2018 Annual American Control Conference (ACC)
June 27–29, 2018. Wisconsin Center, Milwaukee, USA

Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation

Kumar Akash, Tahira Reid, and Neera Jain

Abstract— **Classification algorithms have traditionally been developed based on the assumption of independent data samples characterized by a stationary distribution. However, some data types, including human-subject data, typically do not satisfy the aforementioned assumptions. This is relevant given the growing need for models of human behavior (as they relate to research in human-machine interaction). In this paper, we propose an adaptive probabilistic classification algorithm using a generative model. We model the prior probabilities using a Markov decision process to incorporate temporal dynamics. The conditional probabilities use an adaptive Bayes quadratic discriminant analysis classifier. We implement and validate the proposed algorithm for prediction of human trust in automation using electroencephalography (EEG) and behavioral response data. An improved accuracy is obtained for the proposed classifier as compared to an adaptive classifier that does not consider the temporal dynamics of the process being considered. The proposed algorithm can be used for classification of other human behaviors measured using psychophysiological data and behavioral responses, as well as other dynamic processes characterized by data with non-stationary distributions.**

I. Introduction

Motivation and Problem Definition: In the application of most classification algorithms, it is assumed that data samples are independent, identically distributed, and are characterized by a stationary distribution. Numerous classification algorithms have been developed for data that satisfy these assumptions (see [1] for a review). However, many real-world problems are characterized by data with temporal variations and a non-stationary distribution. One example is the use of human behavioral responses and psychophysiological data for prediction of human behavior.

Human behavior and emotion estimation is becoming an important segment in the fields of modern human-machine interaction, brain-computer interface (BCI) design, and medical care [2], among others. Human behavior inference for decision making is critical for building synergistic relationships between humans and autonomous systems. Researchers have attempted to predict human behavior using dynamic models that rely on the behavioral responses or self-reported behavior of humans [3], [4]. An alternative is the use of psychophysiological signals like the electroencephalogram (EEG) that represents the electrical activity of the brain. In order to infer human behavior from psychophysiological signals, different brain activity patterns must be identified. A common approach for this identification is the use of classification algorithms [5]. However, most of the EEG-based classification algorithms in literature are based on static classifiers that do not account for the dynamic characteristics of human behavior [5]. Therefore, our goal is to use both behavioral responses and psychophysiological measurements to create a more accurate and robust classification algorithm that considers the dynamics of human behavior.

Gaps in Literature: Most existing classification algorithms do not consider the temporal dynamics of the process under consideration. For classification of dynamic processes such as human behavior, inclusion of the temporal dynamics will improve prediction accuracy. However, dynamic classification algorithms (e.g., hidden Markov models) are typically computationally expensive to train adaptively, and therefore, cannot be used for data with non-stationary characteristics [6], [7], [8].

Contribution: In this paper, we propose an adaptive probabilistic classification algorithm which incorporates the temporal dynamics of the underlying process under consideration. We use a generative model with the prior probability modeled using a Markov decision process and the conditional probability modeled using an existing adaptive quadratic discriminant analysis classifier. We implement the proposed algorithm for classification of human trust in automation using psychophysiological measurements along with human behavioral responses. Finally, we cross-validate the classifier and show the improvement in its performance as compared to the adaptive classification algorithm alone.

Outline: This paper is organized as follows. Section II provides background on classification algorithms using EEG. The proposed classification model framework is described in Section III. The implementation of the proposed model for predicting human trust is presented in Section IV. Results and discussions are presented in Section V, followed by concluding statements in Section VI.

II. Background

There are several classification algorithms which are used in BCI applications and human behavior predictions. These include a variety of algorithms, including linear classifiers (e.g. linear discriminant analysis, support vector machines), nonlinear Bayesian classifiers, artificial neural networks, and k-nearest neighbors [5]. These classifiers can be categorized using two taxonomies: Generative vs. Discriminative and Static vs. Dynamic.

*This material is based upon work supported by the National Science Foundation under Award No. 1548616. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907 USA kakash@purdue.edu, tahira@purdue.edu, neerajain@purdue.edu

Generative classifiers, e.g., Bayes quadratic discriminant analysis (QDA), learn the distribution of each class and compute the likelihood of each class for classification. Discriminative classifiers, e.g., support vector machines (SVM), only learn the explicit decision boundaries between the classes, which are then used for classification [9]. Since the EEG signals have non-stationary distributions, data collected on-line may be characterized by different underlying distributions than the training data. Therefore, for an adaptive implementation, it is preferable to identify the changes in the underlying distribution and update a generative model accordingly than to update the decision boundary in a discriminative classifier. Furthermore, generative models are typically specified as probabilistic models; this enables a richer description between features and classes than can be achieved using discriminative models by providing a distribution model of how the data are actually generated.

Static classifiers, e.g., SVM, do not account for temporal information during classification as they classify a single feature vector. In contrast, dynamic classifiers, e.g., hidden Markov models (HMM), account for temporal dynamics by classifying a sequence of feature vectors. HMMs have been used for classification of temporal sequences of EEG features as described in [6], [7], [8]. While these studies showed that they were promising classifiers for BCI systems, the Viterbi algorithm used for training HMM is both computationally expensive and memory intensive [10]. Therefore, HMM is undesirable for use as an adaptive algorithm. Instead, to design an adaptive probabilistic classifier, we will use a generative model, namely, the Bayesian quadratic discriminant analysis (QDA) classifier. To include temporal dynamics in the classification, we propose to supplement the QDA classifier with a dynamic behavioral model using Markov decision process.

III. Probabilistic Classification Algorithm

Probabilistic classifiers predict a probability distribution over the classes, instead of just predicting the most likely class. For predicting the probability of a class label $\mathcal{C}_k$ using the feature vector $\boldsymbol{x}$, we use training data to learn a model for the posterior class probability $P(\mathcal{C}_k|\boldsymbol{x})$. A subsequent decision state uses these posterior class probabilities to assign class labels. *Generative models* initially determine the class-conditional probabilities $P(\boldsymbol{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ and also presume the prior class probabilities $P(\mathcal{C}_k)$. Then, they use Bayes' theorem,

$$P(\mathcal{C}_k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{P(\boldsymbol{x})} \tag{1}$$

to estimate the posterior class probabilities $P(\mathcal{C}_k|\boldsymbol{x})$. The denominator $P(\boldsymbol{x})$ is a normalization constant.

We consider generative models in this work and incorporate dynamic characteristics using the prior class probabilities based on Markov decision process as discussed in Section III-B. In Section III-A, we provide the mathematical foundations for the QDA classifier as well as an adaptive implementation of it based on [11].

A. Adaptive Quadratic Discriminant Analysis Classifier

A Quadratic Discriminant Analysis (QDA) classifier uses a generative approach for classification. The posterior probability that a point $\boldsymbol{x}$ belongs to class $\mathcal{C}_k$ is calculated using (1) as the product of the prior probability ($P(\mathcal{C}_k)$) and the multivariate normal density ($P(\boldsymbol{x}|\mathcal{C}_k)$) [12]. The density function of the multivariate normal distribution with mean $\boldsymbol{\mu_k}$ and covariance $\boldsymbol{\Sigma_k}$ at a point $\boldsymbol{x}$ is

$$P(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma_k}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k})\Big) \ , \tag{2}$$

where $|\boldsymbol{\Sigma_k}|$ is the determinant of $\boldsymbol{\Sigma_k}$ [12]. The Quadratic Discriminant Analysis (QDA) classifies $\boldsymbol{x}$ to a class $\mathcal{C}_k$ so as to maximize a posteriori probability of the class, i.e.,

$$\hat{\mathcal{C}_k} = \underset{i=1,\ldots,K}{\operatorname{argmax}} \hat{P}(\mathcal{C}_i|\boldsymbol{x}) \ . \tag{3}$$

Therefore, to train a QDA classifier, we need to estimate the means ($\boldsymbol{\mu_k}$) and covariance matrices ($\boldsymbol{\Sigma_k}$) for each class label. This estimation is given by the Maximum Likelihood Estimate (MLE) as $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$, and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \hat{\boldsymbol{\mu}}^2$. Moreover, the prior probabilities for each class, $P(\mathcal{C}_k)$, are estimated using the sample frequency of each class in the training data. The parameters are typically estimated using a training dataset offline and then used for prediction. However, an adaptive implementation of the QDA classifier developed by Anagnostopoulos et al. [11] uses online learning with forgetting factor λ as shown in (4).

$$\hat{\boldsymbol{\mu}}_t = \Big(1-\frac{1}{t}\Big)\hat{\boldsymbol{\mu}}_{t-1} + \frac{1}{t}\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}_0 = 0 \tag{4a}$$

$$\hat{\boldsymbol{\Pi}}_t = \Big(1-\frac{1}{t}\Big)\hat{\boldsymbol{\Pi}}_{t-1} + \frac{1}{t}\boldsymbol{x}_t\boldsymbol{x}_t^T, \hat{\boldsymbol{\Pi}}_0 = 0 \tag{4b}$$

$$\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Pi}}_t - \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_t^T \tag{4c}$$

$$n_t = \lambda_{t-1}n_{t-1} + 1 \tag{4d}$$

Here, $\bullet_t$ refers to the t^{th} discrete time value of the variable $\bullet$. The prior probabilities can be calculated as

$$\big(P(\mathcal{C}_k)\big)_t = \Big(1-\frac{1}{n_t}\Big)\big(P(\mathcal{C}_k)\big)_{t-1} + \frac{1}{n_t}\mathbb{I}((\mathcal{C}_k)_t = \mathcal{C}_k), \tag{5}$$

where $\mathbb{I}(x=k)$ is the indicator function that is equal to 1 when the value of x is equal to that of k; else it is 0. A complete derivation can be found in [11].

B. Dynamic probabilistic model for prior probability

Apart from model adaptation, the adaptive QDA classifier is static in nature; that is, the classifier only considers the present data without considering the *dynamics* of the data. Though past data could be used as a part of $\boldsymbol{x}$, it would significantly increase the dimension of parameters to be estimated. Instead, we propose a dynamic probabilistic model to estimate the prior probability $P(\mathcal{C}_k)$ that would supplement the estimation of posterior probability $P(\mathcal{C}_k|\boldsymbol{x})$ using (1). The input to this model could include variables from $\boldsymbol{x}$ and/or other variables that were not used for the classifier. The modeling frameworks for this dynamic probabilistic model

can include state space models (SSM), Markov decision processes (MDP), or HMMs. Here we will consider the use of MDP for modeling the prior probability for classification.

A MDP is a 5-tuple $(\mathcal{S}, \mathcal{A}, T, \mathcal{R}, \gamma)$, with a finite set of states $\mathcal{S}$, a finite set of actions $\mathcal{A}$, state transition probability function $T(s'|s,a) = P[S_{t+1} = s'|S_t = s, A_t = a]$, reward function $\mathcal{R}$, and discount factor $\gamma \in [0,1]$. MDPs are typically used for reinforcement learning to identify the best policy that maximizes the reward. Policy identification is outside the scope of this work. Therefore, for our application of probabilistic dynamic modeling, the reward function $\mathcal{R}$ and the reward discount factor γ will not be considered.

If $T(s'|s,a)$ is not known, it can be empirically estimated, based upon data consisting of actions and corresponding state transitions, using the MLE given as

$$\hat{T}(i,j,k) = \frac{N_{ijk}}{\sum_j N_{ijk}} \qquad (6)$$
$$N_{ijk} = \sum_{t=1}^{n} \mathbb{I}(s_t = i)\mathbb{I}(s_{t+1} = j)\mathbb{I}(a_t = k) \ ,$$

where $\mathbb{I}(s_t = i)$ is the indicator function which is equal to 1 when the state s at time t is i, else it is 0. The other two indicator functions are similarly defined. Once the state transition probability function $T(s'|s,a)$ is known, the probability for the next state s' is based on the present state s and action a as $T(S_t = s, S_{t+1} = s', A_t = a)$. Further, the n step ahead transition matrix T_n can be calculated given the series of actions $a_t, a_{t+1}, ..., a_{t+i}, ..., a_{t+n-1}$, as

$$T_n = \prod_{i=0}^{n-1} T(:,:,a_{t+i}) \ , \qquad (7)$$

and thereafter, the n-step ahead probabilities of states p_n can be calculated as $p_n = p_0 T_n$, where p_0 are the initial probabilities of states. These probabilities p_n will be used as the prior probability $P(\mathcal{C}_k)$ in (1) with each state s of the MDP corresponding to the labels $\mathcal{C}_k$ in the QDA classifier.

IV. Classification of Human Trust in HMI

In this section, we describe the classification of human trust behavior using psychophysiological measurements of participants, specifically EEG, along with their behavioral responses. We used behavioral responses to model the prior probability $P(\mathcal{C}_k)$ as described in Section III-B. The features extracted from the psychophysiological measurements were then used as the input $\boldsymbol{x}$ for the adaptive QDA model described in Section III-A. The framework for our adaptive classification model for human trust is shown in Fig. 1.

A. Methods and Procedures

In our previous work [13], [14], [15], we developed an experiment to elicit human trust dynamics in a simulated autonomous system. The participants interacted with a computer interface in which they were told that they would be driving a car equipped with an image-based *obstacle detection sensor*. The sensor would detect obstacles on the road in front of the car, and the participant would need to evaluate the algorithm reports and choose to either trust or distrust the report based on their experience with the algorithm. The study used a within-subjects design with respect to trust wherein both behavioral and psychophysiological data were collected. We used the data to estimate and validate the classification model for each participant. A detailed description of the study design and methods can be found in [14], [15].

Five hundred eighty-one participants (340 males, 235 females, and 6 unknown) recruited using Amazon Mechanical Turk [16], participated in our study online. The compensation was \$0.50 for their participation, and each participant electronically provided their consent. The Institutional Review Board at Purdue University approved the study. These data only consisted of the behavioral responses and were used to estimate the MDP model parameters.

Forty-eight adults between 18 and 46 years of age (mean: 25.0 years old, standard deviation: 6.9 years) from West Lafayette, Indiana (USA) were recruited using fliers and email lists and participated in an in-lab study. All participants were compensated at a rate of \$15/hr. The group of participants were diverse with respect to their age, professional field, and cultural background (i.e., nationality). Psychophysiological data along with behavioral data were collected from these participants and used for modeling and validation of the proposed trust classification algorithm. We removed data for three participants that had anomalous EEG spectra, possibly due to bad channels or dislocation of EEG electrodes during the study, resulting in 45 participants to analyze.

B. Trust behavior modeling using MDP

At each trial, each participant was presented with a stimuli (obstacle detected or clear road) to which they had to respond 'trust' or 'distrust' based on their previous experience (reliable or faulty trial) and from the feedback they received about the sensor after they responded. For this experiment, we define human trust behavior as the process we will model using an MDP as described below:

- The trust decision of the humans is the finite set of states, i.e., $\mathcal{S}$: {Distrust, Trust}
- The decision process of human trust is influenced by the actions of the machine that lead to the machine performance (experience) as the finite set of actions, i.e., $\mathcal{A}$: {Reliable, Faulty}
- The experience from trial t acts as an action for the new process state at $t+1$. Therefore, the human state s of trust at t moves to a new state s' at $t+1$ due to the action (i.e., machine performance or experience) at t.
- The state transition probability function $T(s'|s,a)$ can be represented as a $2 \times 2 \times 2$ matrix, such that $T(i,j,k)$ represents the transition probability from i^{th} state to j^{th} state given the action k. Therefore, each of $P(:,:,k)$ represents the state transition matrix for the k^{th} action.

We estimated the transition probability function as well as the initial state probabilities using the behavioral data collected from Amazon Mechanical Turk. We used an aggregated data of 581 participants for the estimation, and there-

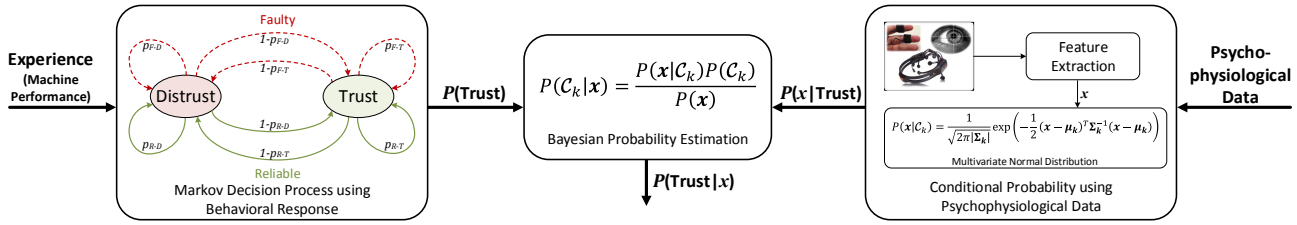


Fig. 1. A framework for adaptive probabilistic classification of human dynamic trust behavior. A Markov decision process model is used for estimating prior probability using the behavioral responses of participants. Psychophysiological measurements from the participants are used for estimating the conditional probability for each trust state.

fore assumed that a single transition probability function is representative of general human trust behavior. The estimated probability matrices are given as

$$T(s, s', a = \text{Faulty}) = \begin{bmatrix} 0.5343 & 0.4857 \\ 0.3131 & 0.6869 \end{bmatrix},$$
$$T(s, s', a = \text{Reliable}) = \begin{bmatrix} 0.3177 & 0.6823 \\ 0.1191 & 0.8809 \end{bmatrix} \quad (8)$$

where s and s' are initial and final states, respectively with each consisting of $\mathcal{S} : \{\text{Distrust}, \text{Trust}\}$. For example, the transition from state Trust to Distrust after a reliable trial has a probability of 0.8809. Estimated initial state probabilities for Distrust and Trust are

$$p_0 = \begin{bmatrix} 0.1985 & 0.8015 \end{bmatrix} . \quad (9)$$

C. Adaptive QDA model using Psychophysiological Data

Adaptive implementation of the classification algorithm inherently requires processing the data and estimating trust in real-time. Therefore, we need to continuously extract features from psychophysiological measurements, which is achieved by continuously considering short segments of signals for calculations. We divided the entire duration of the study into multiple 4-second epochs (segments) with 50% overlap between each consecutive epoch. We assume that the decisive cognitive activity occurs after the participant sees the feedback based upon their previous response. Therefore, we only considered the epochs which were in between each successive beginning of a trial and response (trust/distrust) for training the classifier. All epochs were used for prediction. We extracted an *exhaustive set* of potential features from the data for each epoch. We then reduced the dimension of this feature set to include only the statistically significant variables of trust. This reduced feature set was used for classifier modeling and validation.

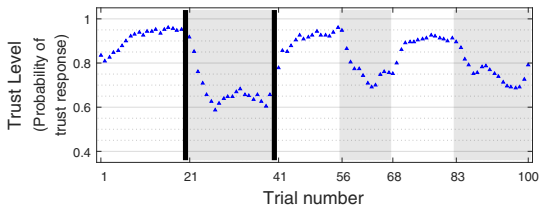
1) Feature Extraction: For each of the seven channels (Fz, C3, Cz, C4, P3, POz, and P4) of EEG data, we extracted both frequency and time domain features from each epoch as described in [15]. For frequency domain features, we decomposed each channel's data into four spectral bands, namely delta (0 Hz - 4 Hz), theta (4 Hz - 8 Hz), alpha (8 Hz - 16 Hz), and beta (16 Hz - 32 Hz) and calculated the mean, variance, and signal energy for each band of each epoch. This introduced 84 $(7 \times 4 \times 3)$ potential features. For time domain features, we included mean, variance, peak-to-peak values, mean frequency, root-mean-square, and signal energy of each epoch, thus introducing 42 (7×6) more potential features. Furthermore, to consider the interaction between different regions of the brain, we calculated the correlation between pairs of channels for each epoch, adding another 21 features.

TABLE I
FEATURES USED AS INPUT VARIABLES FOR TRUST CLASSIFICATION

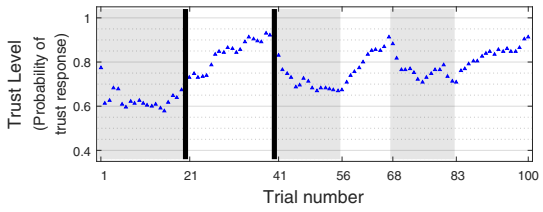
	Feature	Domain
1	Mean Frequency - P4	Time
2	Mean Frequency - C4	Time
3	Mean Frequency - P3	Time
4	Peak-to-peak - C4	Time
5	Peak-to-peak - C3	Time
6	Root Mean Square - Fz	Time
7	Energy - Fz	Time
8	Variation - Fz	Time
9	Correlation - C4 & P4	Time
10	Energy of Beta Band - P3	Frequency
11	Energy of Beta Band - Cz	Frequency
12	Energy of Beta Band - C3	Frequency
13	Variation of Beta Band - P3	Frequency
14	Variation of Beta Band - Cz	Frequency
15	Variation of Beta Band - C3	Frequency

2) Feature Selection: The EEG data resulted in 147 $(84+42+21)$ potential features. To avoid "the curse of dimensionality" [5], these features were reduced to a smaller feature set using a filter approach feature selection algorithm [12]. Participants were randomly divided into two sets, namely, a training-set consisting of 23 participants and a validation-set consisting of 22 participants. Using only training-set participants' data, we selected the best 15 features using the *Scalar Feature Selection* technique [12], [13]. Fisher Discriminant Ratio (FDR) was used as the class separability criterion with a penalty proportional to the cross-correlation between features. This penalty ensures that the selected features are least correlated, therefore reducing redundancy between features. The selected features are shown in Table I.

3) Modeling and validation: The selected feature set was extracted from EEG data to construct the input $\boldsymbol{x}$ to evaluate $P(\boldsymbol{x}|\mathcal{C}_k)$ using (2). It should be noted that for each class label $\mathcal{C}_k$, $\mu_k \in R^{n\times 1}$ and $\Sigma_k \in R^{n\times n}$, where n is the cardinality of the feature set. Therefore, for each class label, $n(n+3)/2$ parameters need to be estimated. This is a relatively large number of parameters given our number of data points. For example, for a two class problem with 15 features, the number of parameters to be estimated is 270 using approximately 270 data points in our study. This leads to significant variations in the estimated covariance matrices and often leads to ill-conditioned matrices which



(a) Group 1



(b) Group 2

Fig. 2. Participants' trust level (blue dots). Faulty trials are highlighted in gray, and black lines mark the breaks between databases.

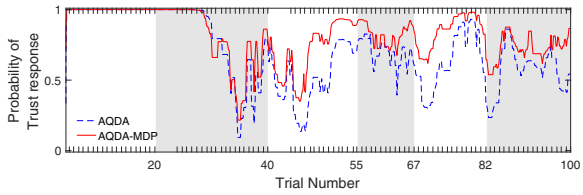
cannot be inverted. This is particularly a challenge during the initial estimation period when even fewer data are available. Therefore, to avoid inversion of ill-conditioned matrices and reduce the number of parameters to be estimated, we assume that the features are independent of each other. This results in covariance matrices that are diagonal and easily invertible. Furthermore, this reduces the number of parameters to be estimated to $2n$ for each class label (i.e. 60 parameters in our example above).

We included psychophysiological measurements in order to identify any latent indicators of trust and distrust. We hypothesized that the trust level would be high in reliable trials and be low in faulty trials, which was validated using responses collected from 581 online participants via Amazon Mechanical Turk [16] as shown in Fig. 2 [14]. Therefore, data from reliable trials were labeled as trust, and data from faulty trials were labeled as distrust. In the next section, we use these features extracted from psychophysiological data, along with the dynamic behavioral model derived in Section III-B, to implement the proposed classification algorithm.

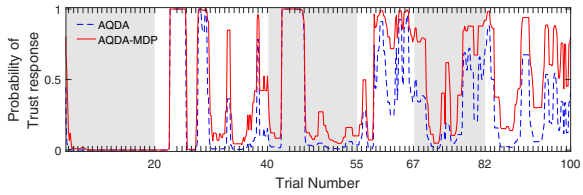
V. Results and Discussions

We implemented the Adaptive Quadratic Discriminant Analysis classifier with Markov Decision Process-based prior probability (hereafter called AQDA-MDP) using the selected features $\boldsymbol{x}$ shown in Table I, class labels $\mathcal{C}_k \in \{\text{Distrust}, \text{Trust}\}$, state transition matrix as given in (8), and the initial state probability as given in (9). For comparison, we also consider the Adaptive Quadratic Discriminant Analysis classifier (hereafter, called AQDA) exclusively with the prior probability estimated using (5). The forgetting factor λ was taken as 1, i.e., no forgetting was used. The algorithms were used for online training and validation of trust classification models from the real-time data for each participant individually.

The results for two different training-set participants and for two different validation-set participants are shown in Fig. 3 and Fig. 4, respectively. Faulty trials are highlighted

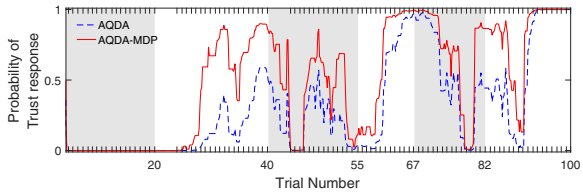


(a) Prediction of trust for participant 5 in the training set.

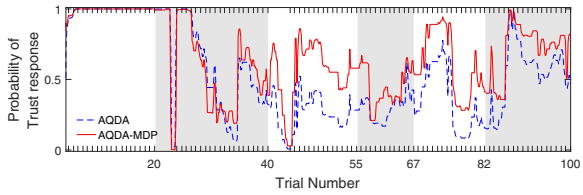


(b) Prediction of trust for participant 7 in the training set.

Fig. 3. Training-set participants' trust level predictions using AQDA-MDP and AQDA algorithms. Faulty trials are highlighted in gray.



(a) Prediction of trust for participant 36 in the validation set.

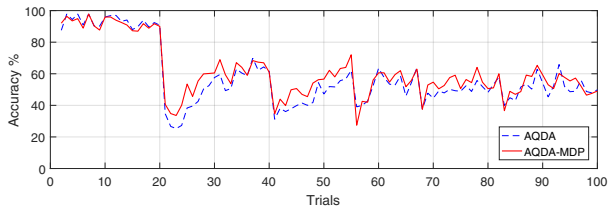


(b) Prediction of trust for participant 34 in the validation set.

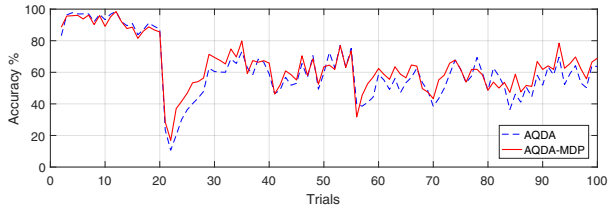
Fig. 4. Validation-set participants' trust level predictions using AQDA-MDP and AQDA algorithms. Faulty trials are highlighted in gray.

in gray, and reliable trials are highlighted in white. A high probability of trust is expected in reliable trials, and a low probability of trust is expected in faulty trials. To observe the benefits of adaptation and to compare the performance of each models, we calculate the mean trial accuracy for each trial. Mean trial accuracy is calculated as the average, across participants, of the percentage of correct prediction for epochs for each trial. The variation of mean trial accuracy for training-set and validation-set participants are shown in Fig. 5(a) and Fig. 5(b), respectively. It can be seen that the performance of the classifier is consistent between training-set and validation-set participants. Therefore, the selected set of features are capable of predicting trust behavior.

We see that the accuracy of the classifier is high for the first 20 trials (see Fig. 5). This is the consequence of the experiment design, which has data for one of the classes (either trust or distrust) initially, therefore making the classifier biased towards the initial training data. Consequently, the classifier accuracy just after the 20^{th} trial is poor, and it takes approximately 4-5 trials to eliminate the bias effect and have a considerable sample size for both classes. After



(a) Training-set participants



(b) Validation-set participants

Fig. 5. Mean Trial accuracy for ADQA and AQDA-MDP algorithms.

the 55^{th} trial, the classifier prediction accuracy decreases as shown in Fig. 5. One of the potential reasons is improper class labeling of the data. We assumed that the participants trusted the obstacle detection sensor during the reliable trials and distrusted it during the faulty trials. However, in the later trials during which the sensor reliability changes more rapidly, participants may have been unsure about the system performance. Therefore, our assumption for class labeling may not hold for data collected during these trials. As a result, the adaptive algorithm incorrectly trains itself in the later trials, resulting in accuracy approximately between 40% and 65% as shown in Fig. 5. A better way to label the trials as trust or distrust could improve the performance of the classifier and is the subject of future work. The mean trial accuracy for AQDA-MDP is, in general, higher than that of AQDA. Despite the limitations of class labeling for our experiment, the proposed algorithm enables the combination of two different types of modeling frameworks, a static QDA classifier and a dynamic MDP, systematically using a Bayesian approach to yield a classifier with improved accuracy. More generally, this algorithm can be used for classification of other human behaviors measured using psychophysiological data and behavioral responses, as well as other dynamic processes characterized by data with non-stationary distributions.

VI. Conclusion

To achieve symbiotic human-machine interactions, human behavior modeling is of utmost importance. This can be accomplished with classification algorithms using psychophysiological measurements and behavioral responses of humans. Traditional classification algorithms, however, do not consider the temporal dynamics of human behavior and the non-stationary characteristics of psychophysiological signals. In this paper, we described an adaptive probabilistic classification algorithm for human behavior which uses a dynamic MDP model to incorporate these temporal dynamics. First, we estimated the parameters for a MDP using behavioral responses. We then extracted an exhaustive set of features from psychophysiological data from 23 training-set participants and reduced the dimension of the feature space using scalar feature selection. We trained a real-time adaptive QDA-based classifier using data collected online for these 23 participants. The classifiers were validated against human subject data from another 22 validation-set participants, and an improved accuracy was obtained with classifier augmented with a dynamic MDP. Future work will include comparing the performance of the proposed classification algorithm against other dynamic classifiers.

Acknowledgment

The authors are extremely grateful and sincerely acknowledge the guidance and help of Dr. Wan-Lin Hu in design of experiments and collection of psychophysiological data.

References

[1] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[2] D. Tan and A. Nijholt, *Brain-Computer Interfaces and Human-Computer Interaction*. London: Springer London, 2010, pp. 3–19.

[3] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer Berlin Heidelberg, 1999, pp. 221–231.

[4] M. Hoogendoorn, S. W. Jaffry, P.-P. Van Maanen, and J. Treur, "Modeling and validation of biased human trust," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02*. IEEE Computer Society, 2011, pp. 256–263.

[5] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, pp. R1–R13, June 2007.

[6] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden Markov models for online classification of single trial EEG data," *Pattern recognition letters*, vol. 22, no. 12, pp. 1299–1309, 2001.

[7] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 9, no. 3, pp. 283–288, 2001.

[8] F. Cincotti, A. Scipione, A. Timperi, D. Mattia, A. Marciani, J. Millan, S. Salinari, L. Bianchi, and F. Bablioni, "Comparison of different feature classifiers for brain computer interfaces," in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*. IEEE, 2003, pp. 645–647.

[9] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 841–848.

[10] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[11] C. Anagnostopoulos, D. K. Tasoulis, N. M. Adams, N. G. Pavlidis, and D. J. Hand, "Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification," *Statistical Analysis and Data Mining*, vol. 5, no. 2, pp. 139–166, Apr. 2012.

[12] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, ser. Pattern Recognition Series. Elsevier Science, 2006.

[13] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-Time Sensing of Trust in Human-Machine Interactions," in *1st IFAC Conference on Cyber-Physical & Human-Systems*, Florianopolis, Brazil, 2016.

[14] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic Modeling of Trust in Human–Machine Interactions," in *2017 American Control Conference*, Seattle, WA, 2017.

[15] K. Akash, W. L. Hu, N. Jain, and T. Reid, "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR," *ACM Transactions on Interactive Intelligent Systems*, 2018. (In Press).

[16] Amazon, "Amazon mechanical turk," 2005. [Online]. Available: https://www.mturk.com/ [Accessed 22 February 2017]