A Theoretical Framework for Independent Classifier Combination

Simon Lucey, Sridha Sridharan and Vinod Chandran
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
slucey@ieee.org, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

Abstract

The combination of classifiers from independent observation domains has a myriad of benefits in practical pattern recognition problems. In this paper we propose a firm theoretical framework from which an upper bound on classifier combination performance can be calculated, based on mismatches between train and test sets. Using this framework, insights can be gained into the conditions under which classifiers can best be combined and where their respective confidence errors stem from. The theoretical framework is presented along with synthetic experiments for empirical validation.

1. Introduction

The combination of an ensemble of classifiers, has been shown empirically to be of great benefit in many practical pattern recognition applications, such as audio-visual speech recognition [3], and multi-modal person identification/verification [2]. In most of these tasks, the errors associated with each observation domain (i.e. modality) can be assumed to be conditionally *independent*. Through the appropriate choice of combination rules, it is possible to dampen the overall effect of the *independent* errors in each observation domain, thus giving superior performance to any of the classifiers individually.

The optimal combination rule for an ensemble of R *independent* classifiers, with respect to classification error, is the product rule [1, 2],

$$F_{pr}(Pr(\omega_i|\mathbf{o}^{\{r\}}), \forall r) = P(\omega_i)^{-(R-1)} \prod_{r=1}^{R} Pr(\omega_i|\mathbf{o}^{\{r\}}) \quad (1)$$

where $P(\omega_i)$ is the *a priori* probability for class ω_i and $Pr(\omega_i|\mathbf{o}^{\{r\}})$ is the *a posteriori* probability of observation $\mathbf{o}^{\{r\}}$ from observation domain r lying in class ω_i. It must be emphasised that $F_{pr}()$ gives a confidence score, *not* a probability, but is equivalent to the probability $Pr(\omega_i|\mathbf{o}^{\{1\}}, \ldots, \mathbf{o}^{\{R\}})$ in terms of the class decision boundaries it realises. Equation 1 holds true if one has access to the true *a posteriori* probabilities from all R independent observation domains. In practice however, one can rarely apply this rule due to the differing decision boundaries realised from the *mismatch* between train and test sets. This mismatch results in a confidence error,

$$Pr(w_i|\mathbf{o}^{\{r\}}) = \hat{Pr}(w_i|\mathbf{o}^{\{r\}}) + \epsilon_i(\mathbf{o}^{\{r\}}) \quad (2)$$

In this paper, we propose a theoretical framework to calculate and remove the mismatch confidence error, thus allowing one to gain the true *a posteriori* probability for class ω_i, and subsequently use the optimal product rule described in Equation 1.

This paper is divided into a number of sections. In Section 2, the train/test mismatch framework is proposed and the concepts of generalised knowledge and context introduced. A theoretical basis for confidence error is presented in Section 3, with the steps required to calculate and remove this mismatch error explained. In Section 4, a synthetic example is given so as to provide *some* empirical validation to the framework proposed in Section 3. Finally, some discussions and conclusions are made about the usefulness of this framework in practical classification problems, and how the framework can be used and possibly expanded in practice.

2. Modelling train/test mismatches

The idea of a train/test mismatch can be formally described if we analyse the problem of determining an *a posteriori* probability in terms of sets. The observation $\mathbf{o}$ exists in the set $\mathcal{S}_{all}$ such that $\mathbf{o} \in \mathcal{S}_{all}$. At any given time, we only have at our disposal observations existing in the subset $\mathcal{S}_{trn} \subset \mathcal{S}_{all}$ or $\mathcal{S}_{tst} \subset \mathcal{S}_{all}$, representing training and testing observations respectively. When one has to gain an *a posterior* probability estimate $\hat{Pr}(\omega_i|\mathbf{o})$ of observation $\mathbf{o} \in \mathcal{S}_{tst}\{i\}$, one has to make a decision based on knowledge gained from observations lying in $\mathcal{S}_{trn}\{i\}$ even though $\mathbf{o}$ may not. A depiction of this situation is shown in

1051-4651/02 $17.00 (c) 2002 IEEE

the Venn diagram in Figure 1(b) where $\mathcal{S}_{all}$, $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$ are subsets. Within a Bayesian framework, one has to allow for the possibility that $\mathbf{o} \notin \mathcal{S}_{trn}$ even though $\mathbf{o} \in \mathcal{S}_{tst}$.

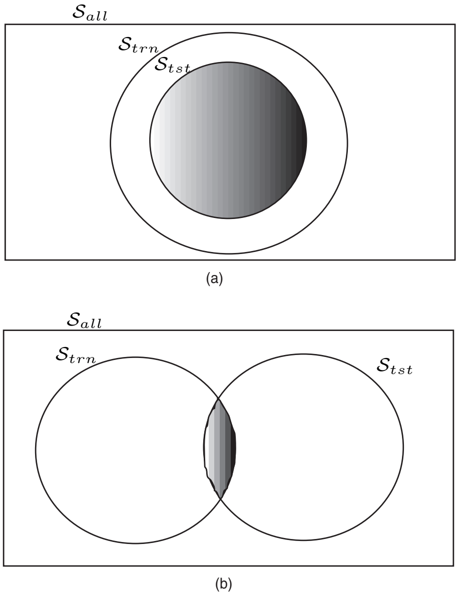

Figure 1. Venn diagram of changes in train/test conditions, (a) $\mathcal{S}_{tst} \subseteq \mathcal{S}_{trn}$ (similar train/test conditions), (b) $\mathcal{S}_{tst} \nsubseteq \mathcal{S}_{trn}$ (different train/test conditions).

Using Bayes [1] rule, when different train/test conditions are encountered, we would ideally use likelihoods $p(\mathbf{o}|\mathcal{S}_{tst}\{i\})$ derived from our knowledge of the test set to gain an *a posteriori* probability,

$$Pr(\mathcal{S}_{tst}\{i\}|\mathbf{o}) = \frac{P(\omega_i)p(\mathbf{o}|\mathcal{S}_{tst}\{i\})}{\sum_{n=1}^{N} P(\omega_n)p(\mathbf{o}|\mathcal{S}_{tst}\{n\})} \qquad (3)$$

However, our knowledge is always restricted to the narrow context of $\mathbf{o} \in \mathcal{S}_{trn}$ which should be reflected in our model thus giving,

$$p(\mathbf{o}|\omega_i) = \underbrace{P(\Omega)p(\mathbf{o}|\mathcal{S}_{trn}\{i\})}_{\mathcal{S}_{tst}\subseteq\mathcal{S}_{trn}} + \underbrace{P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega})}_{\mathcal{S}_{tst}\nsubseteq\mathcal{S}_{trn}} \qquad (4)$$

Equation 4 can be understood by using the concept of context dependent knowledge. There are two terms in Equation 4. The first term $P(\Omega)p(\mathbf{o}|\omega_i)$ represents our knowledge of discerning between classes when we are within our known knowledge context (i.e. $\mathbf{o} \in \mathcal{S}_{trn}$), where $P(\Omega)$ is the prior of the observation coming from that known context. The second term $P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega})$ represents our knowledge for discerning between classes outside the known context. This term is the *same* for all classes, as we have no knowledge for discerning between classes in the unseen context. $P(\overline{\Omega})$ and $p(\mathbf{o}|\overline{\Omega})$ is the prior of the observation coming from the unknown context and the mismatch likelihood respectively. Using this equivalence, we can gain estimates of the *a posteriori* probabilities using Bayes rule by taking into account the likelihoods of *all* classes simultaneously,

$$Pr(\omega_i|\mathbf{o}) \quad = \quad \frac{P(\omega_i)\big[P(\Omega)p(\mathbf{o}|\mathcal{S}_{trn}\{i\})+P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega})\big]}{\sum_{n=1}^{N} P(\omega_n)\big[P(\Omega)p(\mathbf{o}|\mathcal{S}_{trn}\{n\})+P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega})\big]} \qquad (5)$$

under similar train/test conditions one can make the assumption,

$$P(\Omega)p(\mathbf{o}|\mathcal{S}_{trn}\{n\}) \gg P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega}) \qquad 1 \leq n \leq N \qquad (6)$$

which leads to the commonly used estimate,

$$\begin{aligned} \hat{Pr}(\omega_i|\mathbf{o}) &= \frac{P(\omega_i)p(\mathbf{o}|\mathcal{S}_{trn}\{i\})}{\sum_{n=1}^{N} P(\omega_n)p(\mathbf{o}|\mathcal{S}_{trn}\{n\})} \\ &\doteq Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o}) \end{aligned} \qquad (7)$$

Unfortunately, in practice, it is infeasible to gain a model of $P(\overline{\Omega})$ and $p(\mathbf{o}|\overline{\Omega})$, as one requires intimate knowledge of $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$ a priori. However, one can see that if we apply Equation 7 when Equation 6 does not hold (ie. in the case of external noise or an under trained classifier) our *a posteriori* probabilities will be ill-scaled, due to the mismatch class being ignored. This results in a confidence error $\epsilon_i(\mathbf{o})$ as first mentioned in Equation 2.

Nothing can be done about this error $\epsilon_i(\mathbf{o})$ when dealing with a single observation domain with respect to classification error for the common case of equal priors. However, by suitably scaling $\hat{Pr}(\omega_i|\mathbf{o})$, one can convert the mismatch error into Bayesian error, which is intrinsically part of $Pr(\omega_i|\mathbf{o})$, allowing for the optimal use of the product rule. When adjusting for train/test mismatches in the case of unequal priors, the classification error and decision boundary will change for a single observation domain. This is due to the unequal priors becoming dominant as $P(\overline{\Omega})p(\mathbf{o}|\overline{\Omega})$ becomes larger with respect to $P(\Omega)p(\mathbf{o}|\mathcal{S}_{trn}\{i\})$.

3. Form of mismatch likelihood and priors

It is very difficult to parametrically gain a model for $p(\mathbf{o}|\overline{\Omega})$ and its prior $P(\overline{\Omega})$ as they are intrinsically dependent on the decision boundaries, formed as a consequence of the interaction of $Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o})$ and $Pr(\mathcal{S}_{tst}\{i\}|\mathbf{o})$ for all i. However, in the event of having *a priori* knowledge of $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$, one can define the *a posteriori* probability of a mismatch as,

$$Pr(\overline{\Omega}|\mathbf{o}) = 1 - \sum_{n=1}^{N} Pr(\mathcal{S}_{bth}\{n\}|\mathbf{o}) \qquad (8)$$

where,

$$\begin{aligned} Pr(\mathcal{S}_{bth}\{i\}|\mathbf{o}) &= Pr(\mathcal{S}_{trn}\{i\} \cap \mathcal{S}_{tst}\{i\}|\mathbf{o}) \\ &= Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o})Pr(\mathcal{S}_{tst}\{i\}|\mathcal{S}_{trn}\{i\}, \mathbf{o}) \end{aligned} \qquad (9)$$

Using our knowledge of conditional probability [4] for two sets $\mathcal{A}$ and $\mathcal{B}$,

$$\begin{array}{rl} \text{If } \mathcal{A} \subset \mathcal{B} \text{ then} & Pr(\mathcal{B}|\mathcal{A}) = 1 \\ \text{If } \mathcal{B} \subset \mathcal{A} \text{ then} & Pr(\mathcal{B}|\mathcal{A}) = Pr(\mathcal{B})/Pr(\mathcal{A}) \\ \text{If } \mathcal{A} \text{ and } \mathcal{B} \text{ are disjoint then} & Pr(\mathcal{B}|\mathcal{A}) = 0 \end{array} \tag{10}$$

we can define,

$$Pr(\mathcal{S}_{tst}\{i\}|\mathcal{S}_{trn}\{i\}, \mathbf{o}) = \left\{ \begin{array}{ll} 1, & \text{If } \eta_\Omega(\mathbf{o}) > 1 \\ \eta_\Omega(\mathbf{o}), & \text{otherwise} \end{array} \right. \tag{11}$$

where,

$$\eta_\Omega(\mathbf{o}) = \frac{Pr(\mathcal{S}_{tst}\{i\}|\mathbf{o})}{Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o})} \tag{12}$$

Applying this to Equation 5, using *a posteriori* probabilities instead of likelihoods, we can define,

$$Pr(\omega_i|\mathbf{o}) = \left[1 - Pr(\overline{\Omega}|\mathbf{o})\right] Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o}) + P(\omega_i)Pr(\overline{\Omega}|\mathbf{o}) \tag{13}$$

Using Equation 5 one can then calculate the mismatch likelihood $p(\mathbf{o}|\overline{\Omega})$, where the prior $P(\overline{\Omega}) = 1 - P(\Omega)$ can be interpreted as the proportion of $\mathbf{o} \notin \mathcal{S}_{bth}\{i\}$, for all i.

4. A synthetic example

In this section we propose a simple synthetic example to show the benefit of the train/test mismatch framework to classifier combination theory. In this example, we have R independent observation domains of dimensionality $D = 2$. A dimensionality of two was chosen so as to allow for graphical visualisation and cater for non-linear decision boundaries. For simplicity each domain r has two classes ω_1 and ω_2 described by Gaussian likelihood functions with equal priors. Again for simplicity and the ability for seeing the effects of varying R, the likelihood functions have the same parametric form for all domains. Each classifier density function can be parametrically described by,

$$p(\mathbf{o}^{\{r\}}|\omega_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)|_{\mathbf{o}^{\{r\}}} \tag{14}$$

where r is the observation domain, ω_i is the class, $\boldsymbol{\mu}_i$ is the class mean, and $\boldsymbol{\Sigma}_i$ is the class covariance matrix. In each observation domain, the observation train sets $\mathcal{S}_{trn}\{1\}$ and $\mathcal{S}_{trn}\{2\}$ are described by the parameters,

$$\boldsymbol{\mu}_1 = [-2.5, 2.5]^T, \boldsymbol{\mu}_2 = [2.5, -2.5]^T$$

$$\boldsymbol{\Sigma}_1 = \left[\begin{array}{cc} 2 & 0 \\ 0 & 1 \end{array} \right], \boldsymbol{\Sigma}_2 = \left[\begin{array}{cc} 1.5 & -0.5 \\ -0.5 & 1.5 \end{array} \right]$$

A train/test mismatch was introduced in this example, through the shifting of model means, so that the observation test sets $\mathcal{S}_{tst}\{1\}$ and $\mathcal{S}_{tst}\{2\}$ are described by the parameters,

$$\boldsymbol{\mu}'_1 = [-1, -1]^T, \boldsymbol{\mu}'_2 = [1, 1]^T$$

and the same covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, as the train context. A graphical depiction of these two classes can be seen in Figure 2 along with the subsequent decision boundary for the train and test contexts.

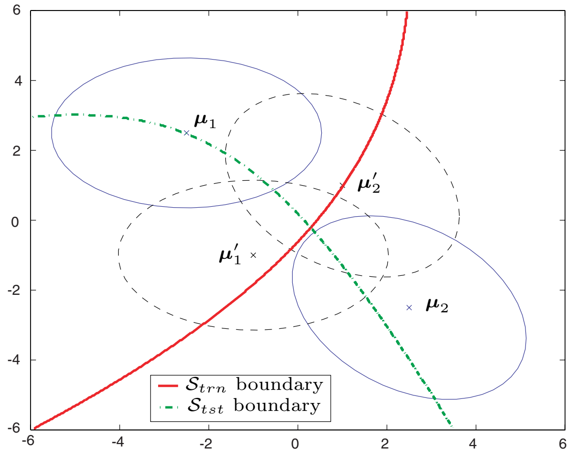

Figure 2. Depiction of synthetic example class models. 90% ellipsoid boundaries shown for both classes and contexts.

Using these train and test models we can synthetically generate independent observations to empirically verify the framework proposed in Section 3. Our experiments took the following form,

Sample size: $N_1 = N_2 = 10,000$ for both $\mathcal{S}_{trn}$ and $\mathcal{S}_{tst}$.
Number of trials: $\tau = 10$

A large number of train and test observations were used to minimise the variation of the classification result due to the finite number observations. The classification task selects the most likely class ω_{i^*}, from a group of N classes for an observation $\mathbf{o}$ such that,

$$i^* = \arg \max_{i=1}^{N} \zeta(\omega_n|\mathbf{o}) \tag{15}$$

where $\zeta(\omega_i|\mathbf{o})$ is the confidence score describing how likely observation $\mathbf{o}$ belongs to class ω_i. The error rate was determined as the percentage of incorrect classifications i^* per trial. For each of the trials the following error rates were acquired,

ϵ_{tst}**:** Error rate for a *single* domain using $\zeta_{tst}(\omega_i|\mathbf{o}) = Pr(\mathcal{S}_{tst}\{i\}|\mathbf{o}^{\{r\}})$ averaged across R domains.

ϵ_{trn}**:** Error rate for a *single* domain using $\zeta_{trn}(\omega_i|\mathbf{o}) = Pr(\mathcal{S}_{trn}\{i\}|\mathbf{o}^{\{r\}})$ averaged across R domains.

ϵ_{pr}**:** Error rate for product rule using $\zeta_{pr}(\omega_i|\mathbf{o}) = F_{pr}(\hat{Pr}(\omega_i|\mathbf{o}^{\{r\}}), \forall r)$, not accounting for mismatches, across *all* R domains.

ϵ^*_{pr}: Error rate for product rule using $\zeta^*_{pr}(\omega_i|\mathbf{o}) = F_{pr}(Pr(\omega_i|\mathbf{o}^{\{r\}}), \forall r)$, accounting for mismatches, across *all* R domains.

these error rates were averaged across the τ trials.

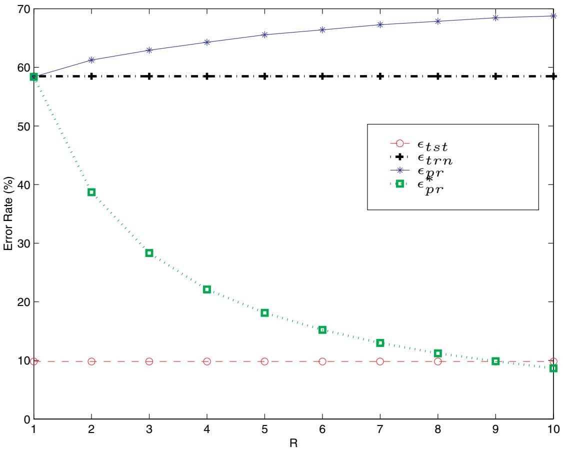

Figure 3. Empirical results for synthetic example.

The results for these different error rates can be seen in Figure 3. One can see the parametric forms of $p(\mathbf{o}|\mathcal{S}_{trn}\{i\})$ and $p(\mathbf{o}|\mathcal{S}_{tst}\{i\})$ were chosen specifically to cause *catastrophic fusion* using the product rule. Catastrophic fusion occurs when the performance of an ensemble of combined classifiers is worse than *any* of the classifiers individually. In Figure 3 it can be seen that as R increases, the error rate of ϵ_{pr} increases from the average train set error ϵ_{trn} for a single domain. However, taking the mismatch into account, the error rate of ϵ^*_{pr} decreases as R increases. It is interesting to note that the performance of ϵ^*_{pr}, given enough independent observation domains, surpasses that of ϵ_{tst} for matched conditions.

It must be emphasised that the classification error for $Pr(\omega_i|\mathbf{o})$ and $\hat{Pr}(\omega_i|\mathbf{o})$ are exactly the *same*, as predicted in Section 3 for equal priors, and shown empirically for ϵ_{pr} and ϵ^*_{pr} at $R = 1$. The mismatch error is not *removed* with respect to classification error, but now manifests itself as Bayesian error, which the product rule can optimally handle. This explains why, given a sufficiently large R, ϵ^*_{pr} can actually fall below ϵ_{tst} as classifier theory [1, 5] dictates, due to the errors being *independent*, the average error shall approach zero as R approaches infinity.

5. Discussion

The theoretical framework presented in this paper for classifier combination requires intimate knowledge of both the train and test observation sets. At first glance this type of approach may seem paradoxical. If one has intimate knowledge of the test observation set, why not adapt the decision boundaries to optimally classify observations for this new context? We concede this is a valid point, in this scenario one should adapt their classifier to match the decision boundaries realised by the test observation set.

However, in practice one rarely has intimate *a priori* knowledge of the test set, making the effects of train/test mismatches an unfortunate part of life. In this paper we have endeavored to give an understanding into why suboptimal, and in many cases catastrophic, fusion between classifiers of independent observation domains occurs. The framework also provides an *theoretical* avenue for removing these mismatch confidence errors. It has to be once again emphasised that the mismatch confidence error with respect to classification error is not removed, but transformed into Bayesian error to be handled optimally by the product rule. This distinction is important, as the mismatch class $\overline{\Omega}$ used to gain the *true* confidence scores, given a train/test mismatch, does *not* provide any class specific knowledge from the test set to improve classification performance. How train/test mismatches affect classifier combination performance, is of important practical value in many pattern recognition problems, as it is often possible to gain a practical estimate of $Pr(\overline{\Omega}|\mathbf{o})$ without prior class specific knowledge of the test set.

A firm theoretical framework for independent classifier combination has been presented. The framework has shown how train/test mismatches can affect classifier combination performance. Empirical results have been also presented that validate this framework using a synthetic example. Further work using our framework on practical problems such as audio visual speaker identification/verification can be found in [3]. The main benefit of our train/test framework is the ability to gain a theoretical understanding of where mismatch errors stem from and how they affect classifier combination performance. Additionally, one can use this understanding in practical situations where there is limited knowledge of the test set to evaluate the effectiveness of different combination rules.

References

[1] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Inc., 24-28 Oval Road, London NW1 7DX, 2nd edition, 1990.

[2] J. Kittler, M. Hatef, R. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[3] S. Lucey, T. Chen, S. Sridharan, and V. Chandran. Integration Strategies for Audio Visual Speech Processing: Applied to Text Dependent Speaker Identification/Verification. *IEEE Trans. on Multimedia*, 2001. awaiting acceptance.

[4] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., Singapore, 3rd edition, 1991.

[5] K. Tumer and J. Ghosh. Classifier Combining: Analytical Results and Implications. In *AAAI'96 - Workshop in Induction of Multiple Learning Models*, Portland, August 1996.