Supra Bayesian Classifier Combination

Article · June 2006

CITATIONS
0

READS
101

1 author:

Mohamed Waleed Fakhr
Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt
115 PUBLICATIONS **251** CITATIONS

SEE PROFILE

www.icgst.com

ICGST – Supra Bayesian Classifier Combination

Waleed Fakhr
Computer Engineering Department, Arab Academy for Science and Technology
Heliopolis, Cairo, Egypt
waleedf@aast.edu
www.aast.edu

Abstract

Model combination in classification, density estimation and clustering and forecasting has become a promising research area where better results are obtained compared to using single models. Bayesian inference offers two approaches for combining models; Bayesian model averaging (BMA), and Supra-Bayesian classifier combination (SBCC). In this paper we briefly overview the BMA, show its drawbacks, and propose a novel idea of Bayesian data/model averaging. Then, the SBCC technique is explained, and its implementation used in this work is presented. A novel approach is proposed where input space is clustered before applying the SBCC. A 2-dimensional, 2-class, 4-experts problem is used to test the SBCC and SBCC-clustering algorithms, where better results are obtained over 300 runs for the data set, compared to an averaging technique, where the 4 experts' performance is averaged.

Keywords: *Bayesian model averaging, Bayesian data averaging, Supra Bayesian classifier combination, Supra Bayesian combination with clustering, K-means clustering, Multilayer perceptrons, Probabilistic neural networks.*

1. Introduction

In decision making problems, it is intuitive that a combination of experts can perform better than any single one alone, conditioned that the decision maker (DM), has the right tools to combine their individual opinions.

Many approaches have emerged in the past for combining experts' opinions, whether these experts are classifiers [1-5], forecasters [6], clustering models [7], or human experts [8].

Among these approaches, the Bayesian ones are the most theoretically motivated, since they produce a formal probabilistic interpretation to the combination process [1,4,8,9-11].

There are two Bayesian model combination schools; the Bayesian model averaging (BMA) [12,13], and the supra-Bayesian classifier combination (SBCC) [3,10,11].

Section 2 in this paper reviews the (BMA) approach and its potential drawbacks. A novel methodology for averaging over different data sets and different models is presented. Section 3 introduces the supra-Bayesian model combination approach, in the context of classification, where it has been called the Bayesian classifier combination (BCC) by Gharamani [3]. We explain the technique we adopt in this paper (supra Bayesian classifier combination SBCC), which is similar to the techniques used by Jacobs [10] and Bahler et al. [11] for evaluating the DM belief, or aggregating the experts' opinions. Experiments are discussed in section 4, and a proposed method to improve the performance of the SBCC using K-means clustering of the input space also evaluated (it is called SBCC-clustering). Finally, section 5 gives a summary and conclusions. Section 6 has the references.

2. Bayesian Model Averaging

Bayesian model averaging (BMA) [12,13] is the natural extension of the Bayesian inference approach, where each model may be viewed as a combination of different models representing the peaks in the parameter posterior distribution.

In the BMA, this idea is extended to the "many" models case. The combined model is the weighted average of the models used, each is weighted by its evidence of having generated the data.

This is summarized in the following.

Bayesian Model Averaging: Over Models

The well known BMA [12,13] assumes that "models" is a discrete space with *L* models. Assuming a data set *D* is used as a reference (training data), and an application domain *E* then our predictions for a test pattern *x* is:

$$P(xi \mid D, E) = \sum_{l=1}^{L} P(M_l \mid D, E) P(xi \mid M_l, D, E)$$

where

$$P(M_l \mid D, E) = \frac{P(Ml \mid E) P(D \mid Ml \mid E)}{P(D \mid E)}$$

$$P(D \mid E) = \sum_{l=1}^{L} P(Ml \mid E) P(D \mid E, Ml)$$

$$P(D \mid E, M_l) = \int_{\theta} P(D \mid M_l, \theta, E) P(\theta \mid E, M_l) d\theta$$

Thus we get a weighted average of models' predictions by their normalized evidences.

Bayesian Model Averaging: Over Data Sets

Assuming that data sets within the application *E* is also a discrete space with *J* sets, then we can extend the Bayesian model averaging to an averaging over all data sets:

$$P(x \mid E) = \sum_{j=1}^{J} P(x \mid Dj, E) P(Dj \mid E)$$

Where,

$$P(x \mid Dj, E) = \sum_{l=1}^{L} P(Ml \mid Dj, E) P(x \mid Ml, Dj, E)$$

And

$$P(M_l \mid Dj, E) = \frac{P(Ml \mid E) P(Dj \mid Ml \mid E)}{P(Dj \mid E)}$$

$$P(Dj \mid E) = \sum_{l=1}^{L} P(Ml \mid E) P(Dj \mid E, Ml)$$

$$P(Dj \mid E, M_l) = \int_{\theta} P(Dj \mid M_l, \theta, E) P(\theta \mid E, M_l) d\theta$$

This novel treatment by averaging over data sets eliminates the conditioning on a particular data set, thus models trained by different data sets may be combined.

Bayesian Model Averaging: The Drawbacks

In conclusion, Bayesian model averaging may be applied for combining probability density estimators, classifiers and predictors, under the condition that the models are probabilistic.

In many applications, particularly in classification, some models may produce discrete outcomes, not associated with a probabilistic model. Thus, a more general framework is needed in this case. The Supra-Bayesian framework would solve this problem.

3. Supra-Bayesian Aggregation of Expert Opinions

Another approach for combining classifiers has been known in literature as the "Supra Bayesian" method. Also known as, Aggregating expert opinions, and Pooling expert opinions: [3,10,11].

In this school of thought, we have *K* experts, each expresses its opinion about a specific event (for example, an input data to be classified into *J* classes). Each opinion is associated with some uncertainty (in the form of a probability distribution).

There is a decision maker (DM) who receives all the opinions (distributions), and aggregates them into one distribution to make the decision.

In other words, from the DM point of view, the opinions expressed by the experts are "data".

Consequently, a DM using a supra-Bayesian method combines the probability distributions provided by the experts with its own (the DM's) prior distribution (prior knowledge) using the Bayes rule.

Let Cj be the quantity of interest (e.g., a class label). Let there be *K* experts.

Each expert expresses opinion about the quantity with a certain degree of uncertainty. This is expressed by a probability distribution $P(Cj|Hi)$ for expert i.

Thus the (DM) actually sees these probabilities, and may not see the inputs of the experts represented here by the knowledge Hi of each expert.

For the DM, the values received can be denoted by $Pi=Pi(Cj|Hi)$ $i=1,2,....,K$

The knowledge at the input of each expert may be input data vectors, or features of any form (numeric, fuzzy, opinions, etc.)

The DM has his own prior probability about the quantity of interest. This is expresses by $P(Cj|\rho)$, where ρ represents the DM prior knowledge about the quantity Cj.

Now we can construct the DM's posterior probability about the quantity of interest: $P(Cj|\rho,P1,P2,.....,PK)$, which is the DM's posterior probability about Cj given its prior knowledge, and the experts opinions.
This posterior probability is given by the Bayes rule:

$$P(Cj|\rho,P1,P2,.....,PK)=\frac{P(P1,P2,.....,PK|\rho,Cj)P(Cj|\rho)}{P(P1,P2,.....,PK|\rho)}$$

or

$$P(Cj|\rho,P1,P2,.....,PK)=\frac{P(P1,P2,.....,PK|\rho,Cj)P(Cj|\rho)}{\sum_{i=1}^{J} P(P1,P2,.....,PK|\rho,Ci)P(Ci|\rho)}$$

where we assume that there are *J* outcomes (classes), and this formula is evaluated for *j=1, 2, ……,J*.
One approximation to the above is to assume that *experts are independent* (classifiers perform independently from each other). In such a case, the DM belief about class Cj becomes;

$$P(Cj\,|\,\rho,P1,P2,.....,PK)=\frac{\prod_{k=1}^{K} P(Pk|Cj,\rho)P(Cj|\rho)}{\sum_{i=1}^{J}\prod_{k=1}^{K} P(Pk|Ci,\rho)P(Ci|\rho)}$$

The left hand side is the DM belief that the true outcome is Cj given the experts' opinions.
Here we have two Approaches to proceed:

First Approach
The first approach treats the outcomes of the experts as discrete values, i.e. based on certain thresholds, the outcome is assigned to one of the *J*-classes (e.g. Pk=A for class-1, Pk=B for class-2, etc.)

In this case, the probabilities $P(Pk\,|\,Cj,\rho)$ define the confusion matrices for the K classifiers.
For example, for a 2-class case, we have 4 probabilities for the outcome of each expert:

P(Pk declares A| True label is A)
P(Pk declares A| True label is B)
P(Pk declares B| True label is A)
P(Pk declares B| True label is B)

In this case, the probabilities $P(Pk\,|\,Cj,\rho)$ can be modelled as multinomial distributions on the Pk's, since they are discrete values taking one of *J* discrete values [3].
The values for those multinomial distributions may be estimated in two techniques. The first is using a validation data in an ML-based sense [11], since for example,

$$P(Pk \text{ declares } A|\ \text{True label is } A)=\frac{\text{number of times Pk declares A when true class}A}{\text{number of times true class is A}}$$

$$P(Pk \text{ declares } A|\ \text{True label is } B)=\frac{\text{number of times Pk declares A when true class}B}{\text{number of times true class is B}}$$

And so on for each expert and for each class label. *(This is the approach taken in this paper).*
For example, for the DM belief for class-1 (event A), given that *P1=A* and *P2=B*, we calculate the following:

$$P(A\,|\,\rho,P1,P2)=\frac{P(P1{=}A|A,\rho)P(P2{=}B|A,\rho)}{P(P1{=}A|A,\rho)P(P2{=}B|A,\rho)+P(P1{=}A|B,\rho)P(P2{=}B|B,\rho)}$$

For DM belief for class-1 (event A) given that P1=A and P2=A;

$$P(A\,|\,\rho,P1,P2)=\frac{P(P1{=}A|A,\rho)P(P2{=}A|A,\rho)}{P(P1{=}A|A,\rho)P(P2{=}A|A,\rho)+P(P1{=}A|B,\rho)P(P2{=}A|B,\rho)}$$

For DM belief for class-2 (event B) given that P1=A and P2=A;

$$P(B\,|\,\rho,P1,P2)=\frac{P(P1{=}A|B,\rho)P(P2{=}A|B,\rho)}{P(P1{=}A|A,\rho)P(P2{=}A|A,\rho)+P(P1{=}A|B,\rho)P(P2{=}A|B,\rho)}$$

And so on.

In the second technique [3,8], we construct a likelihood function over the whole training data, where $P(Pk_n|Cj_n,\rho,\Theta)$, are multinomial distributions with unknown parameters Θ:

$$P(C,Pk's|\Theta)=\prod_{i=1}^{N}\prod_{k=1}^{K} P(Cj)\ P(Pk_n|Cj_n,\rho,\Theta)$$

And we apply the EM algorithm, where the true labels are assumed hidden, and ML is used to estimate the parameters of the multinomial distributions, and the product is over all the data.

Second Approach
The second approach takes the actual values for Pk (which are the expert's output and is the posterior probability of the class-j given

the expert's input Hi), and constructs a probability distribution $P(Pk|Cj,\rho)$ using the continuous values of the Pk.
So, for a two-class case, we construct a distribution for all the values of Pk declaring class-A when true class is A, and one for all the values of Pk declaring class-B when true class is A, and so on for all experts over all classes.
The parametric form for this distribution is assumed to be Gaussian [10], however, it may be very skewed towards 1 for correct distributions and towards 0 for wrong distributions depending on the experts and the data used to estimate the distributions.
A logistic distribution would fit properly in this situation, with a peak at 1 for correct cases and a peak at 0 for wrong cases.
A full Bayesian approach can also be used in this technique by integrating over the parameter space.

Comments on the Proposed Methods

The first approach main strength is that it doesn't care about the expert output value (whether it is a probability or not) and only cares about the expert opinion (decision as a discrete value of *J* classes). This is probably also the weakness here, since it terminates the relation with the expert input space and it doesn't take into account how ***"LOUD"*** the expert is stating its decision (i.e. for a 2-class case, if the expert's output is 0.51 or 0.99 its decision is class A).
The second approach takes into account the actual value of the expert's output. This may be optimum for probability-based experts, but not for a human expert who would just give a discrete opinion, unless some fuzzy measure of confidence is introduced.
Also, it needs more research to find the best way to model the expert's outputs distributions (which seem to me that a logistic distribution may be the best).

4. Experimental Results

Experiment 1

This is an artificial 2-dimensional, 2-classe problem, where each class is made of 2 clusters as shown in figure 1 which shows the test data. We have 4 experts, where each expert sees only one cluster from each class as shown in figure 2. In other words, each expert is trained using only one cluster from each class. This resembles a case when each expert is educated using a different part of a domain, then in testing each expert is asked to give an opinion about its input which may come from any part of the domain. The assumption is that the experts can't be re-trained nor adapted using new data. The DM can only use validation data to estimate the probabilities used in calculating its belief as shown earlier.

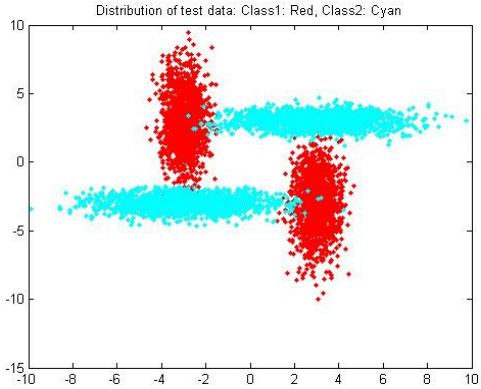

Figure 1. Data distribution of Experiment 1

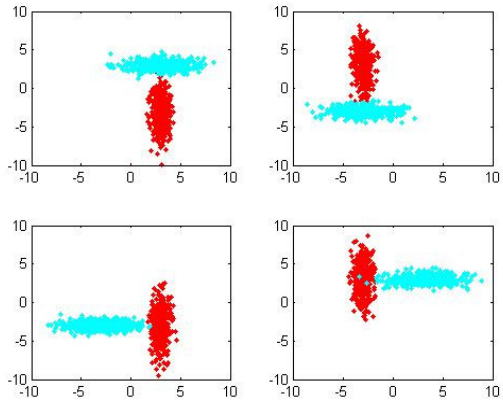

Figure 2. Training data for each expert

Each expert is a MLP with one hidden layer with 10 neurons. Each expert is trained using 800 patterns (400 from each cluster).
Test data contains 6000 patterns (1500 from each cluster), and 6000 patterns are used for validation data (containing data from all 4 clusters equally). The maximum achievable performance using a single classifier in this problem is about 98.5%.
A comparison is made between the supra-Bayesian combination (SBCC) with an averaging technique where the experts' performances are averaged, over 300 runs with the same data distribution (different realizations), and the results are compared in form of histograms.

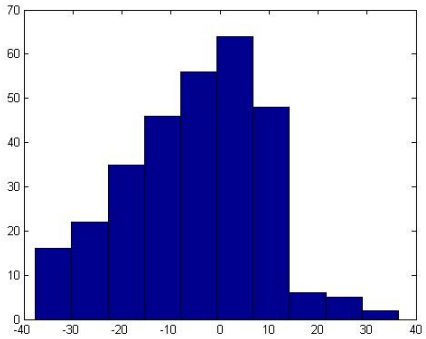

Figure 3. Histogram of %performance difference between SBCC and averaging technique

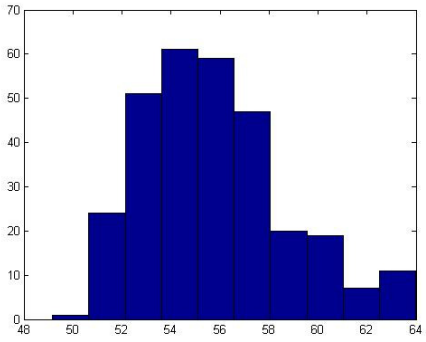

Figure 4. Histogram of averaging %performance

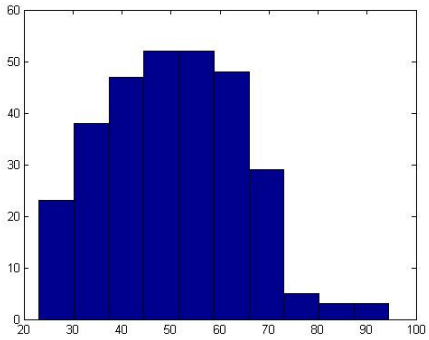

Figure 5. Histogram of SBCC %performance

From figures (3-5), the SBCC did not, in average, perform better than the averaging technique, however, it achieved higher performances. The reason is that in this case, experts' probabilities are calculated over the whole data set domain. Since each expert knows only a part of the domain, their performances in other parts are poor, and hence the DM's overall belief in experts' opinions are not good.

To overcome this problem input space clustering is used, and experts' probabilities are calculated within each cluster.

Extending the DM Knowledge

If we cluster the input space into L clusters, and embed this information into the DM knowledge ρ, then we can have the DM estimates for the values $P(Pk = Ci \mid Cj, \rho)$ to be conditioned on which cluster the input is from. The clustering technique and the identity of the cluster for each data entry is done by a DM-operated clustering engine, which is a standard K-means algorithm.

In this case, the DM has a separate estimate $P(Pk = Ci \mid Cj, cl)$ for each cluster Cl.

This probability is calculated as number of times expert k outcomes class Ci when class Cj is true when the input is from cluster l.

The idea behind this approach is that an expert may be very good in predicting a certain part of the input space, and the DM should take what the expert says about this region very seriously. However, for other parts of the input space, the expert may have very poor predictions, and the DM should give very low weight to the expert's opinion in that region.

In the experiment we have, the data has four clusters, thus we assume knowing the number of clusters and used a K-means algorithm. The resulting clustering is shown in figure 6.

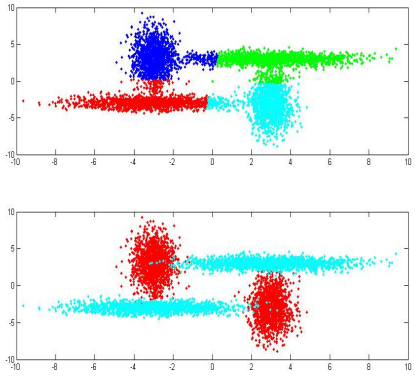

Figure 6. Top: K-means clustering, Down: The classes distributions

All probability estimates are re-calculated based on the cluster identity. The results of this technique, SBCC-clustering, are shown in comparison with the averaging technique in figure 7.

And the histogram of %performance difference between the SBCC-clustering and the averaging is shown in figure 8.

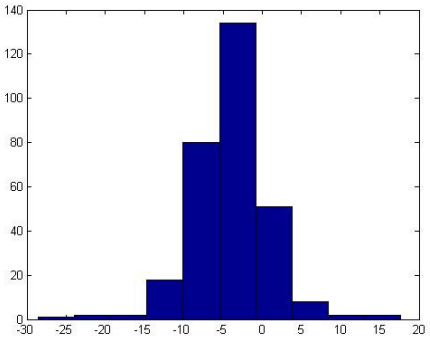

Figure 7. histogram of %performance difference between SBCC-clustering and averaging technique

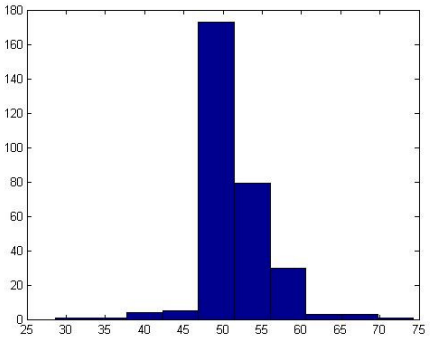

Figure 8. histogram of %performance difference between SBCC-clustering and averaging technique

The clustering has divided the space into 4 parts, where now, each expert has very good knowledge about 2 clusters. And since the probability estimates are now based on the cluster from which the data is coming from, thus, these estimates are either high or low, and that makes the DM belief more accurate. The performance of the SBCC-clustering technique is always over 94%. The effect of the amount of validation data used to estimate the experts' probabilities is shown for the SBCC and the SBCC-clustering in figures (9,10). These are averaged over 20 runs. It is clear that the SBCC-clustering suffers more since the probabilities are conditioned on the cluster and class, thus less data is used in their calculations.

5. Conclusions

The Bayesian philosophy embodies the concept of "averaging" explanations as in the Bayesian model averaging approach. However, averaged models need to be operating in the same knowledge space. Here a novel data averaging methodology is proposed, which overcomes this problem, however, experiments have to be conducted to see its effectiveness. Also, BMA needs to have models described as probabilities, which may not be suited for many classifiers.

The supra-Bayesian approach offers a way of aggregating experts' knowledge even if the experts have different knowledge bases and no probabilistic models. For experiment 1 scenario, when each expert sees only a part of the input domain, clustering may be used so that the DM can separate the confidence of each expert between different parts of the space. This novel technique is called SBCC-clustering, and has shown excellent performance in both experiments. However, it requires more validation data to have sufficient knowledge within each cluster. In other words, the DM should be able to give a higher weight for an expert in certain situations while a lower weight in others within the same class.

6. References

1- Kittler, J.V.[Joseph V.], Combining Classifiers: A Theoretical Framework, PAA(1), No. 1, 1998, pp. 18-27.

2- Kuncheva, L.I.[Ludmila I.], A Theoretical Study on Six Classifier Fusion Strategies, PAMI(24), No. 2, February 2002, pp. 281-286.

3- Ghahramani, Z. and Kim, H.-C. (2003) Bayesian Classifier Combination, Gatsby Technical Report.

4- Didaci, L.[Luca], Giacinto, G.[Giorgio], Roli, F.[Fabio], Marcialis, G.L.[Gian Luca], A study on the performances of dynamic classifier selection based on local accuracy estimation, PR(38), No. 11, November 2005, pp. 2188-2191.

5- Murua, A.[Alejandro], Upper Bounds for Error Rates of Linear Combinations of Classifiers, PAMI(24), No. 5, May 2002, pp. 591-602.

6- V. Petridis, A. Kehagias, L. Petrou, A. Bakirtzis, N. Maslaris, S. Kiartzis; A Bayesian Multiple Models Combination Method for Time Series Prediction. Journal of Intelligent and Robotic Systems ,Volume 31 , Issue 1-3 May -July 2001, Pages: 69 – 89, 2001

7- Ayad, H. and Kamel, M. "Refined Shared Nearest Neighbors Graph for Combining Multiple Data Clusterings", Advances in Intelligent Data Analysis. The 5th International Symposium on Intelligent Data Analysis, IDA 2003, Berlin, Germany, Proceedings. LNCS, Springer. pp. 307-318, 2003.

8- A. Dawid, A. Skene; “Maximum Likelihood Estimation of Observer Error-Rates Using The EM Algorithm”, Applied Statitics, 28, 20-29, 1979.

9- Lindley D.V.; “The Improvement of Probability Judgements”, Journal of the Royal Statistical Society (A), 145. 1982.

10- R. Jacobs; “Methods for Combining Experts’ Probability Assessments”, Neural Computations (7), 1995.

11- Dennis Bahler, Laura Navarro; “Combining Heterogeneous Sets of Classifiers: Theoretical and Experimental Comparison of Methods”,

12- Bayesian Model Averaging (review paper) PDF Version ; Jennifer Hoeting, David Madigan, Adrian Raftery and Chris Volinsky (1999) *Statistical Science* 14, 382-401.