Bayesian Combination of Multiple, Imperfect Classifiers

Edwin Simpson, Stephen Roberts, Ioannis Psorakis
Department of Engineering Science, University of Oxford, UK.
Arfon Smith , **Chris Lintott**
Department of Physics, University of Oxford, UK.

Abstract

Classifier combination methods need to make best use of the outputs of multiple, imperfect classifiers to enable higher accuracy classifications. In many situations, such as when human decisions need to be combined, the base decisions can vary enormously in reliability. A Bayesian approach to such uncertain combination allows us to infer the differences in performance between individuals and to incorporate any available prior knowledge about their abilities when training data is sparse. In this paper we explore Bayesian classifier combination, using the computationally efficient framework of variational Bayesian inference. We apply the approach to real data from a large citizen science project, Galaxy Zoo Supernovae, and show that our method far outperforms other established approaches to imperfect decision combination. We go on to analyse the putative community structure of the decision makers, based on their inferred decision making strategies, and show that natural groupings are formed.

1 Introduction

In many real-world scenarios we are faced with the need to aggregate information from cohorts of imperfect decision making agents (*base classifiers*), be they computational or human. Particularly in the case of human agents, we rarely have available to us an indication of how decisions were arrived at or a realistic measure of agent confidence in the various decisions. Fusing multiple sources of information in the presence of uncertainty is optimally achieved using Bayesian inference, which elegantly provides a principled mathematical framework for such knowledge aggregation. In this paper we provide a Bayesian framework for such imperfect decision combination, where the base classifications we receive are greedy preferences (i.e. labels with no indication of confidence or uncertainty). The classifier combination method we develop aggregates the decisions of multiple agents, improving overall performance. We present a principled framework in which the use of weak decision makers can be mitagated and in which multiple agents, with very different observations, knowledge or training sets, can be combined to provide complementary information. The preliminary application we focus on in this paper is a distributed *citizen science* project, in which human agents carry out classification tasks, in this case identifying transient objects from images as corresponding to potential supernovae or not. This application, *Galaxy Zoo Supernovae* [1], is part of the highly successful *Zooniverse* family of citizen science projects. In this application the ability of our base classifiers can be very varied and there is no guarantee over any individual's performance, as each user can have radically different levels of domain experience and have different background knowledge. As individual users are not overloaded with decision requests by the system, we often have little performance data for individual users (base classifiers). The methodology we advocate provides a scaleable, computationally efficient, Bayesian approach to learning base classifier performance thus enabling optimal decision combinations. The approach is robust in the presence of uncertainties at all levels and naturally handles missing observations, i.e. in cases where agents do not provide any base classifications.

1.1 Independent Bayesian Classifier Combination

Here we present a variant of Independent Bayesian Classifier Combination (IBCC), originally defined in [2]. The model assumes conditional independence between base classifiers, but performed as well as more computationally intense dependency modelling methods [2]. For the ith data point, we assume that the true label t_i is generated from a multinomial distribution with probability $\boldsymbol{\kappa}$ **:** $p(t_i = j|\boldsymbol{\kappa}) = \kappa_j$. We assume that observed classifier outputs, $\boldsymbol{c}$, are discrete and are generated from a multinomial distribution dependent on the class of the true label, with parameters $\boldsymbol{\pi}$: $p(c_i^{(k)}|t_i = j, \boldsymbol{\pi}) = \pi^{(k)}_{jc_i^{(k)}}$. Thus there are minimal requirements on the type of base classifier output, which need not be probabilistic and could be selected from an arbitrary number of discrete values, indicating, for example, greedy preference over a set of class labels. The parameters $\boldsymbol{\pi}$ and $\boldsymbol{\kappa}$ have Dirichlet prior distributions with hyper-parameters $\boldsymbol{\alpha}_{\boldsymbol{j}}^{(\boldsymbol{k})} = [\alpha_{0j,1}^{(k)}, ..., \alpha_{0j,L}^{(k)}]$ and $\boldsymbol{\nu} = [\nu_{01}, ...\nu_{0J}]$ respectively, where L is the number of possible outputs from classifier k and J is the number of classes. The joint distribution over all variables is

$$p(\boldsymbol{\kappa}, \boldsymbol{\pi}, \boldsymbol{t}, \boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\nu}) = \prod_{i=1}^{N}\{\kappa_{t_i} \prod_{k=1}^{K} \pi_{t_i, c_i^{(k)}}\} p(\boldsymbol{\kappa}|\boldsymbol{\upsilon}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}). \tag{1}$$

The graphical model for IBCC is shown in figure 1.

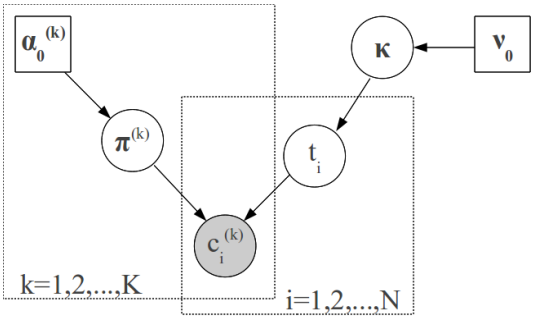


Figure 1: Graphical Model for IBCC. Shaded nodes are observed values, circular nodes are variables with a distribution and square nodes are variables instantiated with point values.

A key feature of IBCC is that $\boldsymbol{\pi}$ represents a *confusion matrix* that quantifies the decision-making abilities of *each* base classifier. This potentially allows us to ignore, or retrain, poorer classifiers and assign experts decision makers to data points that are highly uncertain. Such efficient selection of base classifiers is vitally important when obtaining a classification that has a cost related to the number of decision makers, for example. The IBCC model also allows us to infer values for missing observations of classifier outputs, $\boldsymbol{c}$, so that we can naturally handle cases in which only *partially observed* agents make decisions.

The IBCC model assumes independence between the rows in $\boldsymbol{\pi}$, i.e. the probability of each classifier's outputs is dependent on the *true label class*. In some cases it may be reasonable to assume that performance over one label class may be correlated with performance in another; indeed methods such as *weighted majority* [3] make this tacit assumption. However, we would argue that this is not universally the case, and IBCC makes no such strong assumptions.

The goal of the combination model is to perform inference for the unknown variables $\boldsymbol{t}$, $\boldsymbol{\pi}$, and $\boldsymbol{\kappa}$. The inference technique proposed in [2] was Gibbs Sampling. While this provides some theoretical guarantee of accuracy given the proposed model, it is often very slow to converge and convergence is difficult to ascertain. In this paper we consider the use of a principled approximate Bayesian methods, namely *variational Bayes* (VB) [4] as this allows us to replace non-analytic marginal integrals in the original model with analytic updates in the *sufficient statistics* of the variational approximation. This produces a model that iterates rapidly to a solution in a computational framework which can be seen as a Bayesian generalisation of the *Expectation-Maximization* EM algorithm.

In [2] an exponential prior distribution is placed over $\boldsymbol{\alpha_0}$. However, exponentials are not conjugate to the Dirichlet, and the conjugate prior to the Dirichlet is non-standard and its normalisation constant

is not in closed form [5], requiring the use of an expensive adaptive rejection Gibbs sampling step for $\boldsymbol{\alpha}$ and making even the variational Bayesian solution intractable. We therefore alter the model, so as to using point values for $\boldsymbol{\alpha_0}$, as are used in other VB models [6, 7, 8]. The hyper-parameter values of $\boldsymbol{\alpha_0}$ can hence be chosen to represent any prior level of uncertainty in the values of the agent-by-agent confusion matrices, $\boldsymbol{\pi}$, and can be regarded as pseudo-counts of prior observations, offering a natural method to include any prior knowledge and a methodology to extend the method to sequential, on-line environments.

1.2 Variational Bayes

Given a set of observed data $\boldsymbol{X}$ and a set of latent variables and parameters $\boldsymbol{Z}$, the goal of variational Bayes (VB) is to find a tractable approximation $q(\boldsymbol{Z})$ to the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$ by minimising the KL-divergence between the approximate distribution and the true distribution. We can write the log of the model evidence $p(\boldsymbol{X})$ as

$$\ln p(\boldsymbol{X}) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} \tag{2}$$

$$= L(q) - \mathrm{KL}(q||p). \tag{3}$$

As $q(\boldsymbol{Z})$ approaches $p(\boldsymbol{Z}|\boldsymbol{X})$, the KL-divergence disappears and the lower bound $L(q)$ is maximised. Variational Bayes selects a restricted form of $q(\boldsymbol{Z})$ that is tractable to work with, then seeks the distribution within this restricted form that minimises the KL-divergence. A common restriction is to assume $q(\boldsymbol{Z})$ factorises into single variable factors $q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$. For each factor $q_i(\boldsymbol{Z}_i)$ we then seek the optimal solution $q_i^*(\boldsymbol{Z}_i)$ that minimises the KL-divergence. Mean field theory [9] then shows that the log of each optimal factor $\ln q_i^*(\boldsymbol{Z}_i)$ is the expectation with respect to all other factors of the log of the joint distribution over all hidden and known variables:

$$\ln q_i^*(\boldsymbol{Z}_i) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const.} \tag{4}$$

We can evaluate these optimal factors iteratively by first initialising all factors, then updating each in turn using the expectations with respect to the current values of the other factors. Unlike Gibbs sampling, the each iteration is guaranteed to increase the lower bound on the log-likelihood, $L(q)$, converging to a (local) maximum in a similar fashion to standard EM algorithms. If the factors $q_i^*(\boldsymbol{Z}_i)$ are exponential family distributions, as is the case for the IBCC method we present in the next section, the lower bound is convex with respect to each factor $q_i^*(\boldsymbol{Z}_i)$ and $L(q)$ will converge to a *global* maximum of our approximate, factorised distribution. In practice, once the optimal factors $q_i^*(\boldsymbol{Z}_i)$ have converged to within a given tolerance, we can approximate the distribution of the unknown variables and calculate their expected values.

2 Variational Bayesian IBCC

To provide a variational Bayesian treatment of IBCC, VB-IBCC, we first propose the form for our variational distribution ($q(\boldsymbol{Z})$ in the previous section) that factorises between the parameters and latent variables.

$$q(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}) = q(\boldsymbol{t}) q(\boldsymbol{\kappa}, \boldsymbol{\pi}) \tag{5}$$

This is the only assumption we must make to perform VB on this model; the forms of the factors arise from our model of IBCC. We can use the joint distribution in equation 1 to find the optimal factors $q^*(\boldsymbol{t})$ and $q^*(\boldsymbol{\kappa}, \boldsymbol{\pi})$ it in the form given by equation 4. For the target labels we have

$$\ln q^*(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{\kappa}, \boldsymbol{\pi}}[\ln p(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{c})] + \text{const.} \tag{6}$$

We rewrite this into factors corresponding to independent data points, with any terms not involving t_i being absorbed into the normalisation constant.

$$\ln q^*(t_i) = \mathbb{E}_{\boldsymbol{\kappa}}[\ln \kappa_{t_i}] + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\pi}}[\ln \pi^{(k)}_{t_i, c_i^{(k)}}] + \text{const} \tag{7}$$

To simplify the optimal factors in subsequent equations, we define expectations with respect to $\boldsymbol{t}$ of two statistics: the number of occurrences of each target class is given by

$$N_j = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{t}}[t_i = j] = \sum_{i=1}^{N} q^*(t_i = j) \tag{8}$$

and the counts of each classifier decision, $c_i^{(k)} = l$, given the target label, $t_i = j$, given by

$$N_{jl}^{(k)} = \sum_{i=1}^{N} [c_i^{(k)} = l] \mathbb{E}_{\boldsymbol{t}}[t_i = j] = \sum_{i=1}^{N} [c_i^{(k)} = l] q^*(t_i = j). \tag{9}$$

where $[c_i^{(k)} = l]$ is unity if $c_i^{(k)} = l$ and zero otherwise.

For the parameters of the model we have the optimal factors given by:

$$\begin{aligned} \ln q^*(\boldsymbol{\kappa}, \boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{t}}[\ln p(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{c})] + \text{const} && (10) \\ &= \mathbb{E}_{\boldsymbol{t}}[\sum_{i=1}^{N} \{\ln p_{t_i} + \sum_{k=1}^{K} \ln \pi^{(k)}_{t_i, c_i^{(k)}}\}] + \ln p(\boldsymbol{\kappa} | \boldsymbol{\upsilon_0}) && (11) \\ &\quad + \ln p(\boldsymbol{\pi} | \boldsymbol{\alpha}) + \text{const}. && (12) \end{aligned}$$

In equation 10 terms involving $\boldsymbol{\kappa}$ and terms involving each confusion matrix in $\boldsymbol{\pi}$ are separate, so we can factorise $q^*(\boldsymbol{\kappa}, \boldsymbol{\pi})$ further into

$$q^*(\boldsymbol{\kappa}, \boldsymbol{\pi}) = q^*(\boldsymbol{\kappa}) \prod_{k=1}^{K} \prod_{j=1}^{J} q^*(\boldsymbol{\pi}_j^{(k)}). \tag{13}$$

Considering the prior for $\boldsymbol{\kappa}$ is a Dirichlet distribution, we obtain the optimal factor

$$\begin{aligned} \ln q^*(\boldsymbol{\kappa}) &= \mathbb{E}_{\boldsymbol{t}}[\sum_{i=1}^{N} \ln \kappa_{t_i}] + \ln p(\boldsymbol{\kappa} | \boldsymbol{\upsilon}) + \text{const} && (14) \\ &= \sum_{j=1}^{J} N_j \ln \kappa_j + \sum_{j=1}^{J} (\nu_{0,j} - 1) \ln \kappa_j + \text{const}. && (15) \end{aligned}$$

Taking the exponential of both sides, we obtain a posterior Dirichlet distribution of the form

$$q^*(\boldsymbol{\kappa}) \propto \text{Dir}(\boldsymbol{\kappa} | \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_J) \tag{16}$$

where $\boldsymbol{\nu}$ is updated in the standard manner by adding the data counts to the prior counts ν_0:

$$\nu_j = \nu_{0,j} + N_j. \tag{17}$$

The expectation of $\ln \boldsymbol{\kappa}$ required to update equation 7 is therefore:

$$\mathbb{E}[\ln \kappa_j] = \Psi(\nu_j) - \Psi(\sum_{j'=1}^{J} \nu_{j'}) \tag{18}$$

where $\Psi(z)$ is the standard digamma function.

For the confusion matrices $\boldsymbol{\pi}_j^{(k)}$ the priors are also Dirichlet distributions giving us the factor

$$\begin{aligned} \ln q^*(\boldsymbol{\pi}_j^{(k)}, \boldsymbol{\alpha}_j^{(k)}) &= \sum_{i=1}^{N} \mathbb{E}_{t_i}[t_i = j] \ln \pi^{(k)}_{j, c_i^{(k)}} + \ln p(\boldsymbol{\pi} | \boldsymbol{\alpha}) + \text{const} && (19) \\ &= \sum_{l=1}^{L} N_{jl}^{(k)} \ln \pi_{jl}^{(k)} + \sum_{l=1}^{L} (\alpha_{jl}^{(k)} - 1) \ln \pi_{jl}^{(k)} + \text{const}. && (20) \end{aligned}$$

Again, taking the exponential gives a posterior Dirichlet distribution of the form

$$q^*(\boldsymbol{\pi}_j^{(k)}) \propto \text{Dir}(\boldsymbol{\pi}_j^{(k)} | \alpha_{j1}^{(k)}, ..., \alpha_{jL}^{(k)}) \tag{21}$$

where $\boldsymbol{\alpha}_j^{(k)}$ is updated by adding data counts to prior counts $\alpha_{0,j}^{(k)}$:

$$\alpha_{jl}^{(k)} = \alpha_{0,jl}^{(k)} + N_{jl}^{(k)}. \tag{22}$$

The expectation required for equation 7 is given by

$$\mathbb{E}[\ln \pi_{jl}^{(k)}] = \Psi(\alpha_{jl}^{(k)}) - \Psi(\sum_{l'=1}^{L} \alpha_{jl'}^{(k)}). \tag{23}$$

To apply the VB algorithm to IBCC, we initialise all the expectations over $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$, either randomly or by choosing their prior expectations (if we have domain knowledge to inform this). We then iterate over a two-stage procedure similar to the *Expectation-Maximization* (EM) algorithm. In the variational equivalent of the *E-step* we use the current expected parameters, $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$, to update the variational distribution in equation 5. First we evaluate equation 7, then use the result to update the counts N_j and $N_{jl}^{(k)}$ according to equations 8 and 9. In the variational *M-step*, we update $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$ using equations 18 and 23.

3 Galaxy Zoo Supernovae

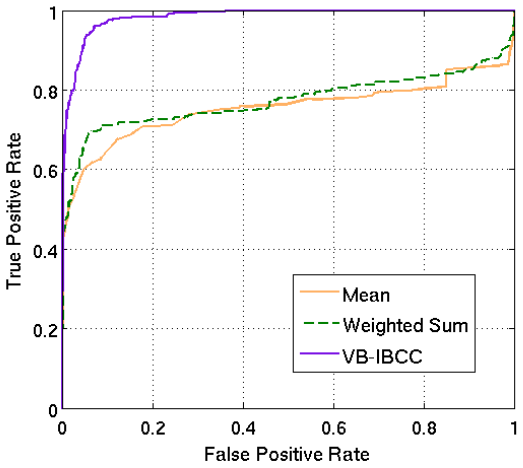
We tested the model using a dataset obtained from the Galaxy Zoo Supernovae citizen science project [1]. The dataset contains scores given by individual volunteer citizen scientists (base classifiers) to candidate supernova images after answering a series of questions, the aim being to classify each data sample (images) as either "supernova" or "not supernova". A set of three linked questions are answered by the users, which are hard-coded in the project repository to scores of -1, 1 or 3, corresponding respectively to decisions that the data point is very unlikely to be a supernova, possibly a supernova and very likely a supernova.

In order to verfiy the efficacy of our approach and competing methods, we use "true" target classifications obtained from full spectroscopic analysis, undertaken as part of the Palomar Transient Factory collaboration [10]. We note that this information, is not available to the base classifiers (the users), being obtained retrospectively. This labelling is not made use of in our algorithms, save for the purpose of measuring performance. We compare IBCC using both variational Bayes (VB-IBCC) and Gibbs sampling (Gibbs-IBCC), using as output the expected values of t_i. We also tested simple majority voting, weighted majority voting & weighted sum [3] and mean user scores, which the Galaxy Zoo Supernovae currently uses to filter results. For majority voting methods we treat both 1 and 3 as a vote for the supernova class.

The complete dataset contains many volunteers that have provided very few classifications, particularly for positive examples, as there are 322 classifications of positive data points compared to 43941 "not supernova" examples. We therefore subsampled the dataset, selecting all positive data points, then selecting only negative data points that have at least 10 classifications from volunteers who have classified at least 50 examples, which produced a data set of some 1000 examples with decisions produced from around 1700 users. We tested all imperfect decision combination methods using five-fold cross validation.

Figure 2a shows the average *Receiver-Operating Characteristic* (ROC) curves taken across all cross-validation datasets for the mean score, weighted sum and VB-IBCC. The ROC curve for VB-IBCC clearly outperforms the mean of scores by a large margin. Weighted sum achieves a slight improvement on the mean by learning to discount base classifiers each time they make a mistake. The performance of the majority voting methods and IBCC using Gibbs sampling is summarised by the area under the ROC curve (AUC) in table 2b. Majority voting methods only produce one point on the ROC curve between 0 and 1 as they convert the scores to votes (-1 becomes a negative vote, 1 and 3 become positive) and produce binary outputs. These methods have similar results to the mean score approach, with the weighted version performing slightly worse, perhaps because too much information is lost when converting scores to votes to be able to learn base classifier weights correctly.

With Gibbs-sampling IBCC we collected samples until the mean of the sample label values converged. Convergence was assumed when the total absolute difference between mean sample labels



(a) Average Reciever operating characteristic (ROC) curves.

Method	AUC
Mean of Scores	0.7543
Weighted Sum	0.7722
Simple Majority Voting	0.7809
Weighted Majority Voting	0.7378
Gibbs-IBCC	0.9127
VB-IBCC	0.9840

(b) Area under the ROC curves (AUCs).

Figure 2: Galaxy Zoo Supernovae: ROC curves and AUCs with 5-fold cross validation.

of successive iterations did not exceed 0.01 for 20 iterations. The mean time taken to run VB-IBCC to convergence was 13 seconds, while for Gibbs sampling IBCC it was 349 seconds. As well as executing significantly faster, VB produces a better AUC than Gibbs sampling with this dataset.

4 Communities of decision makers

In this section we apply a recent community detection methodology to the problem of determining most likely groupings of base classifiers, the imperfect decision makers. Identifying overlapping communities in networks is a challenging task. In recent work [11] we have presented a novel approach to community detection that utilises a Bayesian factorization model to extract *overlapping* communities from a "similarity" or "interaction" network. The scheme has the advantage of soft-partitioning solutions, assignment of node participation scores to communities, an intuitive foundation and computational efficiency. We apply this approach to a similarity matrix calculated over all the citizen scientists in our study, based upon each users' confusion matrix. Denoting π_i as the (3×2) confusion matrix inferred for user i we may define a simple similarity measure between agents i and j as

$$V_{i,j} = \exp\left(-\mathcal{HD}(\pi_i, \pi_j)\right), \tag{24}$$

where $\mathcal{HD}()$ is the *Hellinger distance* between two distributions, meaning that two agents who have very similar confusion matrices will have high similarity.

Application of Bayesian community detection to the matrix $\mathbf{V}$ robustly gave rise to *five* distinct groupings of users. In figure 3 we show the centroid confusion matrices associated with each of these groups of citizen scientists. The labels indicate the "true" class (0 or 1) and the preference for the three scores offered to each user by the Zooniverse questions (-1, 1 & 3). Group 1, for example, indicates users who are clear in their categorisation of "not supernova" (a score of -1) but who are less certain regarding the "possible supernova" and "likely supernova" categories (scores 1 & 3). Group 2 are "extremists" who use little of the middle score, but who confidently (and correctly) use scores of -1 and 3. By contrast group 3 are users who almost always use score -1 ("not supernova") whatever objects they are presented with. Group 4 almost never declare an object as "not supernova" (incorrectly) and, finally, group 5 consists of "non-commital" users who rarely assign a score of 3 to supernova objects, preferring to stick with the middle score ("possible supernova"). It is interesting to note that all five groups have similar numbers of members (several hundred) but clearly each group indicates a very different approach to decision making.

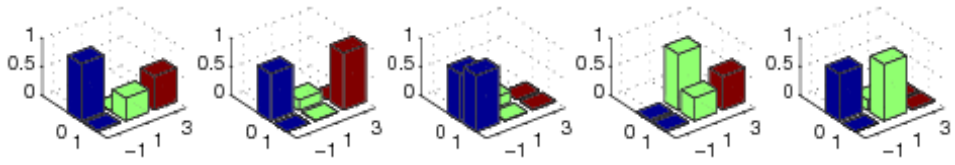


Figure 3: Prototypical confusion matrices for each of the five communities inferred using Bayesian social network analysis (see text for details).

5 Discussion

We present in this paper a very computationally efficient, variational Bayesian, approach to imperpect multiple classifier combination. We evaluated the method using real data from the Galaxy Zoo Supernovae citizen science project, with 963 data points and 1705 base classifiers. In our experiments, our method outperformed all other methods, including weighted sum and weighted majority, both of which are often advocated as they also learn weightings for the base classifiers. For our variational Bayes method the required computational overheads were far lower than those of Gibbs sampling approaches, thus giving much shorter compute time, which is particularly important for applications that need to make regular updates as new data is observed, such as our application here. Furthermore, on this data set at least, the performance was also better than the slower sample based method. We have shown that a sensible structure emerges from the cohort of decision makers via social network analysis and this provides valuable information regarding the decision-making of the groups' members.

Our current work considers how the rich information learned using this method can be exploited to improve the base classifiers, namely the human volunteer users. For example, we can use the confusion matrices, π, to identify users groups who would benefit from more training, potentially from interaction with user groups who perform more accurate decision making (via extensions of *apprenticeship learning*, for example). We also consider, via selective object presentation, ways of producing user specialisation such that the overall performance of the human-agent collective is maximised. We note that this latter concept bears the hallmark traces of *computational mechanism design* and the incorporation of incentives engineering and coordination mechanisms into the model is one of our present challenges.

References

[1] A. M. Smith, S. Lynn, M. Sullivan, C. J. Lintott, P. E. Nugent, J. Botyanszki, M. Kasliwal, R. Quimby, S. P. Bamford, L. F. Fortson15, et al. Galaxy Zoo Supernovae. 2010.

[2] Z. Ghahramani and H. C. Kim. Bayesian classifier combination. *Gatsby Computational Neuroscience Unit Technical Report No. GCNU-T., London, UK:*, 2003.

[3] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[4] Hagai Attias. *Advances in Neural Information ... - Google Books*, chapter A Variational Bayesian Framework for Graphical Models, pages 209–215. 2000.

[5] S. Lefkimmiatis, P. Maragos, and G. Papandreou. Bayesian inference on multiscale models for poisson intensity estimation: Applications to Photon-Limited image denoising. *Image Processing, IEEE Transactions on*, 18(8):1724–1741, 2009.

[6] R. Choudrey and S. Roberts. Variational Mixture of Bayesian Independent Component Analysers. *Neural Computation*, 15(1), 2003.

[7] C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer Science+Business Media, LLC, 4 edition, 2006.

[8] W. D Penny and S. J Roberts. Dynamic logistic regression. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 3, pages 1562–1567, 1999.

[9] G. Parisi and R. Shankar. Statistical field theory. *Physics Today*, 41:110, 1988.

[10] N. M. Law, S. R. Kulkarni, R. G. Dekany, E. O. Ofek, R. M. Quimby, P. E. Nugent, J. Surace, C. C. Grillmair, J. S. Bloom, M. M. Kasliwal, et al. The palomar transient factory: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 121(886):1395–1408, 2009.

[11] I Psorakis, S. Roberts, M. Ebden, and B. Shelden. Overlapping Community Detection using Bayesian Nonnegative Matrix Factorization. *Physical Review E*, 83, 2011.