Progress in Supra-Bayesian Merging of Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

Merging of data 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 data. Little attention is paid to the possible forms they can have; in most of cases arising procedures are working only for a particular type of information. In this article we introduce a merger that brings a solution to two previously mentioned problems by using Supra-Bayesian approach. All is based on a simple idea of unifying all given data into one form and treating the possible incompleteness. In the end it will be shown, that the constructed merger reduces to the Bayesian solution if data calls for this.

1 Introduction

An everyday life of a statistician consists of computing some statistics based on observations. Always hoping for getting fully-fledged data, there is a ghost of chance that this will happen. Therefore the computation, which is in fact a merging process, has been a much discussed topic in past few years. People are trying to eliminate this problem by developing different 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 is of one form. If we introduce a source as an object able to provide some information, we can get a wide range of forms that a given information can have. Sometimes the methods are well prepared from the descriptive point of view, but without giving a constructive tool any kind of computation can hardly be provided. This article brings the answer to previously mentioned problems. Recently, there have also been published articles related to the proposed topic, e.g. [4] and [5].

Precisely talking, we will use a Supra-Bayesian approach [6], where the task of merging the given pieces of information is expressed as a task of constructing a posterior probability mass function (pmf) or probability density function (pdf) for a fictitious decision maker by using Bayes' theorem. In this paper we will focus on the discrete case – built merger will be a posterior pmf. This may lead to a misguiding conclusion, that given data should have a form of pmf. Such a restriction can be easily overcome, since there exist many tools for transforming the forms of input data considered in this paper, namely, (conditional) expected values and realizations of random vectors, into pmfs. In case when (conditional) pmfs of a random vector are given, no transformation step is needed. Once we have treated the incompatibility of the raw information pieces, we face the problem of incompleteness – missing information. This will be solved by inserting the appropriate versions of a not yet constructed merger into the places, where the information is missing. We will end up with complete information in the form of pmfs, so the previously mentioned Supra-Bayesian approach can be easily applied. Summarizing what have been said so far, our method for constructing such a merger consists of three steps. First we focus on the incompatibility of forms of input data and transform them into probabilistic form. Second we fill in the missing information (in the paper it is called extension) so the problem of incompleteness vanishes. After that we will construct the merger of already transformed and extended data.

The description of this construction forms the first part of the paper. It is complemented by 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.

2 Description of the method

In the beginning of this section we introduce the basic terms and notation used through the text and give a list of the main steps of the method. Then, each of the steps will be explained and formalized.

The basic terms we use throughout the text are:

- a source – an object (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) – the environment, about which the source provides the information,

  – we represent it by a (discrete) random vector with a finite number of realizations,
- a neighbour of the source S – another source, domain of which has a nonempty intersection with the domain of source S,

  – we assume that the number of neighbours is finite (for each considered source S); they are labeled by $j = 1, \dots, s-1 < \infty$,

  – altogether, we have a set consisting of s sources (source S and its $s-1$ neighbours),

  – we denote the domain of j^{th} source by $\mathbf{Y}_j$ and assume it has a finite number of different instances $\{\mathbf{y}_j\}$,

  – we treat $\mathbf{Y}_j$ as a discrete random vector, $j = 1, \dots, s$.

Our method consists of the following main steps:

1. expression of different types of given information as probabilistic information (i.e. as probability mass function (pmf) of each domain $\mathbf{Y}_j$, $j = 1, \dots, s$),
2. extension of the pmfs to the union of all domains considered by s sources (using constraints given on the distance between two pmfs),
3. construction of the final merger.

2.1 Transformation of information into probabilistic type

To repeat the important things said so far, we allowed sources to provide information in different forms and we interpreted the domain of particular source as a discrete random vector. Thus we consider following types of information offered by j^{th} source:

I. values of $\mathbf{Y}_j$,

II. an expectation of a function of $\mathbf{Y}_j$,

III. a conditional expectation of a function depending on a part of $\mathbf{Y}_j$ conditioned by the remaining part,

IV. a pmf of $\mathbf{Y}_j$,

V. a conditional pmf of a part of $\mathbf{Y}_j$ conditioned by the remaining part.

Let us discuss constructed transformation in respective cases.

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 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 expectation $\mathrm{E}_{g_{\mathbf{Y}_j}}$ of the function ϕ of $\mathbf{Y}_j$, which can be denoted by

$$\mathrm{E}_{g_{\mathbf{Y}_j}}(\phi(\mathbf{Y}_j)) = \psi, \tag{2}$$

where ϕ and ψ are a function and a value specified by the source. The expectation is taken with respect to a, yet unspecified, pmf $g_{\mathbf{Y}_j}$, existence of which is assumed. We will use the maximum entropy principle (see [7], [8]) to determine the pmf $g_{\mathbf{Y}_j}$. Specifically, we take a set $\{g^*_{\mathbf{Y}_j}\}$ of all possible pmfs describing $\mathbf{Y}_j$ and satisfying (2) and take the one with the highest entropy:

$$g_{\mathbf{Y}_j} = \underset{g^*_{\mathbf{Y}_j}}{\arg\max} \left[-\sum_{\mathbf{y}_j} g^*_{\mathbf{Y}_j}(\mathbf{y}_j) \log g^*_{\mathbf{Y}_j}(\mathbf{y}_j) \right]. \tag{3}$$

III. Let j^{th} source give us the conditional expectation of the function of $\mathbf{Y}_j$. If we decompose source's domain into 2 parts:

A) $\mathbf{P}_j$ denotes a part of $\mathbf{Y}_j$, which is specified by source's past experience with realizations $\{\mathbf{p}_j\}$,

B) $\mathbf{F}_j$ denotes a part of $\mathbf{Y}_j$ expressing source's uncertainty (ignorance) with realizations $\{\mathbf{f}_j\}$,

so that $(\mathbf{P}_j \cup \mathbf{F}_j) = \mathbf{Y}_j$ and $(\mathbf{P}_j \cap \mathbf{F}_j) = \emptyset$, we can express the given information as:

$$\mathrm{E}_{g_{\mathbf{F}_j|\mathbf{P}_j}}(\phi(\mathbf{F}_j, \mathbf{P}_j)|\mathbf{P}_j) = \psi(\mathbf{P}_j), \tag{4}$$

where ϕ, ψ are functions specified by the source. The expectation is taken with respect to a, yet unspecified, pmf $g_{\mathbf{F}_j|\mathbf{P}_j}$, existence of which is assumed. Similarly as in the previous case we will use the maximum entropy principle and construct a set $\{g^*_{\mathbf{F}_j|\mathbf{P}_j}\}$ of all possible pmfs describing $\mathbf{Y}_j$ and satisfying (4). Then we choose the pmf with the highest entropy:

$$g_{\mathbf{F}_j|\mathbf{P}_j} = \underset{g^*_{\mathbf{F}_j|\mathbf{P}_j}}{\arg\max} \left[-\sum_{\mathbf{f}_j} g^*_{\mathbf{F}_j|\mathbf{P}_j}(\mathbf{f}_j|\mathbf{p}_j) \log g^*_{\mathbf{F}_j|\mathbf{P}_j}(\mathbf{f}_j|\mathbf{p}_j) \right]. \tag{5}$$

IV. and V. If the probability vector describing $\mathbf{Y}_j$ (denoted by $g_{\mathbf{Y}_j}$) or conditional probability vector describing conditional relation between $\mathbf{F}_j$ and $\mathbf{P}_j$ (denoted by $g_{\mathbf{F}_j|\mathbf{P}_j}$) is given, there is no need to transform them because they already are in the targeted probabilistic form.

2.2 Extension

We eliminated the problem of different data forms and we have all of given information pieces in the probabilistic form. Thus, we can focus on their extension. It is reasonable to do so, since the selected sources are generally operating on slightly different domains (they are still neighbours) and merging of their information would be difficult without such extension.

Our first step in constructing the extensions is a unification of the domains of considered sources $j = 1, \ldots, s < \infty$, which means construction of random vector $\mathbf{Y}$ involving all different random variables from sources' domains. A set of realizations of $\mathbf{Y}$ will be denoted by $\mathcal{Y} = \{\mathbf{y}\}$ and their number will be finite (since we assumed domain of each source has finite number of realizations). Note that in case of I. and II. type of given information we can use $\mathbf{P}_j$ instead of $\mathbf{Y}_j$, since there is no part expressing uncertainty, only source's past is known.

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

- if j^{th} source has its domain 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 random variables in $\mathbf{Y}$ unconsidered by j^{th} source;
- if the domain of j^{th} source is $\mathbf{Y}_j = \mathbf{P}_j$, 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' domains. The following formula formalizes this idea:

- if the marginal pmf $g_{\mathbf{P}_j}$ is available, then the extension is:

$$g^{(j)}_{\mathbf{Y}}(\mathbf{y}) = \widetilde{h}(\mathbf{u}_j|\mathbf{p}_j)g_{\mathbf{P}_j}(\mathbf{p}_j) \tag{6}$$

- if the conditional pmf $g_{\mathbf{F}_j|\mathbf{P}_j}$ is available, then the extension is:

$$g^{(j)}_{\mathbf{Y}}(\mathbf{y}) = \widetilde{h}(\mathbf{u}_j|\mathbf{f}_j,\mathbf{p}_j)g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)\widetilde{h}(\mathbf{p}_j) \tag{7}$$

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 pmfs of $\widetilde{h}$.

2.3 Final merger

After successfully dealing with the transformation and extension of given information we can derive the merger. First we clarify the form of our merger. It will be expressed as a probability vector, since we use Supra-Bayesian approach, where, as said in the Introduction, the task of merging is expressed as a task of constructing a posterior pmf for a fictitious decision maker by using Bayes' theorem. Specifically, it is an estimate $\widetilde{h}$ of the probability vector h describing $\mathbf{Y}$ based on given information (i.e. the extended pmfs). According to the Bayesian framework [9] our merger will be:

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

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 (selected later),
- $\mathrm{L}(.\,,.)$ 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 [10]) as the loss function. We will then get the following identity (after using Fubini's theorem and a little bit of computation; H is a probabilistic simplex containing h-values):

$$\begin{aligned}
\arg\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h,\hat{h})|D\right] &= \arg\min_{\hat{h}\in\widehat{H}} \int_H \pi(h|D)\mathrm{K}(h,\hat{h})\mathrm{d}h \\
&= \arg\min_{\hat{h}\in\widehat{H}} \int_H \pi(h|D) \sum_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{y})\log\hat{h}(\mathbf{y})\mathrm{d}h \\
&= \arg\min_{\hat{h}\in\widehat{H}} \sum_{\mathbf{y}\in\mathcal{Y}} \left(\int_H \pi(h(\mathbf{y})|D)h(\mathbf{y})\mathrm{d}h\right)\log\hat{h}(\mathbf{y}) \qquad (8) \\
&= \arg\min_{\hat{h}\in\widehat{H}} \sum_{\mathbf{y}\in\mathcal{Y}} \left(\mathrm{E}_{\pi(h(\mathbf{y})|D)}(h(\mathbf{y})|D)\right)\log\hat{h}(\mathbf{y}) \\
&= \arg\min_{\hat{h}\in\widehat{H}} \mathrm{K}\left(\mathrm{E}_{\pi(h|D)}(h|D),\hat{h}\right).
\end{aligned}$$

Since the Kerridge inaccuracy reaches the minimal value if its arguments are equal almost everywhere (a.e.) (see [10]), we see that:

$$\widetilde{h} = \arg\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h,\hat{h})|D\right] = \mathrm{E}_{\pi(h|D)}(h|D). \tag{9}$$

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, which we have already used for transformation of provided information (see Subsection 2.1). But this time it will be a little different – 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

$$\mathrm{E}_{\pi(h|D)}\left(\mathrm{K}(g_{\mathbf{Y}}^{(j)}, h)|D\right) \leq \beta_j(D). \tag{10}$$

Thus, for constructing $\pi(h|D)$ we have to solve following optimization task

$$\underset{\pi(h|D)\in\mathrm{M}}{\arg\max}\left[-\int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h\right],$$

which can be equivalently reformulated as follows:

$$\underset{\pi(h|D)\in\mathrm{M}}{\arg\min}\left[\int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h\right], \tag{11}$$

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

By rearranging the Lagrangian $\mathrm{L}(.\,,.)$ of the task (11) we get (the Lagrangian is given by multipliers $\boldsymbol{\lambda}(D) = (\lambda_1(D), \ldots, \lambda_s(D))$):

$$\begin{aligned}
&\mathrm{L}(\pi(h|D);\boldsymbol{\lambda}(D)) = \int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h + \sum_{j=1}^{s}\lambda_j(D)\left[\mathrm{E}_{\pi(h|D)}\left(\mathrm{K}(g_{\mathbf{Y}}^{(j)}, h)|D\right) - \beta_j(D)\right] \\
&= \int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h - \overbrace{\sum_{j=1}^{s}\int_H}^{\text{Fubini}} \lambda_j(D)\pi(h|D)\sum_{\mathbf{y}\in\mathcal{Y}} g_{\mathbf{Y}}^{(j)}(\mathbf{y})\log(h(\mathbf{y}))\mathrm{d}h \\
&- \sum_{j=1}^{s}\lambda_j(D)\beta_j(D) \\
&= \int_H \pi(h|D)\left(\log\pi(h|D) - \sum_{\mathbf{y}\in\mathcal{Y}}\log(h(\mathbf{y}))^{\sum_{j=1}^{s}\lambda_j(D)g_{\mathbf{Y}}^{(j)}(\mathbf{y})} \pm \log Z(\boldsymbol{\lambda}(D))\right)\mathrm{d}h \\
&- \sum_{j=1}^{s}\lambda_j(D)\beta_j(D) \\
&= \int_H \pi(h|D)\left(\log\pi(h|D) - \log\prod_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{y})^{\sum_{j=1}^{s}\lambda_j(D)g_{\mathbf{Y}}^{(j)}(\mathbf{y})} - \log\frac{1}{Z(\boldsymbol{\lambda}(D))}\right)\mathrm{d}h \\
&- \int_H \pi(h|D)\log Z(\boldsymbol{\lambda}(D))\mathrm{d}h - \sum_{j=1}^{s}\lambda_j(D)\beta_j(D) \\
&= \int_H \pi(h|D)\log\left(\frac{\pi(h|D)}{\frac{\prod_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{y})^{\sum_{j=1}^{s}\lambda_j(D)g_{\mathbf{Y}}^{(j)}(\mathbf{y})}}{Z(\boldsymbol{\lambda}(D))}}\right)\mathrm{d}h - \log Z(\boldsymbol{\lambda}(D))\underbrace{\int_H \pi(h|D)\mathrm{d}h}_{=1} \\
&- \sum_{j=1}^{s}\lambda_j(D)\beta_j(D) \\
&= \mathrm{D}(\pi(h|D)||\widetilde{\pi}(h|D)) - \log Z(\boldsymbol{\lambda}(D)) - \sum_{j=1}^{n}\lambda_j(D)\beta_j(D),
\end{aligned}$$

where $\mathrm{D}(.\,,.)$ stands for Kullback-Leibler divergence (see [11] and [12]), $Z(\boldsymbol{\lambda}(D))$ is a normalizing factor and $\lambda_j(D) \geq 0$ are Lagrange multipliers, $j = 1, \ldots, s$. j^{th} multiplier can be interpreted as a reliability factor of how "good" is the information given by j^{th} source.

We see that the minimum of this Lagrangian is reached for pdf $\pi(h|D) = \widetilde{\pi}(h|D)$ a.e., because:

- the first part – $\mathrm{D}(\pi(h|D)||\widetilde{\pi}(h|D))$ – is the Kullback-Leibler divergence of $\pi(h|D)$ on $\widetilde{\pi}(h|D)$, which reaches its minimum for $\pi(h|D) = \widetilde{\pi}(h|D)$ a.e. (see [12]),
- the remaining part of Lagrangian does not depend on $\pi(h|D)$ and does not influence the minimization.

As we can see the posterior pdf $\widetilde{\pi}(h|D)$ of h is then 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). Using 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:

$$\begin{aligned}
\mathrm{E}_{\widetilde{\pi}(h|D)}(h(\mathbf{y})|D) = \widetilde{h}(\mathbf{y}) = \frac{\nu_{\mathbf{y}}}{\nu_0} &= \frac{1 + \sum_{j=1}^{s} \lambda_j(D) g^{(j)}_{\mathbf{Y}}(\mathbf{y})}{\sum_{\mathbf{y}\in\mathcal{Y}} \nu_{\mathbf{y}}} \\
&= \overbrace{\frac{1}{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)}}^{\lambda^*_0(D)} + \sum_{j=1}^{s} \overbrace{\frac{\lambda_j(D)}{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)}}^{\lambda^*_j(D)} g^{(j)}_{\mathbf{Y}}(\mathbf{y}) \qquad (12) \\
&= \lambda^*_0(D) + \sum_{j=1}^{s} \lambda^*_j(D) g^{(j)}_{\mathbf{Y}}(\mathbf{y}), \quad \forall\, \mathbf{y} \in \mathcal{Y}
\end{aligned}$$

with

$$\sum_{\mathbf{y}\in\mathcal{Y}} \left[\lambda^*_0(D) + \sum_{j=1}^{s} \lambda^*_j(D) g^{(j)}_{\mathbf{Y}}(\mathbf{y}) \right] = \frac{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D) \overbrace{\sum_{\mathbf{y}\in\mathcal{Y}} g^{(j)}_{\mathbf{Y}}(\mathbf{y})}^{=1}}{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)} = 1.$$

If we denote the number of realizations of $\mathbf{Y}$ by n $(< \infty)$, then the final formula (12) will have following form:

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

Note: If there is no data available, the second part of (12) will disappear, while the first part – $\frac{1}{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D) \sum_{\mathbf{y}\in\mathcal{Y}} g^{(j)}_{\mathbf{Y}}(\mathbf{y})}$ – will remain. This can be considered as a prior pmf for $\mathbf{Y}$:

$$\widetilde{h}_0(\mathbf{y}) = \frac{1}{\sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)}.$$

3 Connection to the Bayesian solution

As promised earlier (see Section 1) we will now check if the final merger (12) 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 $X_1, \dots, X_s$, $(s < \infty)$, 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,...,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 $X_1, \dots, X_s$ is then

$$\begin{aligned}\pi(\boldsymbol{\theta}|X_1 = x_1, \dots, X_s = x_s) &\propto q(\boldsymbol{\theta}) \prod_{j=1}^{s} f_{X_j}(x_j|\boldsymbol{\theta}) \\ &= \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y - 1} \prod_{j=1}^{s} \prod_{y\in\mathcal{Y}} \theta_y^{\delta(x_j - y)} = \prod_{y\in\mathcal{Y}} \left(\theta_y^{\alpha_y - 1} \prod_{j=1}^{s} \theta_y^{\delta(x_j - y)} \right) \\ &= \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y) - 1} \qquad (13)\end{aligned}$$

Since the formula (13) 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 $X_1, \dots, X_s$ as follows:

$$\begin{aligned}\mathrm{E}_{\pi(\boldsymbol{\theta}|X_1,\dots,X_s)}(\theta_y|X_1 = x_1, \dots, X_s = x_s) = \widetilde{P}(Y = y) &= \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \left[\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)\right]} \\ &= \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \alpha_y + s} \qquad (14)\end{aligned}$$

Under the following choice:

$$\alpha_y = 1 \quad \forall\, y \in \mathcal{Y} \qquad (15)$$

formula (14) will look as follows

$$\widetilde{P}(Y = y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} 1 + s}. \qquad (16)$$

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 (16) – $\frac{1}{\sum_{y\in\mathcal{Y}} 1 + s}$ – 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 (15) coincides with the statement, that the prior pmf for Y is a pmf of Uniform distribution.

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 discrete random variable Y having realizations $\{y\} = \mathcal{Y}$. This means that their domains are the same (it is just one discrete random variable):

$$Y_1 = \ldots = Y_j = \ldots = Y_s = Y.$$

Since they have the same domain, they are neighbours and so the merging can be applied on them.

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 (12):

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) g_Y^{(j)}(y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)} = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)}, \tag{17}$$

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

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + s}, \tag{18}$$

which coincides with (16). 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 (12) will reduce to the standard Bayesian learning considered in Subsection 3.1.

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

Remark

In the note after the final merger (12) we brought the explanation of what should its first part represent – it stands for the prior pmf for considered random vector $\mathbf{Y}$. In the case considered 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 [8]) 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 introduces 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 it is based on Bayesian framework, we also showed that reduces to a standard Bayesian learning if independent identically distributed data are of disposal for parameter estimation. Still there are some open problems and topics of the future work, for instance the choice of constraints $\beta_j(D)$ in (10) and the extension to the continuous space.

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

[8] 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.

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

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

[11] S. Kullback. *Information theory and statistics. Reprint of the 2nd ed. '68.* Mineola, NY: Dover Publications, Inc. xvi, 399 p. $ 12.95 , 1997.

[12] S. Kullback and R.A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.