Towards a Supra-Bayesian Approach to Merging of Information

Vladimíra Sečkárová*
Department of Adaptive Systems
Institute of Information Theory and Automation
Prague, CZ 182 08
seckarov@utia.cas.cz

Abstract

Merging of information given by different decision makers (DMs) has become a much discussed topic in recent years and many procedures were developed towards it. The main and the most discussed problem is the incompleteness of given information. Little attention is paid to the possible forms in which the DMs provide them; in most of cases arising procedures are working only for a particular type of information. Recently introduced Supra-Bayesian approach to merging of information brings a solution to two previously mentioned problems. All is based on a simple idea of unifying all given information into one form and treating the possible incompleteness. In this article, beside a brief repetition of the method, we show, that the constructed merger of information reduces to the Bayesian solution if information calls for this.

1 Introduction

In this article we bring the answer to a consistency question regarding the final result of information merging method based on Supra-Bayesian approach (introduced in [7]).

Method itself deals with problem of incomplete and incompatible (having different forms) data from sources – decision makers. People are trying to solve the incompleteness by developing various methods, bases of which are, e.g. semantics, entities and trust [1], reduction of the combination space by representing the notion of source redundancy or source complementarity [2] or Bayesian networks and factor graphs [3]. Altogether they often lack one thing – they are usable only if the information has unified form. The Supra-Bayesian merging solves previously mentioned problems in three steps. First, we focus on the incompatibility of forms of input data and transform them into a probabilistic form. Second, we fill in the missing information (in the paper it is called extension) to resolve the problem of incompleteness. After that we will construct the merger of already transformed and extended data. Articles related to the proposed topic can be found in [4] and [5].

Section 2 briefly describes the construction of the merger. Section 3 presents an important check of the solution's logical consistency: the final merger reduces to the standard Bayesian learning when the processed data meets standard conditions leading to it. Throughout the text a discrete case is considered.

2 Basic terms and notation

In the beginning of this section we introduce the basic terms and notation used through the text, then we give the main steps of the method.

*Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

The basic terms we use are as follows:

- a source – a decision maker (e.g. human being), which gave us the information,
  – we now pick one source, denoted by S; the explained setup can be, of course, applied to other sources as well,
- a domain (of the source) – a state space, about which the source provides the information,
  – since a domain can be difficult to describe, we use (discrete) random variable to map it onto preferable space; the range of this mapping consists of finite number of elements
  – every source can describe more than one domain; the relation between them and their ranges maps random vector
- a neighbour of the source S – another source, which has at least one domain (of its considered random variables) same as S
  (note that the range of these variables can differ, so the arising probability measure can be different for each source)
  – we assume that the number of neighbours is finite (for each considered source S); they are labeled by $j = 1, \ldots, s-1 < \infty$,
  – altogether, we have a set consisting of s sources (source S and its $s-1$ neighbours),
  – we denote random vector of j^{th} source by $\mathbf{Y}_j$, set of its realizations by $\{\mathbf{y}_j\}$.

2.1 Transformation of information into probabilistic type

I. Consider that the j^{th} source expressed the information about its domain as a realization of its random vector $\mathbf{Y}_j$ denoted by $\mathbf{x}_j$.

The transformation to probability mass function (pmf) $g_{\mathbf{Y}_j}$ will be done via Kronecker delta as follows:

$$g_{\mathbf{Y}_j}(\mathbf{y}_j) = \delta(\mathbf{x}_j - \mathbf{y}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_j = \text{ a particular realization } \mathbf{y}_j \text{ of } \mathbf{Y}_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

II. Let j^{th} source give us the conditional expectation of the function of $\mathbf{Y}_j$. We would like to determine the pmf $g_{\mathbf{Y}_j}(\mathbf{y}_j) = g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)$, where $\mathbf{P}_j$ denotes a part of $\mathbf{Y}_j$, which is specified by source's past experience with realizations $\{\mathbf{p}_j\}$ and $\mathbf{F}_j$ denotes a part of $\mathbf{Y}_j$ expressing source's uncertainty (ignorance) with realizations $\{\mathbf{f}_j\}$. We will use the maximum entropy principle (see [8], [9]): we construct a set of all possible pmfs describing $\mathbf{Y}_j$ and satisfying the expectations, then we choose the pmf with the highest entropy.

III. Let j^{th} source give us the expectation of the function of $\mathbf{Y}_j$. Similarly as in the previous case we will use the maximum entropy principle to determine $g_{\mathbf{Y}_j}(\mathbf{y}_j) = g_{\mathbf{P}_j}(\mathbf{p}_j)$.

IV. and V. Let the source give a pmf of $\mathbf{Y}_j$ denoted by $g_{\mathbf{P}_j}(\mathbf{p}_j)$ or conditional pmf of a part of $\mathbf{Y}_j$ conditioned by the remaining part denoted by $g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)$. These types of information already are in the targeted probabilistic form.

2.2 Extension

Our first step in constructing the extensions is a unification of the ranges of considered sources $j = 1, \ldots, s < \infty$, which means construction of random vector $\mathbf{Y}$ involving all different random variables considered by sources. A set of realizations of $\mathbf{Y}$ will be denoted by $\mathcal{Y} = \{\mathbf{y}\}$ and their number will be finite (since we assumed range of each source has finite number of realizations).

The decomposition of $\mathbf{Y}$ according to j^{th} source then arises naturally:

- if j^{th} source has its random vector decomposed into two parts $\mathbf{Y}_j = (\mathbf{F}_j, \mathbf{P}_j)$ (as introduced in previous section), the decomposition of $\mathbf{Y}$ will be: $\mathbf{Y} = (\mathbf{U}_j, \mathbf{F}_j, \mathbf{P}_j)$, where $\mathbf{U}_j$ (with realizations $\{\mathbf{u}_j\}$) stands for the remaining realizations in $\mathbf{Y}$ unconsidered by j^{th} source;
- if for j^{th} source holds that $\mathbf{Y}_j = \mathbf{P}_j$, then the decomposition of $\mathbf{Y}$ will be: $\mathbf{Y} = (\mathbf{U}_j, \mathbf{P}_j)$, where again the part $\mathbf{U}_j$ denotes the remaining random variables in $\mathbf{Y}$ unconsidered by the source.

The yet unconstructed merger $\widetilde{h}$ serves us for the extension of pmfs $g_{\mathbf{P}_j}$ and $g_{\mathbf{F}_j|\mathbf{P}_j}$ to $g^{(j)}_{\mathbf{Y}}$ describing the union of neighbours' ranges. If the conditional pmf $g_{\mathbf{F}_j|\mathbf{P}_j}$ is available, then the extension is: $g^{(j)}_{\mathbf{Y}}(\mathbf{y}) = \widetilde{h}(\mathbf{u}_j|\mathbf{f}_j, \mathbf{p}_j) g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)\widetilde{h}(\mathbf{p}_j)$, where $\widetilde{h}(\mathbf{p}_j)$, $\widetilde{h}(\mathbf{u}_j|\mathbf{f}_j, \mathbf{p}_j)$ and $\widetilde{h}(\mathbf{u}_j|\mathbf{p}_j)$ are marginal and conditional versions of $\widetilde{h}$. We proceed similarly if the marginal pmf $g_{\mathbf{P}_j}$ is available.

2.3 Final merger

After successfully dealing with the transformation and extension of given information we can derive the merger. According to the Bayesian framework [10] our merger will be following pmf:

$$\widetilde{h} = \underset{\hat{h}\in\widehat{H}}{\arg\min}\, \mathrm{E}_{\pi(h|D)}[\mathrm{L}(h, \hat{h})|D],$$

where: $\widehat{H}$ denotes a set of all possible estimates $\hat{h}$ of h, D stands for a matrix consisting of extended probability vectors $g^{(j)}_{\mathbf{Y}}$, $\pi(h|D)$ is the posterior pdf of h based on D, $\mathrm{L}(.\,,.)$ is a loss function.

Since h and $\hat{h}$ are pmfs, the loss function should measure the distance between them. In particular, we choose the Kerridge inaccuracy $\mathrm{K}(.\,,.)$ (see [11]). We then get the following identity (after using Fubini's theorem and a little bit of computation; H is a probabilistic simplex containing h-values):

$$\underset{\hat{h}\in\widehat{H}}{\arg\min}\, \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h, \hat{h})|D\right] = \ldots = \underset{\hat{h}\in\widehat{H}}{\arg\min}\, \mathrm{K}\left(\mathrm{E}_{\pi(h|D)}(h|D), \hat{h}\right).$$

Kerridge inaccuracy reaches the minimal value if its arguments are equal almost everywhere (a.e.) (see [11]). Then the following equation holds:

$$\widetilde{h} = \underset{\hat{h}\in\widehat{H}}{\arg\min}\, \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h, \hat{h})|D\right] = \mathrm{E}_{\pi(h|D)}(h|D).$$

The only problem is we do not have the posterior pdf $\pi(h|D)$ of h, so before we actually get to the formula expressing the final merger $\widetilde{h}$ (final estimate of h) we have to choose $\pi(h|D)$. Again we will use maximum entropy principle. This time we are looking for the element with highest entropy subject to additional constraints. The constraints will be connected with the opinion of source S about the distance of j^{th} source from the unknown pmf h using Kerridge inaccuracy (for all $j = 1, \ldots, s$). They are expressed by

$$\mathrm{E}_{\pi(h|D)}\left(\mathrm{K}(g^{(j)}_{\mathbf{Y}}, h)|D\right) \leq \beta_j(D). \tag{2}$$

Thus, to obtain the optimal $\widetilde{\pi}(h|D)$ we have to solve following optimization task:

$$\widetilde{\pi}(h|D) = \underset{\pi(h|D)\in\mathrm{M}}{\arg\min}\left[\int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h\right], \tag{3}$$

where $\mathrm{M} = \left\{\pi(h|D) :\, \mathrm{E}_{\pi(h|D)}(\mathrm{K}(g^{(j)}_{\mathbf{Y}}, h)|D) - \beta_j(D) \leq 0,\ j = 1, \ldots, s,\ \int_H \pi(h|D)\mathrm{d}h - 1 = 0\right\}$.

By constructing and rearranging the Lagrangian $\mathrm{L}(.\,,.)$ of the task (3) we get that its minimum is reached for pdf of Dirichlet distribution $Dir(\{\nu_{\mathbf{y}}\}_{\mathbf{y}\in\mathcal{Y}})$:

$$\widetilde{\pi}(h|D) = \frac{1}{Z(\boldsymbol{\lambda}(D))}\prod_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{y})^{\nu_{\mathbf{y}}-1} \quad \text{with parameters } \nu_{\mathbf{y}} = 1 + \sum_{j=1}^{s}\lambda_j(D) g^{(j)}_{\mathbf{Y}}(\mathbf{y}), \quad \forall\, \mathbf{y}\in\mathcal{Y}.$$

Once we have computed the posterior pdf, we can go back to the expressing the final merger (the optimal estimate $\widetilde{h}$ of h). Denote the number of realizations of $\mathbf{Y}$ by $n\ (<\infty)$ and use the properties of Dirichlet distribution, particularly

$$\mathrm{E}_{\widetilde{\pi}(h|D)}[h(\mathbf{y})|D] = \frac{\nu_{\mathbf{y}}}{\nu_0}, \quad \text{where } \nu_0 = \sum_{\mathbf{y}\in\mathcal{Y}}\nu_{\mathbf{y}} = \sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s}\lambda_j(D)\overbrace{\sum_{\mathbf{y}\in\mathcal{Y}} g^{(j)}_{\mathbf{Y}}(\mathbf{y})}^{=1}.$$

We get following result:

$$\widetilde{h}(\mathbf{y}) = \frac{1 + \sum_{j=1}^{s}\lambda_j(D) g^{(j)}_{\mathbf{Y}}(\mathbf{y})}{n + \sum_{j=1}^{s}\lambda_j(D)}. \tag{4}$$

3 Connection to the Bayesian solution

As promised earlier (see Section 1) we will now check if the final merger (4) reduces to a standard Bayesian learning if merging scenario meets conditions leading to it. First we will derive the empirical pmf via Bayesian approach, second we will reformulate the problem so that our merger can be applied, compute the empirical pmf and compare the results.

3.1 A Bayesian view

Let

- Y be a discrete random variable with finite number of realizations $\{y\} = \mathcal{Y}$,
- $\boldsymbol{\theta}$ be a following random vector: $\boldsymbol{\theta} = (\mathrm{P}(Y = y))_{y\in\mathcal{Y}} = (\theta_y)_{y\in\mathcal{Y}}$. Then let $X_1, \ldots, X_s$, $(s < \infty)$, denote the sequence of observations about Y, which will be considered as independent random variables with the same distribution as Y (depending on $\boldsymbol{\theta}$).

If we assume that

- the prior distribution of $\boldsymbol{\theta} = (\theta_y)_{y\in\mathcal{Y}}$ is Dirichlet distribution $Dir(\{\alpha_y\}_{y\in\mathcal{Y}})$, meaning $q(\boldsymbol{\theta}) \propto \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y - 1}$,
- the conditional probability of X_j, $j = 1, \ldots, s$, conditioned by $\boldsymbol{\theta}$ is $f_{X_j}(x_j|\boldsymbol{\theta}) = \prod_{y\in\mathcal{Y}} \theta_y^{\delta(x_j - y)}$, where $\delta(.)$ stands for Kronecker delta (see (1)),

the posterior pmf of $\boldsymbol{\theta}$ based on $X_1, \ldots, X_s$ is then

$$\begin{aligned}\pi(\boldsymbol{\theta}|X_1 = x_1, \ldots, X_s = x_s) &\propto q(\boldsymbol{\theta}) \prod_{j=1}^{s} f_{X_j}(x_j|\boldsymbol{\theta}) \\ &= \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y - 1} \prod_{j=1}^{s} \prod_{y\in\mathcal{Y}} \theta_y^{\delta(x_j - y)} = \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y) - 1} \end{aligned} \tag{5}$$

Since the formula (5) is the pdf of Dirichlet distribution $Dir\left(\left\{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)\right\}_{y\in\mathcal{Y}}\right)$, we can easily compute the conditional expectation of θ_y conditioned by $X_1, \ldots, X_s$ as follows:

$$\begin{aligned}\mathrm{E}_{\pi(\boldsymbol{\theta}|X_1,\ldots,X_s)}(\theta_y|X_1 = x_1, \ldots, X_s = x_s) = \widetilde{P}(Y = y) &= \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \left[\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)\right]} \\ &= \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \alpha_y + s} \end{aligned} \tag{6}$$

Under the following choice:

$$\alpha_y = 1 \quad \forall\, y \in \mathcal{Y} \tag{7}$$

formula (6) will look as follows

$$\widetilde{P}(Y = y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} 1 + s}. \tag{8}$$

If n denotes the number of realizations of Y, then: $\widetilde{P}(Y = y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{n + s}$.

Note: The first part of (8) – $\frac{1}{\sum_{y\in\mathcal{Y}} 1+s}$ – can be considered as the prior pmf of Y, because if there is no available information, we will get: $\widetilde{P}_0(Y = y) = \frac{1}{\sum_{y\in\mathcal{Y}} 1+s}$. Then, the choice (7) coincides with the statement, that the prior pmf for Y is a pmf of Uniform distribution.

Illustrative example:

Assume we are interested in the changes of the stock price. Y will now be an identity mapping from the set consisting of 3 elements: 1 (increase), 0 (stagnation), -1 (decrease). We now get a sequence of data - opinions from independent experts: $\{1, -1, 1, 1, 0, 1, 1, -1, 1, 1\}$. Then the estimate of the probabilities will be (regarding the (7)):

$$\widehat{P}(Y=1) = \tfrac{1+7}{3+10} = \tfrac{8}{13}, \quad \widehat{P}(Y=0) = \tfrac{1+1}{3+10} = \tfrac{2}{12}, \quad \widehat{P}(Y=-1) = \tfrac{1+2}{3+10} = \tfrac{3}{13}.$$

3.2 Merging approach

Now we reformulate and handle the same information scenario as in Subsection 3.1 by using the proposed information merging.

Let us have a group of s (independent) sources, all of them describing the same domain and range. Therefore sources are neighbours and so the merging can be applied on them. The relation between domain and range maps discrete random variable Y, realizations of which are denoted by $\{y\} = \mathcal{Y}$.

Assume also that the information they gave are the values of Y, denoted by $x_1, \ldots, x_s$. Now we can follow the steps introduced in the previous sections:

1. transformation: (non-probability form into probability form)

– x_j will be expressed (in the probability form) as follows: $g_{Y_j=Y}(y_j = y) = \delta(x_j - y)$,

2. extension: (from particular domains to the union of all considered domains)

– since the sources have the same domain, Y, the union is also Y,

– because of that, the extended version of probabilistic form of given information will be:

$g_Y^{(j)}(y) = \delta(x_j - y)$,

3. merging: now that we have probabilistic information extended on Y, we can use the merger (4):

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) g_Y^{(j)}(y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)} = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)},$$

which for particular choice $\lambda_1(D) = \ldots = \lambda_j(D) = \ldots = \lambda_s(D) = 1$ has following form:

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + s}, \tag{9}$$

which coincides with (8). That is if we assume that the sources have the same reliability factor (see subsection 2.3) and it is equal to 1, the final merger (4) will reduce to the standard Bayesian learning considered in Subsection 3.1.

Illustrative example:

Our results can be easily applied on the example in the previous section: we have 9 independent sources, which have the same domain. Therefore they are neighbors. All given information are just values, we have to transform them into probabilities (see Section 2.1). Since they also have the same range no extension is needed, so we can directly proceed to the merging. According to (9) results are the same as in the example in Subsection 3.1.

Remark

In the note after the final merger (8) we brought the explanation of what should its first part represent – generally, it stands for the prior pmf for considered random vector $\mathbf{Y}$ (see (4)). In this paper, the prior pmf of $\mathbf{Y}$ is a pmf of uniform distribution. But we will be allowed to use another prior distribution if we choose constrained minimum cross entropy principle (see [9]) for determination of the posterior pdf (see subsection 2.3) instead of constrained maximum entropy principle. It is because the maximum entropy principle coincides with minimum cross entropy principle when prior distribution is uniform.

4 Conclusion

This paper brings an important conclusion regarding a new method for merging of information, which successfully deals with the different types of given partially overlapping information and also with problem of missing data. Since the method is based on Bayesian framework, we showed that it reduces to a standard Bayesian learning if independent identically distributed data are at disposal for parameter estimation. Still there are some open problems and topics of the future work, e.g. the choice of constraints $\beta_j(D)$ in (2), choice of prior distribution (see previous remark) and the extension to the continuous space.

Acknowledgement

This research has been partially supported by GAČR 102/08/0567 and DAR 1M0572.

References

[1] L. Šubelj, D. Jelenc, E. Zupančič, D. Lavbič, D. Trček, M. Krisper, and M. Bajec. Merging data sources based on semantics, contexts and trust. *The IPSI BgD Transactions on Internet Research*, 7(1):18–30, 2011.

[2] B. Fassinut-Mombot and J.B. Choquel. A new probabilistic and entropy fusion approach for management of information sources. *Information Fusion*, 5(1):35–47, 2004.

[3] G. Pavlin, P. de Oude, M. Maris, J. Nunnik, and T. Hood. A multi-agent systems approach to distributed bayesian information fusion. *Information Fusion*, 11(3):267–282, 2010.

[4] M. Kárný, T. Guy, A. Bodini, and F. Ruggeri. Cooperation via sharing of probabilistic information. *International Journal of Computational Intelligence Studies*, pages 139–162, 2009.

[5] M. Kárný and T.V. Guy. Sharing of Knowledge and Preferences among Imperfect Bayesian Participants. In *Proceedings of the NIPS Workshop 'Decision Making with Multiple Imperfect Decision Makers'*. UTIA, 2010.

[6] C. Genest and J. V. Zidek. Combining probability distributions: a critique and an annotated bibliography. With comments, and a rejoinder by the authors. *Stat. Sci.*, 1(1):114–148, 1986.

[7] V. Sečkárová. Supra-Bayesian Approach to Merging of Incomplete and Incompatible Data. In *Decision Making with Multiple Imperfect Decision Makers Workshop at 24th Annual Conference on Neural Information Processing Systems*, 2010.

[8] E.T. Jaynes. Information theory and statistical mechanics. I, II. 1957.

[9] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory*, 26:26–37, 1980.

[10] M. H. DeGroot. *Optimal statistical decisions.* Wiley-Interscience; Wiley Classics Library. Hoboken, NJ: John Wiley and Sons. xx, 489 p., 1970.

[11] D.F. Kerridge. Inaccuracy and inference. *J. R. Stat. Soc., Ser. B*, 23:184–194, 1961.