals were given clear instructions what to do. We may assume that individual differences will be more pronounced in tasks that lack clear rules and put more emphasis on subjective behavior, such as e.g. judging aesthetics (Thumfart et al., 2011). This is clearly an open, but promising research question. Furthermore, there is a lack of machine learning databases that include information about how the training data were created, in particular whether the ground truth data were generated by one or more individuals. This information should be included in databases to allow an assessment of individual differences and to quantify what is lost by averaging over all the individual trainers.

We believe that ideally, training of machine learning structures should be performed on a per-subject basis. If training input from multiple subjects is treated as a uniform data set, systematic differences between subjects cannot be resolved. Those differences will appear as unresolvable conflicts in the data and lead to unavoidable modeling error. We propose training from multiple subjects should be combined only at the output stage of individually trained machine learning structures. By making the diversity of trainers explicit, this approach not only accommodates the potentially conflicting data of individuals, but also allows for improved system performance. Indeed, instead of mere majority voting of individual models for the overall system output, weighted combinations can emphasize reliable, consistent trainers. The weights need not be set a-priory but can be determined automatically from estimates of the prediction error on unlabeled and thus impartial data (Donmez et al., 2010). The weights can either be based on individual expected errors or determined in a joint optimization procedure guided by the expected error of the combined vote.

Clearly, segregating an otherwise larger data set into smaller per-subject chunks could lead to higher prediction error and possibly to over-fitting of the individual structures. However, the success of Random Forests (Breiman, 2001) has shown, that the combination of classifiers trained on independent (or independently sampled) subsets of data can rival and even surpass other state-of-the-art models trained on the whole set.

Once we have learned more about how such difference reflect in machine learning systems, we may be able to judge the validity of particular models for a particular task and (group of) trainer(s). Machine learning methods could be biased to favor models that are more likely to reproduce the behavior in simple structures and thus improve training efficiency and performance.

Our main conclusion is that research in artificial intelligence should be aware that there is no single 'correct' machine learning structure for particular task and that the results obtained may be substantially influenced by the individual that is modeled in this structure.

ACKNOWLEDGEMENTS

This work has been supported by the FEMtech program of The Federal Ministry for Transport, Innovation and Technology under FFG grant No. 318113. It reflects only the authors' views.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1993). *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, FL.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, 2 edition.

Donmez, P., Lebanon, G., and Balasubramanian, K. (2010). Unsupervised supervised learning i: Estimating classification and regression errors without labels. *J. Mach. Learn. Res.*, 11:1323–1351.

Drury, C. G. (1978). Integrating human factors models into statistical quality control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 20(12):561–572.

Eitzinger, C., Heidl, W., Lughofer, E., Raiser, S., Smith, J., Tahir, M., Sannen, D., and Van Brussel, H. (2009). Assessment of the influence of adaptive components in trainable surface inspection systems. *Machine Vision and Applications*, 21(5):613–626.

Garrett, S. K., Melloy, B. J., and Gramopadhye, A. (2001). The effects of per-lot and per-item pacing on inspection performance. *International Journal of Industrial Ergonomics*, 27(5):291–302.

Heidl, W., Thumfart, S., Lughofer, E., Eitzinger, C., and Klement, E. P. (2010). Classifier-based analysis of visual inspection: Gender differences in decision-making. In *Proceedings of SMC2010, IEEE Conference on Systems, Man and Cybernetics*, pages 113–120.

Heidl, W., Thumfart, S., Lughofer, E., Eitzinger, C., and Klement, E. P. (2011). Decision tree-based analysis suggests structural gender differences in visual inspection. In *Proceedings of AIA2011, IASTED International Conference on Artificial Intelligence and Applications*, pages 142–149.

Jain, A. K., Prabhakar, S., and Hong, L. (1999). A multichannel approach to fingerprint classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):348–359.

Moffat, S. D., Hampson, E., and Hatzipantelis, M. (1998). Navigation in a "virtual" maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior*, 19(2):73–87.

Peacock, A., Ke, X., and Wilkerson, M. (2004). Typing patterns: a key to user identification. *Security Privacy, IEEE*, 2(5):40–47.

Ruggieri, S., Pedreschi, D., and Turini, F. (2010). Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data*, 4(2):9:1–9:40.

Stevens, R. and Soller, A. (2005). Machine learning models of problem space navigation: The influence of gender. *Computer Science and Information Systems/ComSIS*, 2(2):83–98.

Thumfart, S., Jacobs, R. A., Lughofer, E., Eitzinger, C., Cornelissen, F. W., Groißböck, W., and Richter, R. (2011). Modelling human aesthetic perception of visual textures. *Accepted for publication in ACM Trans. on Applied Perception*.

Yu, K., Wang, Y., and Tan, T. (2004). Writer identification using dynamic features. In Zhang, D. and Jain, A. K., editors, *Biometric Authentication*, volume 3072 of *Lecture Notes in Computer Science*, pages 1–8. Springer Berlin / Heidelberg.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458.