A maximum entropy approach to multiple classifiers combination

Francois Fouss & Marco Saerens
Information Systems Research Unit (ISYS)
IAG – Université catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium
Email: {saerens,fouss}@isys.ucl.ac.be

March 23, 2004

Abstract

In this paper, we present a maximum entropy (maxent) approach to the fusion of experts opinions, or classifiers outputs, problem. The maxent approach is quite versatile and allows us to express in a clear, rigorous, way the a priori knowledge that is available on the problem. For instance, our knowledge about the reliability of the experts and the correlations between these experts can be easily integrated: Each piece of knowledge is expressed in the form of a linear constraint. An iterative scaling algorithm is used in order to compute the maxent solution of the problem. The maximum entropy method seeks the joint probability density of a set of random variables that has maximum entropy while satisfying the constraints. It is therefore the "most honest" characterization of our knowledge given the available facts (constraints). In the case of conflicting constraints, we propose to minimise the "lack of constraints satisfaction" or to relax some constraints and recompute the maximum entropy solution. The maxent fusion rule is illustrated by some simulations.

1. Introduction

The fusion of various sources of knowledge has been an active subject of research since more than three decades (for some review references, see [2], [5], [7]). It has recently been successfully applied to the problem of classifiers combination or fusion (see for instance [12]).

Many different approaches have been developped for experts opinions fusion, including weighted average (see for instance [2], [7]), Bayesian fusion (see for instance [2], [7]), majority vote (see for instance [1], [11], [15]), models coming from incertainty reasoning: fuzzy logic, possibility theory [13] (see for instance [3]), standard multivariate statistical analysis techniques such as correpondence analysis [17], etc. One of these approaches is based on maximum entropy modeling (see [16], [18]). Maximum entropy is a versatile modeling technique allowing to easily integrate various constraints, such as correlation between experts, reliability of these experts, etc.

In this work, we propose a new model of experts opinions integration, based on a maximum entropy model (for a review of maximum entropy theory and applications, see for instance [6], [8], [9] or [10]). In this paper, we use the term "experts opinions", but it should be clear that we can use exactly the same procedures for "classifiers combination". In other words, we could substitute "experts" by "classifiers" everywhere.

Here is the rationale of the method. Each expert expresses his opinion about the outcome of a random event, $y = i$, in the form of an a posteriori probability density, called a score. These scores are subjective expectations about this event. We also suppose that we have access to a reliability measure for each expert, for instance in the form of a probability of success, as well as a measure of the correlation between the experts. Each of these measures are combined properly by maximum entropy in order to obtain a joint probability density. Let us recall that the maximum entropy density is the density that is "least informative" while satisfying all the constraints; i.e. it does not introduce "extra ad hoc information" that is not relevant to the problem. Once this joint density is found, we compute the a posteriori probability of the event by averaging all the possible situations that can be encountered, i.e. by computing the marginal $\mathrm{P}(y = i|\mathbf{x})$, where $\mathbf{x}$ is the feature vector on which we base our prediction.

While the main idea is similar, our model differs from [18] in the formulation of the problem (we focus on quantities that are relevant to classification problems, and can easily be computed for classifiers: success rate, degree of agreement, etc) and in the way the individual opinions are aggregated. Furthermore, we also tackle the problem of incompatible constraints; that is, when there is no feasible solution to the problem, a situation that is not mentionned by [18].

Section 2 introduces the problem and our notations. Section 3 develops the maximum entropy solution. Section 4 presents some simulations results. Section 5 is the conclusion.

2. Statement of the problem

Suppose we observe the outcome of a set of events, $\mathbf{x}$, as well as a related event, y, whose outcomes belong to the set $\{1, 2, \ldots, n\}$. We hope that the random vector $\mathbf{x}$ provides some useful information that allows to predict the outcome of y with a certain accuracy.

We also assume that domain experts (m experts in total) have expressed their opinion on the event y, based on the observation of $\mathbf{x}$: We denote by $d(k) = i$, with $i \in \{1, 2, \ldots, n\}$ and $k = 1, \ldots, m$, the fact that expert k chooses the outcome or alternative i – in other words, he takes decision i. In this framework, $\mathrm{P}(d(k) = i|\mathbf{x})$ will be interpreted as the personal expectation of the expert, i.e. the proportion of times a given expert k would choose alternative i, when observing $\mathbf{x}$ (for a general introduction to the concept of subjective probabilities, see [15]).

Our objective is to seek the joint probability density of the event, $y = i$, as well as the experts opinions, $d(k) = i_k$:

$$\mathrm{P}(y = i, d(1) = i_1, d(2) = i_2, \ldots, d(m) = i_m|\mathbf{x}) \tag{2.1}$$

This joint probability density will be estimated by using a maximum entropy argument that will be presented in the next section.

Prior knowledge on the problem, including expert's opinions, will be expressed as linear constraints on this joint density (2.1). In our case, there will be four different types of constraints:

1. Constraints ensuring that (2.1) is a **probability density** (it sums to one);
2. Constraints related to the **opinion** of the experts;
3. Constraints related to the **reliability** of the experts;
4. Constraints related to the **correlation** between experts.

These constraints are detailled in the four following subsections.

2.1. Constraints inducing a probability density

The first constraint simply states that the joint density sum to one:

$$\sum_{i,i_1,\ldots,i_m=1}^{n} \mathrm{P}(y=i, d(1)=i_1, d(2)=i_2, \ldots, d(m)=i_m|\mathbf{x}) = 1 \tag{2.2}$$

This constraint will be called the sum (*sum*) constraint. Of course, we should also impose that the joint density is always positive, but this is not necessary (maximum entropy estimation leads to positive values, so that this constraint will be automatically satisfied).

2.2. Constraints related to the opinions of the experts

Here, we provide information related to the expert's opinions. We will consider that each expert expresses his opinion about the outcomes, according to the observation $\mathbf{x}$:

$$\mathrm{P}(d(k)=i_k|\mathbf{x}) = \pi(d(k)=i_k|\mathbf{x}) \text{ for } k=1\ldots m, i_k=1\ldots n \tag{2.3}$$

where $\pi(d(k)=i_k|\mathbf{x})$, the likelihood of choosing alternative i_k, is provided by expert k for each outcome $i_k \in \{1,2,\ldots,n\}$. In other words, each expert provides his likelihood of observing outcome i_k according to his subjective judgement. It indicates that, in average, expert k would choose alternative i_k with probability $\pi(d(k)=i_k|\mathbf{x})$ when he observes evidence $\mathbf{x}$. This constraint will be called the opinion (*op*) constraint.

Notice that (2.3) can be rewritten as

$$\sum_{i,i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_m} \mathrm{P}(y=i, d(1)=i_1, \ldots, d(m)=i_m|\mathbf{x}) = \pi(d(k)=i_k|\mathbf{x}) \tag{2.4}$$

for $k=1\ldots m$ and $i_k=1\ldots n$. Or, equivalently,

$$\sum_{i,i_1,\ldots,i_m} \delta(i_k-j_k)\,\mathrm{P}(y=i, d(1)=i_1, \ldots, d(m)=i_m|\mathbf{x}) = \pi(d(k)=j_k|\mathbf{x}) \tag{2.5}$$

where δ is the delta of Kronecker.

2.3. Constraints related to the reliability of the experts

Some experts may be more reliable than others. We can express this fact by, for instance, recording the success rate of each expert. This can be expressed formally by

$$\sum_i \mathrm{P}(y = i, d(k) = i|\mathbf{x}) = \Delta(k|\mathbf{x}) \text{ for } k = 1 \ldots m \tag{2.6}$$

$\Delta(k|\mathbf{x})$ can be interpreted as the success rate for expert k, the probability of taking the correct decision (the probability that the opinion of the expert and the outcome of the event agree) when observing $\mathbf{x}$. If $\Delta(k|\mathbf{x}) = 1$, expert k is totally reliable in the sense that the judgement of the expert and the outcome of the experiment always agree. On the other hand, if $\Delta(k|\mathbf{x}) = 0$, expert k is always wrong (he always disagrees with the outcome of the experiment). Now, if the reliability is only known without reference to the context $\mathbf{x}$ (or we do not have access to this detailled information) we could simply state that it is independent of $\mathbf{x}$ which, of course, is much more restrictive:

$$\sum_i \mathrm{P}(y = i, d(k) = i|\mathbf{x}) = \Delta(k) \text{ for } k = 1 \ldots m \tag{2.7}$$

Where we do not require knowledge of the probability of success for all situations $\mathbf{x}$. This constraint will be called the reliability (*rel*) constraint.

(2.7) can be rewritten as

$$\sum_{i,i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_m} \mathrm{P}(y = i, d(1) = i_1, \ldots, d(k) = i, \ldots, d(m) = i_m|\mathbf{x}) = \Delta(k) \tag{2.8}$$

or

$$\sum_{i,i_1,\ldots,i_m} \delta(i - i_k)\, \mathrm{P}(y = i, d(1) = i_1, \ldots, d(m) = i_m|\mathbf{x}) = \Delta(k) \tag{2.9}$$

2.4. Constraints related to the correlations between experts

It is well known that experts opinions can be correlated. A possible choice for modeling experts correlations would be to provide

$$\sum_i \mathrm{P}(d(k) = i, d(l) = i|\mathbf{x}) = \sigma(k, l|\mathbf{x}) \text{ for } k, l = 1 \ldots m \tag{2.10}$$

It corresponds to the probability that expert k and expert l agree. If $\sigma(k, l|\mathbf{x}) = 1$, expert k and expert l always agree (they are totally correlated), while if $\sigma(k, l|\mathbf{x}) = 0$, they always disagree. If we only know the correlation without reference to the context $\mathbf{x}$, we must postulate independence with respect to the context, i.e.

$$\sum_i \mathrm{P}(d(k) = i, d(l) = i|\mathbf{x}) = \sigma(k, l) \text{ for } k, l = 1 \ldots m \tag{2.11}$$

This constraint will be called the correlation (*cor*) constraint. Once more, we can rewrite (2.11) as

$$\sum_{i,i_1,\ldots,i_m} \delta(i_k - i_l)\,\mathrm{P}(y = i, d(1) = i_1, \ldots, d(m) = i_m|\mathbf{x}) = \sigma(k,l) \tag{2.12}$$

We will now see how to compute the joint probability distribution satisfying the set of constraints (2.2), (2.3), (2.6), (2.10).

In the case of classifiers combination, the values of Δ and σ should be readily available based on statistics recorded on a training set or previous classification tasks.

3. The maximum entropy approach

3.1. A score of aggregation for expert's opinions

As already stated, we would like to estimate the joint probability density

$$\mathrm{P}(y = i, d(1) = i_1, d(2) = i_2, \ldots, d(m) = i_m|\mathbf{x}) \tag{3.1}$$

satisfying the set of constraints (2.2), (2.3), (2.6), (2.10). The maximum entropy estimate of (3.1) will be denoted by

$$\widehat{\mathrm{P}}(y = i, d(1) = i_1, d(2) = i_2, \ldots, d(m) = i_m|\mathbf{x}) \tag{3.2}$$

with a hat. From this joint density, (3.2), we will compute the a posteriori probability of the true outcome $y = i$

$$\widehat{\mathrm{P}}(y = i|\mathbf{x}) = \sum_{i_1,\ldots,i_m} \widehat{\mathrm{P}}(y = i, d(1) = i_1, \ldots, d(m) = i_m|\mathbf{x}) \text{ for } i = 1 \ldots n \tag{3.3}$$

and this score will define our **score of aggregation** for expert's opinions. It represents the probability of outcome $y = i$ satisfying all the constraints provided by the experts and based on the estimated density that has maximum entropy. In equation (3.3), we average on all the possible situations that can appear in the context $\mathbf{x}$; that is, on all the different decisions of the experts, where each situation is weighted by its probability of appearance.

Notice that, in a different framework, Myung et al. [18] proposed to compute the a posteriori probability of y conditional on expert's probability of taking a given decision. This is, however, not well-defined since the experts provide a subjective probability density and not a clear decision: $\pi(d(k) = i_k|\mathbf{x})$ is not a random variable; the authors are therefore conditioning on a probability density and not an event.

If we define $\mathbf{d} = [d(1), d(2), \ldots, d(m)]^{\mathrm{T}}$ and $\mathbf{i} = [i_1, i_2, \ldots, i_m]^{\mathrm{T}}$, we can rewrite (3.3) in a more compact way as

$$\widehat{\mathrm{P}}(y = i|\mathbf{x}) = \sum_{\mathbf{i}} \widehat{\mathrm{P}}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) \text{ for } i = 1 \ldots n \tag{3.4}$$

We will now see how to compute these scores thanks to the maximum entropy principle.

3.2. The maximum entropy estimate

Our aim is to estimate $\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x})$ by seeking the probability density that has maximum entropy

$$I = -\sum_{i,\mathbf{i}} \mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) \log \left[\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x})\right] \tag{3.5}$$

among all the densities satisfying the constraints (2.2), (2.3), (2.6), (2.10). This problem has been studied extensively in the litterature. We show in Appendix A that $\widehat{\mathrm{P}}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x})$ takes the form

$$\begin{aligned}\widehat{\mathrm{P}}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) \quad &= \quad G_{sum} \prod_{k=1}^{m} G_{op}(k, i_k) \\ &\times \prod_{k=1}^{m} \left[G_{rel}(k)\right]^{\delta(i-i_k)} \prod_{k=1}^{m} \prod_{l=1}^{m} \left[G_{cor}(k, l)\right]^{\delta(i_k - i_l)}\end{aligned} \tag{3.6}$$

where the parameters G_{sum}, G_{op}, G_{rel}, G_{cor} can be estimated iteratively by an iterative scaling procedure (see next section).

3.3. Computing the maximum entropy estimate

The first step is to verify that there is a feasible solution to the problem at hand. Since all the constraints are linear, a linear programming procedure can be used to solve this problem.

Once we have verified that there is indeed a feasible solution, an iterative scaling procedure allowing to estimate the G_{sum}, G_{op}, G_{rel}, G_{cor} can easily be derived (see for instance [4]). The iterative scaling procedure aims to satisfy in turn each constraint, and iterate on the set of constraints (as proposed by many authors, for instance [14]). It has been shown that this iterative procedure converges to the solution provided that there exists a feasible solution to the problem (that is, the set of constraints can be satisfied). Indeed, the entropy criterion is convex and the constraints are linear so that convex programming algorithms can be used in order to solve the problem.

3.4. The case where there is no feasible solution

It can be the case that no solution satisfying the constraints (2.2), (2.3), (2.6), (2.10) exists. This means that there is a conflict between the different estimates π, Δ, σ, so that this situation cannot normally appear in reality. In that case, the user of the system should revise his different pieces of knowledge or data.

However, despite this conflicting situation, if the user nevertheless wants to compute an aggregated score, we have to relax in some way the set of constraints. There are two different ways of doing this: (1) by introducing slack variables that compute the lack of constraint satisfaction, or (2) to relax the equality constraints by providing intervals instead of exact values. These two approaches are introduced in the two next sections.

3.4.1. Introduction of slack variables.

In this case, some equality constraints are relaxed. For instance, let us consider that we are willing to relax the reliability and correlation constraints for all experts. By introducing slack variables, ξ_k^{r+}, ξ_k^{r-}, ξ_{kl}^{c+}, ξ_{kl}^{c-}, measuring the lack of constraint satisfaction for each constraint, we have

$$\sum_i \mathrm{P}(y = i, d(k) = i|\mathbf{x}) + \xi_k^{r+} - \xi_k^{r-} = \Delta(k|\mathbf{x}) \text{ for } k = 1 \ldots m \tag{3.7}$$

$$\sum_i \mathrm{P}(d(k) = i, d(l) = i|\mathbf{x}) + \xi_{kl}^{c+} - \xi_{kl}^{c-} = \sigma(k, l|\mathbf{x}) \text{ for } k, l = 1 \ldots m \tag{3.8}$$

$$0 \leq \xi_k^{r+}, \xi_k^{r-}, \xi_{kl}^{c+}, \xi_{kl}^{c-} < \theta \tag{3.9}$$

where θ is a treshold provided by the user: the slack variables are not allowed to exceed this treshold. Consequently we want to minimize the lack of constraint satisfaction

$$\min \left[\sum_k (\xi_k^{r+} + \xi_k^{r-}) + \sum_{k,l} (\xi_{kl}^{c+} + \xi_{kl}^{c-}) \right] \tag{3.10}$$

subject to constraints (2.2), (2.3), (3.7), (3.8), (3.9). This is a standard linear programming problem.

3.4.2. Introduction of intervals.

Another alternative would be to relax the equality constraints by providing intervals instead of exact values. Once again, let us consider that we are willing to relax the reliability and correlation constraints for all experts. The problem would be reformulated as a maximum entropy problem with inequality constraints:

$$\Delta^-(k|\mathbf{x}) \leq \sum_i \mathrm{P}(y = i, d(k) = i|\mathbf{x}) \leq \Delta^+(k|\mathbf{x}) \text{ for } k = 1 \ldots m \tag{3.11}$$

$$\sigma^-(k, l|\mathbf{x}) \leq \sum_i \mathrm{P}(d(k) = i, d(l) = i|\mathbf{x}) \leq \sigma^+(k, l|\mathbf{x}) \text{ for } k, l = 1 \ldots m \tag{3.12}$$

Once more, numerical procedures related to iterative scaling can be used in order to compute the maximum entropy solution [4]. We have to maximize (3.5) subject to constraints (2.2), (2.3), (3.11), (3.12).

3.5. Some extensions

The maximum entropy model presented in Section 2 can be extended in several ways. For instance, if we know the full "confusion matrix" for each expert, we could exploit this knowledge by defining the constraints

$$\mathrm{P}(y = i, d(k) = j) = \Delta(y = i, d(k) = j) \text{ for } k = 1 \ldots m \text{ and } i, j = 1 \ldots n \tag{3.13}$$

In addition, we also know the full correlation matrix between the experts:

$$\mathrm{P}(d(k) = i, d(l) = j) = \sigma(d(k) = i, d(l) = j) \text{ for } k, l = 1 \ldots m \text{ and } i, j = 1 \ldots n \tag{3.14}$$

The maximum entropy joint density can be computed in the same way as before.

4. Simulation results

For illustration purposes, we used the proposed combination rule in three different conditions. For each condition, we compute the following values:

Maximum entropy. If there is a feasible solution, we compute the *maximum entropy* solution to the problem.

Linear programming. If there is no feasible solution, we compute the *linear programming* solution to the problem (see (3.4.1)).

Weighted average. We also compute a *weighted average* solution, for comparison.

The weighted average is computed as follows. For each expert, we associate a weight, $w(k)$, which is a normalised (it sums to one) measure of his reliability and we assume that each reliability, $\Delta(k|\mathbf{x})$, is greater than 0.5 (the expert performs better than a random guess): $w(k) = (\Delta(k) - 0.5)/\sum_k (\Delta(k) - 0.5)$

The weighted average score is: $Score(i) = \sum_k w(k)\, \pi(d(k) = i|\mathbf{x})$

Notice that we did not introduce correlation constraints in this set of simulations. For all simulations, we consider the case where there are three experts (1, 2, 3) and two outcomes (0, 1).

4.1. First set of simulations

We set, for experts' opinions, $\pi(d(1) = 0|\mathbf{x}) = 0.3$, $\pi(d(1) = 1|\mathbf{x}) = 0.7$, $\pi(d(2) = 0|\mathbf{x}) = 0.3$, $\pi(d(2) = 1|\mathbf{x}) = 0.7$, $\pi(d(3) = 0|\mathbf{x}) = 0.8$, $\pi(d(3) = 1|\mathbf{x}) = 0.2$. In other words, the two first experts agree, while the third one has an opposite opinion. For experts' reliability, we set, $\Delta(1) = 0.7$, $\Delta(2) = 0.7$, $\Delta(3) = z$, where $0.5 < z < 1$. The results are shown in Figure 4.1, where we display $\widehat{\mathrm{P}}(y = 0|\mathbf{x})$ and $Score(0)$ in terms of the reliability of expert 3, i.e. z.

We observe that when the reliability of expert 3 is high, the fusion rules (that is, the experts combination rules) favour outcome 0. Notice that when $z > 0.8$, there is no feasible solution, that is, the constraints cannot be satisfied. Notice also that the maximum entropy solution is always in favour of outcome 0, in comparison with the weighted average ($\widehat{\mathrm{P}}(y = 0|\mathbf{x}) > Score(0)$).

4.2. Second set of simulations

In this second example, we set, for experts' opinions, $\pi(d(1) = 0|\mathbf{x}) = 0.85$, $\pi(d(1) = 1|\mathbf{x}) = 0.15$, $\pi(d(2) = 0|\mathbf{x}) = 0.8$, $\pi(d(2) = 1|\mathbf{x}) = 0.2$, $\pi(d(3) = 0|\mathbf{x}) = z$, $\pi(d(3) = 1|\mathbf{x}) = (1 - z)$ and, for the reliability, $\Delta(1) = 0.7$, $\Delta(2) = 0.75$, $\Delta(3) = 0.8$. The results are shown in Figure 4.2, where we display $\widehat{\mathrm{P}}(y = 0|\mathbf{x})$ and $Score(0)$ in terms of the opinion of expert 3, i.e. z.

We observe that the weighted average rule is linear in the opinion of expert 3, while maxent is nonlinear. Below the value of $z = 0.35$, we see that the constraints are incompatible; in this case, the procedure described in 3.4.1 (linear programming) is used in order to compute the aggregation score.

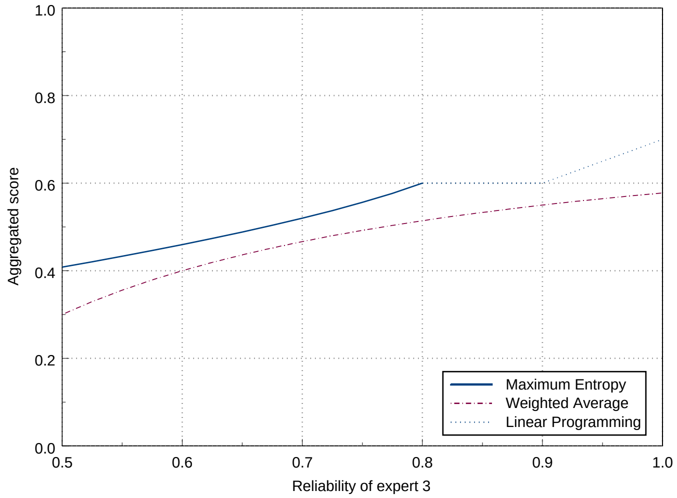

Figure 4.1: A first example of simulation of the fusion rule (see Section 4.1 for details). We display the results of the three combination rules, maximum entropy, weighted average and linear programming in terms of the reliability of expert 3.

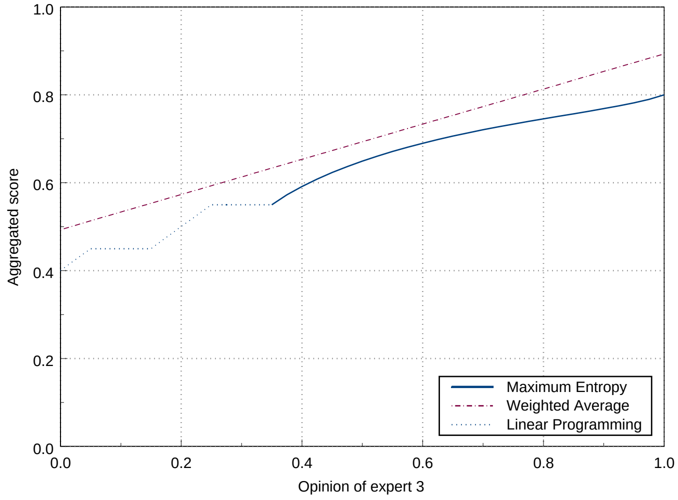

Figure 4.2: Another example of simulation of the fusion rule (see Section 4.2 for details). We display the results of the three combination rules, maximum entropy, weighted average and linear programming in terms of expert 3's opinion.

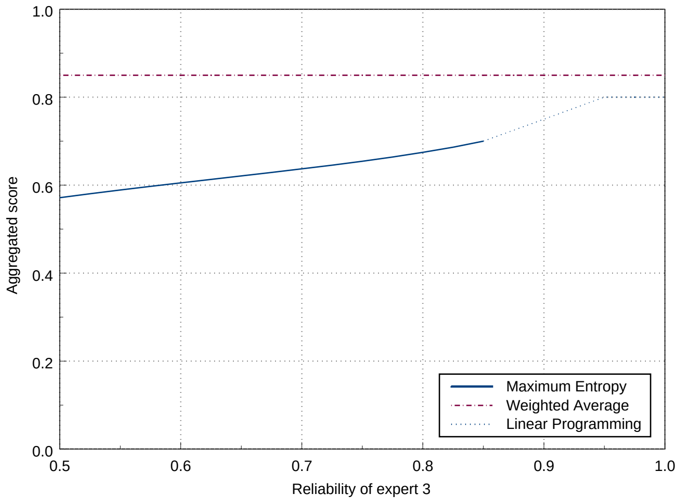

Figure 4.3: A last example of simulation of the fusion rule (see Section 4.3 for details). We display the results of the three combination rules, maximum entropy, weighted average and linear programming in terms of the reliability of expert 3.

4.3. Third set of simulations

In this last example, we set, for experts' opinions, $\pi(d(1) = 0|\mathbf{x}) = 0.9$, $\pi(d(1) = 1|\mathbf{x}) = 0.1$, $\pi(d(2) = 0|\mathbf{x}) = 0.8$, $\pi(d(2) = 1|\mathbf{x}) = 0.2$, $\pi(d(3) = 0|\mathbf{x}) = 0.85$, $\pi(d(3) = 1|\mathbf{x}) = 0.15$ and, for the reliability, $\Delta(1) = 0.6$, $\Delta(2) = 0.6$, $\Delta(3) = z$. This time, we vary the reliability score of expert 3. The results are shown in Figure 4.3, where we display $\widehat{\mathrm{P}}(y = 0|\mathbf{x})$ and $Score(0)$ in terms of the reliability of expert 3, i.e. z.

5. Conclusion

We introduced a new way of combining experts opinions or classifiers outputs. It is based on the maximum entropy framework; maximum entropy seeks the joint probability density of a set of random variables that has maximum entropy while satisfying the constraints, i.e. it does not introduce any "additional ad hoc information". It is therefore the "most honest" characterization of our knowledge given the available facts. The available knowledge is expressed through a set of linear constraints on the joint density including a measure of reliability of the experts, and of the correlation between them.

Iterative mathematical programming methods are used in order to compute the maximum entropy, with guaranteed convergence to the global maximum if, of course, there is a feasible solution. If there is a conflict between the available facts, i.e. between the constraints, so that there is no feasible solution, we could still compute the solution that is "closest" in some way to the constraints satisfaction.

By using the concept of maximum entropy, the different constraints (representing the a priori knowledge about the problem) are properly incorporated within a single measure. However, even if this approach seems promising, it has not been evaluated in the context of classifiers combination. Further work will thus be devoted to the experimental comparison with more standard techniques such as those that were mentioned in the introduction.

References

[1] D. Chen and X. Cheng. An asymptotic analysis of some expert fusion methods. *Pattern Recognition Letters*, 22:901–904, 2001.

[2] R. M. Cooke. *Experts in uncertainty*. Oxford University Press, 1991.

[3] D. Dubois, M. Grabisch, H. Prade, and P. Smets. Assessing the value of a candidate: Comparing belief function and possibility theories. *Proceedings of the Fifteenth international conference on Uncertainty in Artificial Intelligence*, pages 170–177, 1999.

[4] S.-C. Fang, J. Rajasekera, and H.-S. J. Tsao. *Entropy optimization and mathematical programming*. Kluwer Academic Publishers, 1997.

[5] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 36:114–148, 1986.

[6] A. Golan, G. Judge, and D. Miller. *Maximum entropy econometrics: Robust estimation with limited data*. John Wiley and Sons, 1996.

[7] R. A. Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7:867–888, 1995.

[8] F. Jelinek. *Statistical methods for speech recognition*. The MIT Press, 1997.

[9] J. N. Kapur and H. K. Kesavan. *The generalized maximum entropy principle (with applications)*. Sandford Educational Press, 1987.

[10] J. N. Kapur and H. K. Kesavan. *Entropy optimization principles with applications*. Academic Press, 1992.

[11] J. Kittler and F. Alkoot. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):110–115, 2003.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[13] G. J. Klir and T. A. Folger. *Fuzzy sets, uncertainty, and information*. Prentice-Hall, 1988.

[14] H. Ku and S. Kullback. Approximating discrete probability distributions. *IEEE Transactions on Information Theory*, 15(4):444–447, 1969.

[15] F. Lad. *Operational subjective statistical methods*. John Wiley and Sons, 1996.

[16] W. B. Levy and H. Delic. Maximum entropy aggregation of individual opinions. *IEEE Transactions on Systems, Man and Cybernetics*, 24(4):606–613, 1994.

[17] C. Merz. Using correspondence analysis to combine classifiers. *Machine Learning*, 36:226–239, 1999.

[18] I. J. Myung, S. Ramamoorti, and J. Andrew D. Bailey. Maximum entropy aggregation of expert predictions. *Management Science*, 42(10):1420–1436, 1996.

Acknowledgments

We thank Prof. Philippe Chevalier, from the "Institut d'Administration et de Gestion" of the Université catholique de Louvain for insightful discussions.

A. Appendix: Derivation of the maximum entropy expression

The problem is to estimate the probability density $\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x})$ that has maximum entropy and verifies the constraints (2.2), (2.3), (2.9), (2.12). We therefore build the Lagrange function

$$\begin{aligned}
\pounds \quad = \quad & -\sum_{i,\mathbf{i}=1}^{n} \mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) \log\left[\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x})\right] \\
& +\lambda_{sum}\left[\sum_{i,\mathbf{i}=1}^{n} \mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) - 1\right] \\
& +\sum_{k'=1}^{m}\sum_{i^{'}_{k'}=1}^{n} \lambda_{op}(k^{'}, i^{'}_{k'})\left[\sum_{i,\mathbf{i}=1}^{n} \delta(i^{'}_{k'} - i_{k'})\,\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) - \pi(d(k^{'}) = i^{'}_{k'})\right] \\
& +\sum_{k'=1}^{m} \lambda_{rel}(k^{'})\left[\sum_{i,\mathbf{i}=1}^{n} \delta(i - i_{k'})\,\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) - \Delta(k^{'})\right] \\
& +\sum_{k'=1}^{m}\sum_{l'=1}^{m} \lambda_{cor}(k^{'}, l^{'})\left[\sum_{i,\mathbf{i}=1}^{n} \delta(i_{k'} - i_{l'})\,\mathrm{P}(y = i, \mathbf{d} = \mathbf{i}|\mathbf{x}) - \sigma(k^{'}, l^{'})\right]
\end{aligned}$$

Now, if we derive this expression with respect to $\mathrm{P}(y = j, \mathbf{d} = \mathbf{j}|\mathbf{x})$ and set the result equal to zero, we obtain

$$\begin{aligned}
\frac{\partial \pounds}{\mathrm{P}(y = j, \mathbf{d} = \mathbf{j}|\mathbf{x})} &= -\log\left[\mathrm{P}(y = j, \mathbf{d} = \mathbf{j}|\mathbf{x})\right] - 1 + \lambda_{sum} \\
&\quad + \sum_{k'=1}^{m}\sum_{i^{'}_{k'}=1}^{n} \delta(i^{'}_{k'} - j_{k'})\lambda_{op}(k^{'}, i^{'}_{k'}) \\
&\quad + \sum_{k'=1}^{m} \delta(j - j_{k'})\lambda_{rel}(k^{'}) \\
&\quad + \sum_{k'=1}^{m}\sum_{l'=1}^{m} \delta(j_{k'} - j_{l'})\lambda_{cor}(k^{'}, l^{'}) \\
&= 0
\end{aligned}$$

So that we find

$$\begin{aligned}
\log\left[\mathrm{P}(y = j, \mathbf{d} = \mathbf{j}|\mathbf{x})\right] &= \lambda_{sum} - 1 \\
&\quad + \sum_{k'=1}^{m}\sum_{i^{'}_{k'}=1}^{n} \delta(i^{'}_{k'} - j_{k'})\lambda_{op}(k^{'}, i^{'}_{k'})
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{k'=1}^{m} \delta(j - j_{k'}) \lambda_{rel}(k^{'}) \\
&+ \sum_{k'=1}^{m} \sum_{l'=1}^{m} \delta(j_{k'} - j_{l'}) \lambda_{cor}(k^{'}, l^{'})
\end{aligned}
$$

Or, equivalently,

$$
\begin{aligned}
\log\left[\mathrm{P}(y = j, \mathbf{d} = \mathbf{j} | \mathbf{x})\right] \quad = \quad & \lambda_{sum} - 1 \\
&+ \sum_{k'=1}^{m} \lambda_{op}(k^{'}, j_{k'}) \\
&+ \sum_{k'=1}^{m} \delta(j - j_{k'}) \lambda_{rel}(k^{'}) \\
&+ \sum_{k'=1}^{m} \sum_{l'=1}^{m} \delta(j_{k'} - j_{l'}) \lambda_{cor}(k^{'}, l^{'})
\end{aligned}
$$

By taking the exponential of both sides, we find

$$
\begin{aligned}
\mathrm{P}(y = j, \mathbf{d} = \mathbf{j} | \mathbf{x}) \quad = \quad & \exp\left[\lambda_{sum} - 1\right] \prod_{k'=1}^{m} \exp\left[\lambda_{op}(k^{'}, j_{k'})\right] \\
&\times \prod_{k'=1}^{m} \exp\left[\delta(j - j_{k'}) \lambda_{rel}(k^{'})\right] \\
&\times \prod_{k'=1}^{m} \prod_{l'=1}^{m} \exp\left[\delta(j_{k'} - j_{l'}) \lambda_{cor}(k^{'}, l^{'})\right]
\end{aligned}
$$

So that by defining the following parameters

$$
\begin{cases}
G_{sum} = \exp\left[\lambda_{sum} - 1\right] \\
G_{op}(k^{'}, j_{k'}) = \exp\left[\lambda_{op}(k^{'}, j_{k'})\right] \\
G_{rel}(k^{'}) = \exp\left[\lambda_{rel}(k^{'})\right] \\
G_{cor}(k^{'}, l^{'}) = \exp\left[\lambda_{cor}(k^{'}, l^{'})\right]
\end{cases}
$$

We easily obtain

$$
\begin{aligned}
\mathrm{P}(y = j, \mathbf{d} = \mathbf{j} | \mathbf{x}) \quad = \quad & G_{sum} \prod_{k'=1}^{m} G_{op}(k^{'}, j_{k'}) \\
&\times \prod_{k'=1}^{m} \left[G_{rel}(k^{'})\right]^{\delta(j - j_{k'})} \\
&\times \prod_{k'=1}^{m} \prod_{l'=1}^{m} \left[G_{cor}(k^{'}, l^{'})\right]^{\delta(j_{k'} - j_{l'})}
\end{aligned}
$$

where the G parameters must be computed by imposing the constraints (2.2), (2.3), (2.9), (2.12). This leads to a set of nonlinear equations that can be solved numerically by an iterative scaling procedure.. ■