Journal of Machine Learning Research 20 (2019) 1-18 Submitted 4/18; Revised 1/19; Published 3/19

Bayesian Combination of Probabilistic Classifiers using Multivariate Normal Mixtures

Gregor Pirš GREGOR.PIRS@FRI.UNI-LJ.SI
Erik Štrumbelj ERIK.STRUMBELJ@FRI.UNI-LJ.SI
University of Ljubljana
Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia

Editor: Ryan Adams

Abstract

Ensemble methods are a powerful tool, often outperforming individual prediction models. Existing Bayesian ensembles either do not model the correlations between sources, or they are only capable of combining non-probabilistic predictions. We propose a new model, which overcomes these disadvantages. Transforming the probabilistic predictions with the inverse additive logistic transformation allows us to model the correlations with multivariate normal mixtures. We derive an efficient Gibbs sampler for the proposed model and implement a regularization method to make it more robust. We compare our method to related work and the classical linear opinion pool. Empirical evaluation on several toy and real-world data sets, including a case study on air-pollution forecasting, shows that the method outperforms other methods, while being robust and easy to use.

Keywords: correlated classifiers, ensemble, probabilistic models, additive logistic transformation, Bayesian inference

1. Introduction

Classifier combination (or ensemble learning) is an important area of applied machine learning and statistics. The idea behind these methods is that combining several prediction models should achieve better performance. Intuitively, if the models make different mistakes, we can combine them and overcome their individual drawbacks. Zhou (2012) highlights the importance of diversity between individual models and argues that understanding such diversity is an open problem in ensemble learning. A theoretical result on how the generalization error of a combination (weighted average of individual methods) decreases as the correlations between individual errors decrease can be found in Ueda and Nakano (1996). That ensembles achieve better performance in practice proved to be the case, as they are extensively used in various machine learning challenges and competitions to great effect—for example an ensemble method won the KDD Cup 2017 (SIGKDD, 2017). The classifiers in an ensemble can be trained on the same data, or on different data sets. In practice, one might want to combine predictions from various sources, for example, a classifier and expert predictions.

Our work was motivated by Kim and Ghahramani (2012) who proposed a Bayesian approach to combining classifiers—the independent Bayesian classifier combination (IBCC) and its dependent extension, the dependent Bayesian classifier combination (DBCC). Their method is able to combine categorical predictions of classifiers, humans, and other sources, not necessarily trained on the same data. They used Markov networks to model the dependencies between classifiers. One drawback of their method is that it can only be used to combine non-probabilistic (categorical) classifiers. Another drawback is that the method's time complexity grows exponentially in the number of classifiers. Simpson et al. (2013) applied the IBCC to a crowdsourcing problem. They observed that the sampler for IBCC can have difficulties with convergence, and proposed a variational IBCC. Additionally Simpson (2014) presented a simplification of Kim and Ghahramani's IBCC, which led to a more efficient Gibbs sampler for parameter inference. Venanzi et al. (2014) extended the IBCC specifically for crowdsourcing problems. They proposed CommunityBCC, where each source (worker) is assumed to belong to some community with a common confusion matrix. The method is able to model the dependencies implicitly through community membership. The results show that CommunityBCC outperforms IBCC on sparse data sets. Nazábal et al. (2016) extended the IBCC model to combine probabilistic models by modeling the predictions with the Dirichlet distribution, however they do not model correlations explicitly.

Another approach that is closely related to IBCC are supra-Bayesian methods, which rely on Bayes' rule to combine the information gathered from several experts and thus improve the predictions. They assume a multivariate normal (MVN) distribution of the predictions and are therefore able to model correlations. Lindley (1985) presented a method for combining probabilistic predictions of categorical data. The author proposed modeling of log-odds instead of original data to deal with the non-normality. The method also uses a common covariance matrix, independent of true labels. Due to the use of a MVN distribution and the assumption of a common covariance matrix the described method lacks flexibility to model the latent space well.

Hoeting et al. (1999) presented Bayesian model averaging that combines several probabilistic models, weighted by their posterior probability. However this approach assumes that one of the combining models is the true data generating model and that the models represent mutually exclusive situations. These assumptions can lead to severe loss of performance, as Cerquides and De Mántaras (2005) have shown.

Lacoste et al. (2014) proposed the agnostic Bayesian learning of ensembles, where they produced ensembles of predictors based on holdout estimations of their generalization performances. Their ideas are based on the inductive learning paradigm and they use Bayesian treatment to find the posterior probability of each hypothesis being the best. They defined the risk of each hypothesis as its expected loss and examined various priors over the joint risk. The inputs to the model can be probabilistic and they account for correlations between inputs. Agnostic Bayes differs from IBCC, its extensions, and supra-Bayesian methods, as it focuses on finding the best performing models and weighting them accordingly, as opposed to relying on finding a latent structure of the predictions.

We propose a new model based on the IBCC that is able to combine probabilistic predictions by learning the latent structure of the sources. We extend the IBCC to probabilistic predictions with a change of the model's distribution for the predictions. First we transform the predictions with the inverse of the additive logistic transformation and then model them with MVN mixtures to account for correlation explicitly. The correlation matrices grow quadratically in the number of classifiers and the number of classes. We construct the model with a fully Bayesian framework and use Gibbs sampling (Casella and George, 1992) for parameter inference. Since the complexity of the model grows with the number of classifiers, we implement a regularization method, which makes the model more robust. Our method can also be viewed as an extension of the supra-Bayesian method, where we use mixtures of normals as the likelihood and condition the means and covariance matrices on the true label—increasing the method's flexibility.

We empirically evaluated our method on several toy and real-world data sets, and compared it to related methods. The results on toy data sets highlight that none of the ensembles works well for all data sets. Overall, our method compares favourably to related methods. Additionally, we provide a case study on combining predictions of air-pollutant concentration in Slovenia, where we show that our method is well suited for combining machine learning models with human expert predictions.

The paper is organized as follows. In Section 2 we formulate the problem, present our main methodological contribution, and provide a description of related work. We present the data sets, empirical results, and a case study in Section 3. We discuss the results and provide directions for future work in Section 4.

2. Methods

Let $\{t_i\}_{i=1}^n$ be our data set of n observations that can take one of m different values $t_i \in \{1, \dots, m\}$. Additionally, we have r sources of probabilistic predictions and, for each source, a probabilistic prediction for each observation $\{p_i^{(1)}, \dots, p_i^{(r)}\}_{i=1}^n$, where $p_i^{(k)} = \begin{bmatrix} p_{i1}^{(k)} & \cdots & p_{im}^{(k)} \end{bmatrix}^T$ is the $k-$th source's prediction for the $i-$th observation.

The task is to learn how to combine individual sources into more accurate probabilistic predictions, so that we are able to produce probabilistic predictions $\{\hat{p}_i\}_{i=1}^{n+n^*}$ for n observed and n^* future/unobserved data, represented with categorical random variables $\{T_i\}_{i=n+1}^{n+n^*}$, for which only the sources' predictions $\{p_i^{(1)}, \dots, p_i^{(r)}\}_{i=n+1}^{n+n^*}$ are known.

2.1. MVN Mixture Conditional Likelihood Model (MM)

Aitchison (1982) proposed the modeling of correlated data on a simplex with the logit-normal distribution. The data on a simplex are considered as data drawn from a MVN distribution, transformed by the additive logistic transformation. This transformation transforms a vector $x \in \mathbb{R}^z$ into a vector $f(x) \in S^{z+1}$, where S^{z+1} represents a $(z+1)$-dimensional simplex. To model the correlations between probabilistic predictions we first transform each source's prediction with the inverse of the additive logistic transformation. Applying it to $\{p_i^{(1)}, \dots, p_i^{(r)}\}_{i=1}^{n+n^*}$, we get $\{u_i^{(1)}, \dots, u_i^{(r)}\}_{i=1}^{n+n^*}$, where $u_i^{(k)} = \begin{bmatrix} u_{i1}^{(k)} & \cdots & u_{i(m-1)}^{(k)} \end{bmatrix}^T$. Let $u_i = \begin{bmatrix} u_i^{(1)T} & \cdots & u_i^{(r)T} \end{bmatrix}$, be the concatenated vector of $u_i^{(k)}$s, where $k = 1, ..., r$. The dimension of u_i is then $r(m-1)$, the number of sources times the number of classes minus one. Then the vectors u_i, can be modeled by a MVN distribution, as described above. However, the transformation does not guarantee a MVN distribution of u_i and in practice multi-modal distributions can often arise. Therefore we model the transformed data with MVN mixtures of dimension $r(m-1)$.

Our generative model can be described as follows. The probabilistic predictions $p_i^{(k)} = \begin{bmatrix} p_{i1}^{(k)} & \cdots & p_{im}^{(k)} \end{bmatrix}^T$ are generated by applying the additive logistic transformation to realizations of MVN mixtures, whose parameters depend on the true label. The probability for a new observation to be in true label j is then proportional to the density of the corresponding MVN mixture. The true labels $\{t_i\}_{i=1}^n$ and $\{T_i\}_{i=n+1}^{n+n^*}$ are assumed to be generated by a categorical distribution with parameter ρ, and we set a Dirichlet prior on ρ, same as Nazábal et al. (2016). The mixture memberships g_i are assumed to be generated by a categorical distribution of dimension d. For the parameters of MVN distributions we set semi-conjugate MVN and inverse-Wishart priors. We write the likelihood and the priors as

$$\begin{aligned} u_i|t_i = j, g_i = h, \mu, \Sigma &\sim \mathcal{N}_{r(m-1)}(\mu_{jh}, \Sigma_{jh}), & (1)\\ t_i|\rho &\sim \text{Cat}(\rho), & (2)\\ \rho|\gamma_0 &\sim \text{Dir}(\gamma_0), & (3)\\ g_i &\sim \text{Cat}(\tau_0), & (4)\\ \mu_{jh} &\sim \mathcal{N}_{r(m-1)}(\mu_0, \Sigma_0), & (5)\\ \Sigma_{jh} &\sim \text{inv-Wishart}(\nu_0, S_0^{-1}), & (6) \end{aligned}$$

where $\mu = \{\mu_{jh} : j = 1, ..., m \wedge h = 1, ..., d\}$, $\Sigma = \{\Sigma_{jh} : j = 1, ..., m \wedge h = 1, ..., d\}$. Figure 1 shows the proposed model in plate notation. Let vectors $u_i = \begin{bmatrix} u_i^{(1)T} & \cdots & u_i^{(r)T} \end{bmatrix}$ form the rows of matrix U. Our goal is to sample from

$$\int p(T, \rho, g, \mu, \Sigma|U, t) d\rho = p(T, g, \mu, \Sigma|U, t). \quad (7)$$

To simplify the sampling we marginalize over ρ. Therefore, we need to sample T, μ, Σ, and g. In the remainder of the section we derive full-conditional distributions for these variables and construct a Gibbs sampler. We infer the parameters μ, Σ, and g only on the data where the true label is known, however, the method would also allow inference over unlabelled data.

First we observe that full-conditional distributions of μ and Σ are conditionally independent of ρ and proportional to the product of likelihood and the respective prior. Let $U^{(jh)}$ be the matrix of observations where the true label is j and that belong to the $h-$th mixture. Using the standard formulas for semi-conjugate priors and using Eq. (1), (5), and (6), we get the following full-conditionals

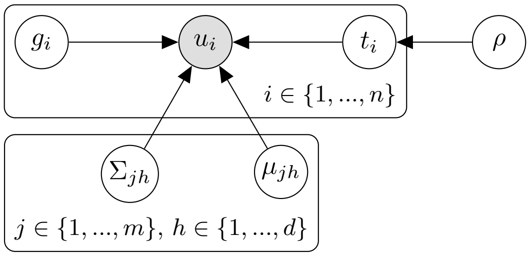

Figure 1: Bayesian plate notation for the proposed model. The transformed predictions u_i for the $i-$th observation are assumed to be generated conditional on the true label $t_i = j$ from the corresponding MVN distribution with parameters μ_{jh} and Σ_{jh}, which also depend on the latent group $g_i = h$. The true label is distributed categorically with parameter ρ. Priors are omitted for brevity.

$$\mu_{jh}|U, t, \rho, g, \Sigma \sim \mathcal{N}_{r(m-1)}(\mu^*, \Sigma^*), \tag{8}$$

$$\Sigma_{jh}|U, t, \rho, g, \mu \sim \text{inv-Wishart}(\nu^*, S^{*-1}), \tag{9}$$

$$\begin{aligned}
\mu^* &= \Sigma^*(\Sigma_0^{-1}\mu_0 + n_{jh}\Sigma_{jh}^{-1}\bar{u}^{(jh)}),\\
\Sigma^* &= (\Sigma_0^{-1} + n_{jh}\Sigma_{jh}^{-1})^{-1},\\
\nu^* &= \nu_0 + n_{jh},\\
S^* &= S_0 + \sum_{l=1}^{n_{jh}}(u_l^{(jh)} - \mu_{jh})(u_l^{(jh)} - \mu_{jh})^T,
\end{aligned}$$

where n_{jh} is the number of true labels equal to j in group h and $\bar{u}^{(jh)}$ are the column means of $U^{(jh)}$.

The mixture membership variable g is also conditionally independent of ρ. Using Eq. (1) and (4), the full-conditional of g is

$$\begin{aligned}
p(g|U, t, \rho, \mu, \Sigma) &\propto p(U|g, t, \rho, \mu, \Sigma)p(g, t, \rho, \mu, \Sigma)\\
&\propto \Big(\prod_{i=1}^{n} p(u_i|g_i, t_i, \rho, \mu, \Sigma)\Big)p(g)\\
&\propto \prod_{i=1}^{n} p(u_i|g_i, t_i, \mu, \Sigma)p(g_i).
\end{aligned} \tag{10}$$

For the predictive distribution of a new observation T_i, we do not have samples of g_i. Therefore we need to marginalize over g_i. Using Eq. (1), (2), and (3), the full-conditional for T is

$$
\begin{aligned}
p(T_i = j|U, t, \rho, g, \mu, \Sigma) &\propto \int \sum_{l=1}^{d} p(T_i = j, g_i = l, U, t, \rho, g, \mu, \Sigma) d\rho \\
&\propto \int \sum_{l=1}^{d} p(u_i|T_i = j, g_i = l, t, \rho, g, \mu, \Sigma) p(T_i = j, g_i = l, t, \rho, g, \mu, \Sigma) d\rho \\
&\propto \int \sum_{l=1}^{d} p(u_i|\mu_{jl}, \Sigma_{jl}) p(g_i = l|T_i = j, g) p(T_i = j, t, \rho) d\rho \\
&\propto \sum_{l=1}^{d} p(u_i|\mu_{jl}, \Sigma_{jl}) p(g_i = l|T_i = j, g) \int p(T_i = j, \rho) p(t|\rho) p(\rho) d\rho,
\end{aligned}
\tag{11}
$$

where we integrate over all $\rho \in S^m$. The integral in Eq. (11) can be solved by observing that the expression is the multivariate beta function. Some algebra with gamma functions leads to $(\gamma_{0j} + n_j)$, where $n_j = \sum_{l=1}^{n} \mathrm{I}(t_l = j)$. Additionally let $n_{g_{jl}} = p(g_i = l|T_i = j, g) = |\{g_s = l \wedge t_s = j : s = 1, \ldots, n\}|/|\{t_s = j : s = 1, \ldots, n\}|$, which represents the probability of the new observation being in a specific mixture. Inserting this into Eq. (11) and using Eq. (8), (9), and (10) we construct a Gibbs sampler for Eq. (7)

$$
\begin{aligned}
\mu_{jh}|U, t, \rho, g, \Sigma &\sim \mathcal{N}_{r(m-1)}(\mu^*, \Sigma^*), \\
\Sigma_{jh}|U, t, \rho, g, \mu &\sim \text{inv-Wishart}(\nu^*, S^{*-1}), \\
p(g_i = h|U, t_i = j, \rho, \mu, \Sigma) &\propto p(u_i|\mu_{jh}, \Sigma_{jh}) \tau_{0h}, \\
p(T_i = j|U, t, \rho, g, \mu, \Sigma) &\propto \sum_{l=1}^{d} n_{g_{jl}} p(u_i|\mu_{jl}, \Sigma_{jl})(\gamma_{0j} + n_j),
\end{aligned}
$$

where we use the conditional independence of group memberships in Eq. (10) and sample each g_i separately. Due to an efficient group collapsing property, the number of mixture groups d can be set to some arbitrary high number and the model finds the suitable number automatically. Some groups tend to be closer together at the beginning. This causes the data points to interchange between groups and pairs of groups often merge into one, resulting in fewer mixture components.

If we constrain the covariance matrices Σ to be diagonal, the model is still able to find mixture-components where the correlations are relatively low. This way the correlations are still modelled, while the method becomes less complex. Therefore, we can exchange some flexibility for simplicity, which results in faster inference (inverse of the covariance matrices becomes trivial), while still remaining flexible enough to model the correlations well in most situations. We included this method in the empirical evaluation (MM-diag).

2.1.1. Regularization

To make the method more robust, we implemented regularization by discounting individual dimensions in transformed observation space U. Let λ^* be a vector of length $r(m-1)$ and

$\lambda_i^* \geq 0$. We modify Eq. (11) by changing $p(u_i|\mu_{jl}, \Sigma_{jl})$ to $p(u_i|\mu_{jl}, \Sigma_{jl} + \mathrm{diag}(\lambda^*))$. Adding a positive number to an element on the diagonal increases the variance of the density, reducing the variance of the data in that dimension (and their covariance with all other dimensions). This effectively decreases the differences in density (between observations) in that dimension. That is, it reduces the influence of that dimension on the distribution of T_i.

For each Gibbs iteration, the λ^* for that iteration is determined before predicting for the unobserved data T using

$$p(\lambda|U, t, \rho, g, \mu, \Sigma) \propto p(t_i = j_i|U, t, \rho, g, \mu, \Sigma, \lambda)p(\lambda),$$

where j_i is the true class label for example i.
In general, we could make our predictions by integrating over λ. However, the posterior distribution for λ is intractable and in practice typically very difficult to sample from efficiently. For the purposes of this work, we use only the MAP estimate

$$\lambda^* = \arg\max_{\lambda} \log p(\lambda) + \sum_{i=1}^{n} \log p(t_i = j_i|U, t, \rho, g, \mu, \Sigma, \lambda).$$

2.2. Related Work

To empirically evaluate our method, we also implemented four Bayesian approaches to combining classifiers and the classical linear opinion pool.

2.2.1. Supra-Bayesian

Lindley (1985) presented the supra-Bayesian method for combining probabilistic predictions of a categorical variable of r experts. The method relies on Bayes' theorem to calculate the posterior distribution of predictions of the decision maker, given the predictions of the experts. Let $A_1, A_2, ..., A_m$ be the possible outcomes of the categorical variable and let $S_1, S_2, ..., S_r$ be classifiers. Let H be the decision maker's prior probability for the response and $q_{ij} = \log(\mathrm{Pr}_i(A_j))$ the log-probability assigned to true label A_j by the $i-$th classifier. Then the decision maker updates his probabilities via the Bayes' theorem

$$p(A_j|Q, H) \propto p(Q|A_j, H)p(A_j|H). \tag{12}$$

To model the correlations between given predictions, Lindley proposes a MVN distribution for $p(Q|A, H)$. Additionally, Lindley argues that the classifiers' belief in their prediction is independent of the label of their prediction, therefore the MVN distributions share a common covariance matrix between predictions for different labels. The author also proposes the modeling of log-odds instead of log-probabilities, as log-odds are expected to be distributed normally. We implemented a ML estimation of the multivariate parameters for Eq. (12), where the matrix Q represented log-odds.

Note that the described method is the one proposed by Lindley (1985) and that supra-Bayesian methods are a very general term for combination methods that rely on the Bayes' theorem. By selecting the appropriate likelihoods we arrive at IBCC and (unregularized) MM as special cases.

2.2.2. IBCC

IBCC (Kim and Ghahramani, 2012) assumes that the true labels t are generated by a categorical distribution with parameter κ. The prediction $c_i^{(k)}$ of individual classifier k is also generated by a categorical distribution with parameters $\pi_j^{(k)}$ for each possible value j of the true label. Let ν be the prior for κ and $\alpha_{0,j}^{(k)}$ prior for $\pi_j^{(k)}$. Kim and Ghahramani additionally put an exponential prior on $\alpha_{0,j}^{(k)}$, but Simpson (2014) observed that the model without this additional flexibility performs comparably well and derived the Gibbs sampler for κ, Π, and the true labels

$$p(\kappa|t,\nu_0) = \frac{1}{B(\nu)}\prod_{j=1}^{m}\kappa_j^{\nu_j},$$

$$p(\pi_j^{(k)}|t,c,\alpha_{0,j}^{(k)}) = \frac{1}{B(\alpha_{t_i}^{(k)})}\prod_{l=1}^{m}\left(\pi_{j,l}^{(k)}\right)^{\alpha_{j,l}^{(k)}},$$

$$p(t_i = j|\Pi,\kappa,c) \propto \kappa_j \prod_{k=1}^{r}\pi_{j,c_i^{(k)}}^{(k)},$$

where $\nu_j = \nu_{0,j} + \sum_{i=1}^{n+n^*}\delta(t_i - j)$ and $\alpha_{j,l}^{(k)} = \alpha_{0,j,l}^{(k)} + \sum_{i=1}^{n+n^*}\delta(t_i - j)\delta(c_i^{(k)} - l)$.

2.2.3. IBCC QPI

The IBCC method can be extended to allow for quasi-probabilistic inputs (QPI). Probabilistic classifiers provide us with probabilities of each outcome $p_i^{(k)} = \left[p_{i1}^{(k)} \;\cdots\; p_{im}^{(k)}\right]^T$. For the IBCC to exploit probabilistic predictions, we can first take a sample from a categorical distribution with parameter $p_i^{(k)}$ for each source k. The categorical samples can then be represented as binary vectors, which are of appropriate form for the inputs of IBCC. We then take n_{tsamp} such samples and use them to create a new training set of size $n \times n_{\text{tsamp}}$.

For prediction, we use the same procedure. We sample according to the sources' probabilistic predictions. We retrieve the final prediction as the average probabilistic prediction over all samples. This allows the IBCC to take advantage of probabilistic predictions, instead of simply using the class with the highest probability, as the sources' predictions.

2.2.4. Agnostic Bayes

The agnostic Bayes approach (Lacoste et al., 2014) combines sources based on the probability of being the *best* source, where best is defined in terms of some task-dependent measure. The combined prediction is defined as

$$\hat{p}_i = \sum_{k=1}^{r} P(k^\star = k|p,t)p_i^{(k)},$$

where $P(k^\star = k|p,t)$ is the probability of source k being the best source, given the available data (all predictions p and observed values t). In our case of categorical target variables and log-score performance measure, we can simplify $P(k^\star = k|p,t) = P(\forall k' \in 1..r : E[l^{(k)}] \geq E[l^{(k')}]|l) = P(k|l)$, where $E[l^{(k)}]$ is the $k-$th source's expected log-score and l all the observed log-scores (for this model, a sufficient statistic for p and t).

We draw samples from $P(k|l)$ with bootstrap inference (this was shown to be very effective in Lacoste et al. 2014, at least as effective as placing priors over joint risk and performing Bayesian inference). That is, we draw a replacement set of observations and compute the average log-score. This gives us the best source on this replacement set—one sample from $P(k|l)$. The process is repeated for several iterations to obtain a set of samples that approximate $P(k|l)$.

2.2.5. Linear Opinion Pool (LOP)

The linear opinion pool (Cooke et al., 1991) is the classical approach to combining predictions. The combined prediction $\hat{p}_i$ is expressed as a linear combination of individual sources

$$\hat{p}_i = X_i \beta,$$

where $X_i = \begin{bmatrix} p_i^{(r)} & p_i^{(r-1)} & \cdots & p_i^{(1)} & p_\cdot^{(0)} \end{bmatrix}$ and $\beta = \begin{bmatrix} \beta_r & \beta_{r-1} & \cdots & \beta_1 & \beta_0 \end{bmatrix}^T$.

The parameters β_i and $p_\cdot^{(0)}$ were fit by maximizing the likelihood, subject to $0 \leq \beta_i \leq 1$, $\sum_{i=0}^{r} \beta_i = 1$, $0 \leq p_{ij}^{(0)} \leq 1$, and $\sum_{j=1}^{m} p_{\cdot j}^{(0)} = 1$.

We also included a combination of LOP and MM, where the probabilistic predictions obtained by MM are added to the sources before LOP is applied (LOP+MM).

2.2.6. Discussion of Related Work

We can divide the methods for classifier combination into two groups. Methods that aim to estimate the performance of each classifier and then weigh their predictions by their performance (LOP, agnostic-Bayes, etc.). And methods that aim to learn the latent structure of the classifications and then provide probabilistic classifications (IBCC and extensions, supra-Bayesian, etc.).

The former are expected to perform poorly when there is a latent structure to learn, for example, biased classifiers or systematic errors in classifiers. On the other hand, these methods have an advantage that they are not expected to perform discernibly worse than the best source. The latter are the exact opposite—they can learn complex latent structure and perform better than the first group, but can also perform extremely poorly, discernibly worse than the best source. In Section 3 we present 5 toy data sets, which serve to show the differences in performance of these groups, depending on the structure of the data.

The main difference between IBCC and our method (MM) lies in the conditional likelihood for the true label. IBCC models sources' predictions with a multinomial distribution, while MM assumes a latent space of transformed predictions and models it with multivariate normal mixtures. The advantage of this is that it allows us to model both probabilistic predictions and correlations between sources, while the computational complexity is quadratic in the number of sources and outcomes (as opposed to DBCC, where it is exponential).

IBCC QPI is similar to MM in that it is able to exploit probabilistic nature of predictions, however it does not model correlations.

Nazábal et al. (2016) used a Dirichlet distribution to model probabilistic outputs of the sources. A similar (but not equivalent) model would result if we used diagonal covariances without mixtures in the proposed MM. Therefore the proposed MM improves on the drawbacks of Nazábal et al. as it allows for the modeling of correlations between sources, and additionally adds flexibility through the use of mixtures.

MM could also be interpreted as a regularized supra-Bayesian method with MVN-mixture likelihood. The modeling of log-odds described by Lindley (1985) is equivalent to the modeling of the transformed classifiers by MM. A major difference between the methods is that MM does not assume the same covariances over all true labels. Furthermore, each mixture component has its own covariance matrix, which further improves the flexibility of our method.

3. Empirical Evaluation

We empirically evaluated and compared the methods on several toy and real-world data sets. We estimated out-of-sample log-score using train-test splits. To compare the best-performing method with the rest, we used the differences between log-scores. Let a be the vector of length n_t of log-scores of the best performing method, b the vector of log-scores of another method, and $d = a - b$. If

$$\left| \frac{1}{n_t} \sum_{i=1}^{n_t} d_i \right| > 2\sqrt{\frac{\mathrm{VAR}[d]}{n_t}},$$

we argue that there is a discernible difference in performance. Details on how the data were generated and split are provided below.

For IBCC and its extension, we selected priors that assume each class has the same prior probability. We used the same priors ($\alpha_{0,j} = 1$) for all confusion matrices, which represents a weak belief that the models are random. For MM we selected vague priors that put prior belief that the mean values of transformed predictions are zero with large variances. However, MM has proved to be somewhat robust to the prior selection. We set the same priors for Bayesian methods over all data sets. For λ, we use uniform priors $\lambda_i \sim_{\text{iid}} \mathrm{U}(0, 10^5)$. We set the maximum number of mixture components for MM to 15, however this number dynamically falls, depending on the problem, as we described in Section 2.1.

3.1. Data Sets

In this section we describe the data sets we used for empirical evaluation.

3.1.1. Toy Data Sets

Each of the 5 toy data sets is a combination of the same underlying data set and two sources of probabilistic predictions from a set of sources with different properties. The purpose of the toy data sets is to illustrate some of the shortcomings of the methods for combining categorical predictions.

The data set has a target variable with three possible outcomes. The target variable's value is generated for each sample i separately by first drawing the underlying probability vector p_i^* from Dirichlet$(1.5, 0.9, 0.6)$ and then drawing the target value from Categorical(p_i^*). We generated 500 samples for training and 5000 samples for testing.

We then generated the following sources of probabilities, based on the underlying probability vectors p^*:

- **Random 1 & 2:** The probabilities are drawn independently for each sample from Dirichlet$(1, 1, 1)$.
- **Good 1 & 2:** The probabilities are drawn independently for each sample from Dirichlet$(10p_i^*)$.
- **Permuted 1:** The probabilities are drawn independently for each sample from Dirichlet$(100p_i^*)$ and permuted (2, 3, 1). This results in a source that gives very accurate probabilities, but does not correctly assign them to the possible values of the target variable.
- **Permuted 2:** Identical to Good 2, but permuted (2, 3, 1).

The two sources, Source 1 and Source 2, were assigned as follows: **toy A** (Random 1, Random 2), **toy B** (Good 1, Random 1), **toy C** (Good 1, Good 2), **toy D** (Good 1, Permuted 1), and **toy E** (Good 1, Permuted 2).

3.1.2. Bookies

Each sample in this data set is a football game and the four sources of probabilistic predictions are four major online bookmakers. Probabilistic predictions were calculated by normalizing the reciprocals of odds offered for the outcome of the game (home, draw, away). Therefore, the target variable has 3 outcomes. The data set has 5434 samples, we used 2000 for training and 3434 for testing. Bookmakers are very good sources of probabilistic predictions (Forrest et al., 2005) and their predictions are highly correlated—the lowest correlation is 0.92. We do not expect any ensemble method to outperform the best source.

3.1.3. DNA

Two data sets were constructed using the StatLog DNA data set in a way similar to the experiments in Kim and Ghahramani (2012).

DNA A: 2000 samples were partitioned into 5 partitions, 400 samples each, and a C4.5 decision tree classifier was trained on each partition. These classifiers were used to classify the remaining 1186 samples, 400 of which were used for training and 786 for testing.

DNA B: 2000 samples were used to train 5 different classifiers: a multinomial logistic regression (Source 1), linear discriminant analysis (Source 2), a decision tree classifier (Source 3), a random forest classifier (Source 4), and a k-nearest-neighbour classifier (Source 5, k = 50). These classifiers were used to classify the remaining 1186 samples, 400 of which were used for training and 786 for testing.

3.2. Results

The results are shown in Table 1. No method clearly outperforms other methods over all data sets. LOP and MM perform well on most data sets. For LOP the exceptions are data sets with a more complex relationship between the sources' predictions and the outcome (toy D). And for MM the exceptions are the two simple toy data sets with a continuous relationship between the sources' predictions and the probability of the outcome (toy B, toy C). MM-diag performs worse than the MM on all data sets, however, the differences are small on average. A combination of LOP and MM nullifies the drawbacks of each individual method and performs the best on average. Toy D, with systematic errors in otherwise good sources, is a good example of a data set where the methods that rely on finding the best source are inferior to models that learn the latent structure, providing a good argument for the latter.

Agnostic Bayes will assign a lot of weight to the best source, unless it is difficult to discern the best source. The latter is in cases where there is not a lot of data, relative to the magnitude of the differences between the sources in terms of performance (toy A, DNA A). Therefore, as expected, it performs well when the most accurate source performs well (toy B, toy C, toy E, bookies), and performs poorly, when it does not (DNA B). Note that there is no correlation between the true labels and sources in the toy A data set—the best possible model would assign the same probabilities to all outcomes. Agnostic Bayes performs poorly because it tries to identify the best performing model, but none of the sources perform well. The same would happen to LOP if we did not include an intercept term.

IBCC performs poorly on the toy data sets but performs well on the DNA data sets. However, it is strictly outperformed by MM. The quasi-probabilistic IBCC outperforms the IBCC on DNA data sets and performs similarly on most of the toy data sets. Furthermore, it performs well on the toy A data set, where it is the best performing model, with the proposed MM showing a similar level of performance. IBCC QPI therefore proved to be better suited for probabilistic input than the IBCC, as expected. The supra-Bayesian approach of Lindley (1985), which can be viewed as a very simple case of the proposed approach, performs the worst on average, among all the Bayesian approaches.

As expected, all methods fail to outperform the best source on the bookies data set, which has very accurate and highly correlated sources. This relates heavily to the theoretical results mentioned in the introduction—the data set lacks diversity to exploit. Methods that do not model correlations perform exceptionally poorly. However, methods that weigh/select sources perform at least as good as the best source.

toy A	toy B	toy C	toy D	toy E	bookies	DNA A	DNA B
IBCC QPI	agnostic	LOP	MM	LOP+MM	Source 1	LOP+MM	LOP+MM
−1.03 ± 0.005	−0.831 ± 0.009	−0.81 ± 0.008	−0.001 ± 0	−0.821 ± 0.008	−0.965 ± 0.008	−0.197 ± 0.022	−0.129 ± 0.017
MM	Source 1	LOP+MM	MM-diag	agnostic	agnostic	MM	MM
−1.031 ± 0.005	−0.831 ± 0.009	−0.811 ± 0.008	−0.002 ± 0	−0.826 ± 0.009	−0.966 ± 0.008	−0.206 ± 0.025	−0.132 ± 0.028
MM-diag	LOP	agnostic	LOP+MM	Source 1	LOP+MM	LOP	IBCC QPI
−1.032 ± 0.006	−0.831 ± 0.009	−0.822 ± 0.009	−0.007 ± 0	−0.826 ± 0.009	−0.967 ± 0.008	−0.22 ± 0.026	−0.134 ± 0.023
LOP	LOP+MM	Source 1	IBCC	MM	LOP	agnostic	MM-diag
−1.032 ± 0.005	−0.831 ± 0.009	−0.826 ± 0.009	−0.018 ± 0	−0.83 ± 0.007	−0.967 ± 0.007	−0.22 ± 0.02	−0.139 ± 0.027
LOP+MM	Supra	MM	Supra	LOP	Source 3	MM-diag	LOP
−1.032 ± 0.006	−0.892 ± 0.008	−0.828 ± 0.007	−0.111 ± 0.001	−0.83 ± 0.008	−0.968 ± 0.007	−0.222 ± 0.03	−0.157 ± 0.016
IBCC	MM	Source 2	IBCC QPI	MM-diag	MM	IBCC QPI	IBCC
−1.043 ± 0.006	−0.893 ± 0.006	−0.832 ± 0.009	−0.204 ± 0.001	−0.838 ± 0.007	−0.968 ± 0.008	−0.248 ± 0.033	−0.181 ± 0.033
Supra	MM-diag	MM-diag	agnostic	Supra	Source 4	IBCC	agnostic
−1.107 ± 0.002	−0.906 ± 0.006	−0.839 ± 0.008	−0.824 ± 0.009	−0.883 ± 0.007	−0.968 ± 0.007	−0.271 ± 0.037	−0.182 ± 0.033
agnostic	IBCC	Supra	Source 1	IBCC QPI	Source 2	Supra	Source 2
−1.259 ± 0.009	−0.921 ± 0.008	−0.883 ± 0.007	−0.824 ± 0.009	−0.908 ± 0.005	−0.969 ± 0.007	−0.275 ± 0.017	−0.182 ± 0.033
Source 2	IBCC QPI	IBCC QPI	LOP	IBCC	MM-diag	Source 4	Supra
−1.469 ± 0.015	−0.953 ± 0.005	−0.907 ± 0.005	−0.835 ± 0.008	−0.92 ± 0.012	−0.969 ± 0.008	−0.461 ± 0.055	−0.214 ± 0.013
Source 1	Source 2	IBCC	Source 2	Source 2	IBCC QPI	Source 5	Source 4
−1.508 ± 0.016	−1.469 ± 0.015	−0.919 ± 0.012	−3.081 ± 0.006	−1.43 ± 0.011	−0.993 ± 0.006	−0.767 ± 0.074	−0.311 ± 0.011
					Supra	Source 3	Source 3
					−1.008 ± 0.007	−0.784 ± 0.075	−0.321 ± 0.046
					IBCC	Source 2	Source 5
					−1.245 ± 0.02	−0.846 ± 0.078	−0.587 ± 0.01
						Source 1	Source 1
						−0.879 ± 0.079	−4.305 ± 0.717

Table 1: Estimated log-scores and standard errors on toy and real-world data sets. For each data set, the methods are ordered in descending order of performance. Highlighted methods are not discernibly worse than the best-performing method for that data set.

3.3. Case Study—Prediction of Air-Pollutant Concentration

Air pollution poses a great risk for public health (World Health Organization, 2014). Forecasting excessive air pollution is an important task, also regulated by a EU directive (European Council, 2008). In Slovenia, the Slovenian Environment Agency (ARSO) is tasked with forecasting daily average concentration of particulate matter (PM_{10}) and daily maximum concentration of tropospheric ozone (O_3) for the same day (in the morning) and the next day. Currently, models and human experts are used for this task. Faganeli Pucer et al. (2018) proposed a Bayesian methodology for this problem and empirically compared several machine learning methods and human experts. Gaussian processes performed the best on average, but the question remains whether an ensemble method can be used to improve on individual models. With that aim, we evaluate methods from Section 2 on four air-pollution data sets used in Faganeli Pucer et al. (2018).

The data sets consist of measurements of daily levels of PM_{10} and O_3 on several stations across Slovenia from 2013 to 2016. The data were provided by ARSO. According to reporting guidelines, PM_{10} levels are categorized into three categories: 0-35 $\mu g/m^3$, 35-50 $\mu g/m^3$, greater than 50 $\mu g/m^3$, and O_3 levels are categorized into four categories: 0-60 $\mu g/m^3$, 60-120 $\mu g/m^3$, 120-180 $\mu g/m^3$, greater than 180 $\mu g/m^3$.

Three prediction models—Bayesian lasso, random forests, and Gaussian processes—and human expert predictions were used to predict the levels for the same and next day, depending on covariates available at the current day. Expert predictions were non-probabilistic (categorical). The models were trained in a time-respecting manner (predictions for 2014 were obtained by training the models on data from 2013, predictions for 2015 were obtained by training the models on data from 2013 and 2014, etc.). We used two thirds of observations for training and the rest for testing.

The empirical results for the case study are shown in Table 2. The combination of MM and LOP performs the best over all data sets. MM and LOP are the only methods that show no discernible differences to LOP+MM over all data sets. This provides a strong argument for our method, compared to other methods which learn the latent structure of the data. MM also strictly outperforms MM-diag, as expected. IBCC QPI performs well on the PM_{10} data sets, however it does not outperform MM. Agnostic Bayes performs slightly better than the best source on all data sets, however it is discernibly worse than the best method on two data sets. The other methods are unable to outperform the best performing individual classifier on all data sets.

The major arguments for our method are twofold. First, it is a step forward in the state-of-the-art methods which rely on learning the latent structure of the data (IBCC, IBCC QPI, Supra). This can be seen as it is the only such method which is never discernibly worse than the best performing method. Second, LOP+MM performs the best on all data sets in the case study. It also outperforms LOP on all real-world data sets, which is discernibly better—it is highly unlikely that this is due to chance.

4. Conclusion

We proposed a new Bayesian method for combining probabilistic predictions, based on MVN mixtures. Our method improves on the IBCC (and DBCC) as it is able to model probabilistic classifiers, and on the work of Nazábal et al. (2016) as it also models the

PM_{10} tomorrow	**PM_{10} today**	**O_3 today**	**O_3 tomorrow**
LOP+MM	LOP+MM	LOP+MM	LOP+MM
−0.613 ± 0.022	−0.406 ± 0.019	−0.375 ± 0.025	−0.439 ± 0.027
MM	MM	LOP	LOP
−0.613 ± 0.023	−0.409 ± 0.021	−0.378 ± 0.024	−0.442 ± 0.028
MM-diag	LOP	MM	MM
−0.62 ± 0.023	−0.419 ± 0.019	−0.389 ± 0.028	−0.446 ± 0.029
IBCC QPI	MM-diag	agnostic	agnostic
−0.62 ± 0.024	−0.419 ± 0.021	−0.393 ± 0.027	−0.463 ± 0.028
LOP	IBCC QPI	Source 1	Source 1
−0.634 ± 0.02	−0.436 ± 0.024	−0.393 ± 0.027	−0.463 ± 0.028
agnostic	agnostic	MM-diag	MM-diag
−0.684 ± 0.026	−0.442 ± 0.023	−0.393 ± 0.027	−0.471 ± 0.034
Source 2	Source 1	IBCC QPI	IBCC QPI
−0.684 ± 0.026	−0.442 ± 0.023	−0.394 ± 0.028	−0.49 ± 0.032
Source 1	Source 2	Source 2	Source 2
−0.703 ± 0.019	−0.516 ± 0.015	−0.462 ± 0.02	−0.508 ± 0.021
Supra	Supra	Source 3	Source 3
−0.732 ± 0.017	−0.533 ± 0.018	−0.511 ± 0.033	−0.53 ± 0.022
IBCC	Source 3	IBCC	Supra
−0.791 ± 0.038	−0.641 ± 0.041	−0.563 ± 0.053	−0.808 ± 0.031
Source 3	IBCC	Supra	IBCC
−0.896 ± 0.042	−0.673 ± 0.041	−0.683 ± 0.036	−0.83 ± 0.067
Source 4	Source 4	Source 4	Source 4
−2.23 ± 0.09	−1.578 ± 0.081	−1.18 ± 0.101	−1.371 ± 0.108

Table 2: Estimated log-scores and standard errors on air-pollution data sets. For each data set, the methods are ordered in descending order of performance. Highlighted methods are not discernibly worse than the best-performing method for that data set. The order of magnitude of the differences is in percent, which shows a practical significance in the usefulness of the best performing methods.

correlations between the classifiers. We derived an efficient Gibbs sampler for our method, along with a regularization method for robustness.

The results on toy data sets highlighted that there is no single method that performs well over several diverse data sets. However, our method outperforms related Bayesian methods on all but one real-world data sets. The method proved to be especially useful in the case study, where we combined three machine learning methods and human expert predictions for air-pollutant concentration forecasting. This case study illustrates the main use-case for our method—a relatively small number of probabilistic or crisp (0/1) sources, which are potentially flawed (biased, etc.) and correlated. Our method is also very robust and requires practically no tuning. Additionally, a combination of linear opinion pool and our method proved to be very successful, outperforming all methods on all but one real-world data sets, and also being robust on the difficult toy data sets.

Our method still has some drawbacks, however. The original model is based on full covariance matrices, so the number of the parameters grows quadratically with the number of possible outcomes and the number of classifiers. While this is an improvement over the exponential growth in the number of classifiers of the DBCC model, it still leads to very complex models in high dimensions. One possibility of at least partially regulating this drawback is to constrain the covariance matrices to being diagonal—as in our empirical evaluation. Alternatively, we could assume the same covariance matrix over all mixture components in a true label, or assume the same variances/covariances over several predictions. We leave the analysis of possible constraints for future work. As part of future work, we will also explore if the predictive performance or robustness of the method could be improved by using Gaussian processes and/or variational autoencoders instead of multivariate normal mixtures to model the sources' transformed probabilities. Additionally, we could limit the computational complexity by using sparse Gaussian processes.

As is the case in other areas of learning, there is no single best method of combining probabilistic predictions. In some cases, the best approach is to combine the sources based on their predictive performance or to directly maximize the predictive performance of the combination, such as agnostic Bayes or the linear opinion pool. However, sometimes there are more complex relationships between classifiers and the target variable and methods that are able to learn these patterns will perform better. In such cases, our method is in several ways a superior alternative to existing Bayesian approaches. It is able to model complex relationships, while still being robust and easy to tune, thus providing us with a useful approach for combining probabilistic predictions.

Acknowledgments

The authors would like to acknowledge Jana Faganeli Pucer and the Slovenian Environment Agency (ARSO), who provided environmental pollution data sets. Our research was partially supported by the Slovenian Research Agency (ARRS project grant L1-7542 and research core funding P5-0410). Gregor Pirš is supported by the ARRS Young researcher grant.

References

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Jesús Cerquides and Ramon López De Mántaras. Robust Bayesian linear classifier ensembles. In *European Conference on Machine Learning*, pages 72–83. Springer, 2005.

Roger Cooke et al. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press on Demand, 1991.

European Council. Directive 2008/56/EC of the European Parliament and of the Council. *Decision of Council*, 2008.

Jana Faganeli Pucer, Gregor Pirš, and Erik Štrumbelj. A Bayesian approach to forecasting daily air-pollutant levels. *Knowledge and Information Systems*, pages 1–20, 2018.

David Forrest, John Goddard, and Robert Simmons. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3):551–564, 2005.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian Classifier Combination. In *International Conference on Artificial Intelligence and Statistics*, pages 619–627, 2012.

Alexandre Lacoste, Mario Marchand, François Laviolette, and Hugo Larochelle. Agnostic Bayesian learning of ensembles. In *International Conference on Machine Learning*, pages 611–619, 2014.

Dennis V. Lindley. *Bayesian Statistics 2*, chapter Reconciliation of discrete probability distributions, pages 375–391. North-Holland, 1985.

Alfredo Nazábal, Pablo García-Moreno, Antonio Artés-Rodríguez, and Zoubin Ghahramani. Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1342–1351, 2016.

SIGKDD. KDD Cup 2017: Highway tollgates traffic flow prediction. `http://www.kdd.org/kdd2017/News/view/announcing-kdd-cup-2017-highway-tollgates-traffic-flow-prediction`, 2017. Accessed: 2018-04-05.

Edwin Simpson. *Combined decision making with multiple agents.* PhD thesis, University of Oxford, 2014.

Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. Dynamic Bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks*, pages 90–95. IEEE, 1996.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based Bayesian aggregation models for crowdsourcing. In *International Conference on World Wide Web*, pages 155–164. ACM, 2014.

World Health Organization. Ambient (outdoor) air quality and health. *Fact sheet*, (313), 2014.

Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012.