Fusion Rule Estimation Using Vector Space Methods †

Nageswara S. V. Rao
Center for Engineering Systems Advanced Research
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-6364
email: raons@ornl.gov

Submitted to *SPIE Conference on Sensor Fusion: Architecture and Applications*, April 21-25, 1996, Orlando, Florida.

---

†Research sponsored in part by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp., and by ORNL Seed Money Program.

DISCLAIMER

Fusion Rule Estimation Using Vector Space Methods

Nageswara S. V. Rao
Center for Engineering Systems Advanced Research
Oak Ridge National Laboratory
Oak Ridge, TN 37831
raons@ornl.gov

ABSTRACT

In a system of N sensors, the sensor S_j, $j = 1, 2 \ldots, N$, outputs $Y^{(j)} \in \Re$, according to an unknown probability distribution $P_{(Y^{(j)}|X)}$, corresponding to input $X \in [0, 1]$. A training n-sample (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) is given where $Y_i = (Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(N)})$ such that $Y_i^{(j)}$ is the output of S_j in response to input X_i. The problem is to estimate a fusion rule $f : \Re^N \mapsto [0, 1]$, based on the sample, such that the expected square error is minimized over a family of functions $\mathcal{F}$ that constitute a vector space. The function f^* that minimizes the expected error cannot be computed since the underlying densities are unknown, and only an approximation $\hat{f}$ to f^* is feasible. We estimate the sample size sufficient to ensure that $\hat{f}$ provides a close approximation to f^* with a high probability. The advantages of vector space methods are two-fold: (a) the sample size estimate is a simple function of the dimensionality of $\mathcal{F}$, and (b) the estimate $\hat{f}$ can be easily computed by well-known least square methods in polynomial time. The results are applicable to the classical potential function methods and also (to a recently proposed) special class of sigmoidal feedforward neural networks.

Subject Terms: Sensor fusion, fusion rule estimation, empirical estimation, vector space methods.

1 INTRODUCTION

Over the past decade, the area of sensor fusion has witnessed a tremendous growth due to: (a) an increasing number of applications that require solutions to difficult sensor fusion problems, and (b) advances in computational systems and methods that make it possible to process large volumes of data. The sensor fusion problems have particular relevance to engineering systems, where the fundamental limitations of single sensor systems have been realized in many applications. By employing multiple sensors: (i) replicated sensors can be employed for fault tolerance, and (ii) sensors of different modalities can be used to achieve tasks that cannot be performed by a single sensor. In either case, the fusion method must be designed carefully, since an inappropriate fuser can make the system worse than the worst individual sensor.

Several existing sensor fusion methods require either independence of sensor distributions or closed-form analytical expressions for error densities. In the former case, a general majority fusion rule suffices, while in the latter a fusion rule can be computed using Bayesian methods. Several popular distributed decision fusion methods belong to the latter class.[4,28] In engineering systems, however, independence can seldom be assured and, in fact, may not be satisfied. Also, the problem of obtaining the probability densities which are required by Bayesian methods can be more difficult than the fusion problem itself.[26] Thus practical solutions to fusion problems must exploit the empirical data available from observation and/or experimentation. Recently, such "learning" methods that estimate fusion rules based on recent advances in empirical estimation and non-linear

computational methods have been developed[12] within the framework of Probably and Approximately Correct (PAC) learning.[27,25] These methods are suited for engineering systems where the sensor system is available for operation/experimentation, but, it is difficult to obtain detailed sensor error densities.

Consider a system [1] of N sensors such that corresponding to input $X \in [0,1]$, the sensor S_i, $i = 1, 2, \ldots, N$, outputs $Y^{(i)} \in \Re$ according to an *unknown* distribution $P_{Y^{(i)}|X}$. A independently and identically distributed (iid) n-sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is given where $Y_i = (Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(N)})$ and $Y_i^{(j)}$ is the output of S_j in response to input X_i. We consider the expected square error

$$I(f) = \int [X - f(Y))]^2 dP_{Y|X} dP_X, \tag{1.1}$$

where $Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(N)})$, to be minimized over a family of fusion rules $\mathcal{F} = \{f : \Re^N \mapsto [0,1]\}$, based on the given n-sample. For simplicity, we consider the quadratic cost, but the approach is valid for general costs if suitable boundedness or smoothness conditions are satisfied (see Section 2).

Let $f^* \in \mathcal{F}$ minimize $I(.)$. In general, f^* cannot be computed since the underlying distributions are unknown. Furthermore, since no restrictions are placed on the distributions, it will not be possible to infer f^* (with probability one) based on *only* a finite sample. Consequently, only an approximation $\hat{f}$ to f^* is feasible in general. If $\mathcal{F}$ forms a finite dimensional vector space, then we show that an estimator $\hat{f}$ can be computed which satisfies

$$P[I(\hat{f}) - I(f^*) > \epsilon] < \delta, \tag{1.2}$$

where $\epsilon > 0$ and $0 < \delta < 1$. Informally, this condition states that the "error" of $\hat{f}$ is within ϵ of the optimal error (of f^*) with arbitrary high probability $1 - \delta$, given a sufficiently large sample. Such criteria have been extensively used in a number of machine learning and empirical estimation problems (see Vapnik[27] for more details). The sample size sufficient to ensure the criterion (1.2) is shown to be

$$\frac{512}{\epsilon^2}\left[d \ln\left(\frac{64e}{\epsilon} + \ln\frac{64e}{\epsilon}\right) + \ln(8/\delta)\right],$$

where d is dimension of $\mathcal{F}$. The advantages of vector space methods are two-fold: (a) the sample size estimate is a simple function of the dimensionality of $\mathcal{F}$, and (b) the estimate $\hat{f}$ can be easily computed by well-known least square methods in polynomial time. Also, our results provide new perspectives on the classical potential function method of Aizerman *et al.*[1] and recently proposed feedforward sigmoidal networks of Kurkova.[8]

The sensor fusion problem (1.1) under the criterion (1.2) was first formulated in Rao[12] and was further developed in Rao.[13,15,16] The special case of decision fusion where $Y_i \in \{0,1\}^N$ has been solved using majority rules,[20,18] empirical Bayesian rules,[14,19] and nearest neighbor rules.[17]

The paper is organized as follows. In Section 2, we show that for a sufficiently large sample, the criterion (1.2) can be satisfied when $\mathcal{F}$ is a finite-dimensional vector space. We then discuss computational issues and some well-known examples of $\mathcal{F}$ in Section 3.

2 SAMPLE SIZE ESTIMATION

Let S be a set equipped with a pseudometric d_n. The *covering number* $N(\epsilon, d_n, S)$ is defined as the smallest number of closed balls of radius ϵ, and centers in S, whose union covers S. Let $Y_1, Y_2, \ldots, Y_n$ denote the iid sample and $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. We define $\mathcal{F}_{\vec{Y}} = \{(f(Y_1), f(Y_2), \ldots, f(Y_n)) : f \in \mathcal{F}\} \subseteq [0,1]^n$. Consider the random variable $N(\epsilon, d_n, \mathcal{F}_{\vec{Y}})$ where $d_n : [0,1]^n \times [0,1]^n \mapsto [0,1]$ is defined as $d_n(\vec{x}, \vec{z}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - z_i|$, for $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{z} = (z_1, z_2, \ldots, z_n)$. This cover size plays an essential role in the convergence of

[1] The approach of this paper can be directly extended to the case where $Y^{(i)} \in \Re^d$ and $X \in [-\tau, \tau]$, for $0 < \tau < \infty$.

empirical values of the functions to their expectations. We state a result which is an adaptation of Pollard's result[11] by Lugosi and Zeger.[10]

LEMMA 2.1. *Let $\mathcal{F}$ be a class of measurable functions from A into $[0,1]$, and P be a probability measure defined on A. Then*

$$P\left\{\sup_{f\in\mathcal{F}}|P_nf - Pf| > \epsilon\right\} \leq 4E\left[N(\epsilon/16, d_n, \mathcal{F}_{\vec{Y}})\right] e^{-\epsilon^2 n/128}$$

where $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, $Pf = \int f(y)dP$ and $P_nf = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)$.

Now consider a cover size for function classes. Let $\mathcal{F} = \{f : \Re^N \mapsto [0,1]\}$. Consider $N(\epsilon, d_P, \mathcal{F})$, where

$$d_P(f_1, f_2) = \int_{y\in\Re^N} |f_1(y) - f_2(y)|dP,$$

for a probability distribution defined on $\Re^d$. If $\mathcal{F}$ forms a vector space of dimensionality d, then its cover size can be upperbounded as follows as a direct consequence of results of Cover[2] and Haussler.[7]

LEMMA 2.2. *Let $\mathcal{F}$ denote d dimensional vector space of functions defined on A with range $[0,1]$. Then for any probability measure P defined on A, we have*

$$N(\epsilon, d_P, \mathcal{F}) \leq 2\left(\frac{2e}{\epsilon}\ln\frac{2e}{\epsilon}\right)^d.$$

Proof: This bound is obtained by first showing that the VC-dim of the sets of the form $\{\{x : f(x) \geq 0\} : f \in \mathcal{F}\}$ to be d using the result of Cover[2] (also see[11,6]). The VC-dim is used by Haussler[7] to obtain the bound on the cover size. □

Since the bound in Lemma 2.2 is valid for any P defined on A, we have $N(\epsilon/16, d_n, \mathcal{F}_{\vec{Y}}) \leq \left(\frac{2e}{\epsilon}\ln\frac{2e}{\epsilon}\right)^d$, by noting that d_n specifies a discrete uniform probability meaure on $\vec{Y}$ with mass $1/n$ at each Y_i.

THEOREM 2.3. *Let f^* and $\hat{f}$ denotes the expected best and empirical best fusion rules chosen from a vector space $\mathcal{F}$ of dimension d and range $[0,1]$. Given an iid sample of size*

$$\frac{512}{\epsilon^2}\left[d\ln\left(\frac{64e}{\epsilon}+\ln\frac{64e}{\epsilon}\right)+\ln(8/\delta)\right],$$

we have $P[I(\hat{f}) - I(f^*) > \epsilon] < \delta$.

Proof: By the result of Vapnik[26] we have

$$P\left\{I(\hat{f}) - I(f^*) > \epsilon\right\} \leq P\left\{\sup_{g\in\mathcal{G}}|P_ng - Pg| > \epsilon/2\right\},$$

where $\mathcal{G} = \{g(x,y) = (x - f(y))^2 : f \in \mathcal{F}\}$, $Pg = \int g(x,y)dP$ and $P_ng = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$. Consider $g_1, g_2 \in \mathcal{G}$ such that $g_i(x,y) = (x - f_i(y))^2$, $i = 1,2$. Now we have

$$|g_1(x,y) - g_2(x,y)| \leq |[2x - f_1(y) - f_2(y)][f_1(y) - f_2(y)]| \leq 2|f_1(y) - f_2(y)|$$

which implies, by Lemma 2.2

$$N(\epsilon, d_n, \mathcal{G}) \leq N(\epsilon/2, d_n, \mathcal{F}) \leq 2\left(\frac{4e}{\epsilon}\ln\frac{4e}{\epsilon}\right)^d.$$

By Lemma 2.1, we have

$$P\left\{I(\hat{f}) - I(f^*) > \epsilon\right\} \leq 8\left(\frac{64e}{\epsilon}\ln\frac{64e}{\epsilon}\right)^d e^{-\epsilon^2 n/512}.$$

The right hand side is upperbounded by δ for the sample size n given in the theorem. □

This result can be generalized to functions with range $[-\tau, \tau]$, for $\tau < \infty$, without changing the overall functional dependence on ϵ, δ and d. Also, more general cost functions can be considered. Consider $\mathcal{F} = \{f : \Re^N \mapsto [0,1]\}$. Given $f_1, f_2 \in \mathcal{F}$, we say that $f_1 \leq f_2$ if $f_1(y) \leq f_2(y)$ for all $y \in \Re^d$. The function $|f_1 - f_2|$ is defined as $|f_1 - f_2|(y) = |f_1(y) - f_2(y)|$ at every $y \in \Re^N$. The cost function $\Theta(.)$ defined on $\mathcal{F}$ satisfies *Lipschitz property* if there exists a positive constant Γ_Θ such that $|\Theta(f_1) - \Theta(f_2)| \leq \Gamma_\Theta |f_1 - f_2|$ for all $f_1, f_2 \in \mathcal{F}$. The square error cost defined above is a special case with $\Gamma_\Theta = 2$, for $x \in [0,1]$. The Theorem 2.1 can be generalized to account for the Lipschitz cost functions.

3 FUNCTION SPACES AND COMPUTATIONAL PROBLEM

Let $\{f_1, f_2, \ldots, f_d\}$ be the basis of $\mathcal{F}$ such that $f \in \mathcal{F}$ can be written as $f(y) = \sum_{i=1}^{d} a_i f_i(y)$ for $a_i \in \Re$. Then consider $\hat{f} = \sum_{i=1}^{d} \hat{a}_i f_i(y)$ such that $\hat{a} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_d)$ minimizes the cost expressed as (with abuse of notation)

$$I_{emp}(\vec{a}) = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \sum_{i=1}^{d} a_i f_i(Y_k)\right)^2,$$

where $\vec{a} = (a_1, a_2, \ldots, a_d)$. This is the well-known least squares problem, which can be solved by a number of methods (for example see,[9] Chapter 10). Now $I_{emp}(\vec{a})$ can be written in the following form:

$$I_{emp}(\vec{a}) = \frac{1}{n}\sum_{k=1} X_k^2 + \sum_{i=1}^{d}\sum_{j=1}^{d} a_i c_{ij} a_j + \sum_{i=1}^{d} a_i d_i$$

where

$$c_{ij} = \frac{1}{n}\sum_{k=1} f_i(Y_k) f_j(Y_k)$$

$$d_i = \frac{-2}{n}\sum_{k=1}^{n} f_i(Y_k) X_k.$$

Thus $\hat{a} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_d)$ is obtained by minimizing the quadratic form $a^T C a + a^T d$, where $C = [c_{ij}]$ is a positive definite symmetric matrix, and $D = [d_i]$. This problem is known to be polynomial-time solvable using quadratic programming methods.[29]

One of the earliest candidates for $\mathcal{F}$ is the set of polynomials of fixed degree l (which form a vector space of dimension $l+1$).

The potential functions of Aizerman *et al.*,[1] where $f_i(y)$ of the form $exp((y-\alpha)^2/\beta)$, for suitably chosen constants α and β, constitute another example of the vector space methods. An incremental algorithm was originally proposed for the computation of the coefficient vector $\vec{a}$, for which finite sample results have been derived recently[21] under certain conditions. The sample size estimate of this paper is simpler and is proportional to the number of component functions, as opposed to the complicated form of the existing finite sample results (e.g. dependence[21] on eigenvalues of the correlation matrix). Note that the sample size of this paper is valid only for the method that minimizes I_{emp} and is not valid for the original incremental algorithm of the potential functions.

More recent examples of vector space methods are the two-layer sigmoidal networks of Kurkova,[8] where the only unknown weights are in the output layer (also see[5]). The specific form of these networks enables us to express each network in the form $\sum_{k=1}^{d} a_i \eta_i(y)$ where $\eta_i(.)$'s are universal. These networks have been shown to approximate classes of continuous functions with arbitrarily specified precision, in a manner similar to the general single layer sigmoidal networks (shown by Cybenko[3]). We are unaware of any previous finite sample and computational results for function estimation based on this method. Based on the results presented in this paper, we have a simple bound for these networks based on a polynomial-time computable solution. This is in contrast with the general feedforward sigmoidal networks, where the sample size estimate is fairly complicated,[22,23] and the computational problem is very hard.[24]

4 CONCLUSIONS

We presented a class of solutions to a general sensor fusion problem, where the underlying sensor error distributions are not known but a sample is available. The advantages of vector space methods are two-fold: (a) the sample size estimate is a simple function of the dimensionality of $\mathcal{F}$, and (b) the estimate $\hat{f}$ can be easily computed by well-known least square methods in polynomial time. In addition, this work provides a new perspective on the computational and finite sample aspects of the classical potential function methods[1] and a special type of sigmoidal neural networks.[8]

Several issues of the fusion rule estimation are open problems. In our sample size estimates no efforts are made to optimize the constants; we believe much smaller values for the constants can be obtained. It would be interesting to obtain lower bounds for the sample sizes in order to judge the tightness of bounds proposed here.

ACKNOWLEDGEMENTS

This research is sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.

5 REFERENCES

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. *Extrapolative problems in automatic control and method of potential functions*, volume 87 of *American Mathematical Society Translations*, pages 281–303. 1970.

[2] T. Cover. Geometric and statistical properties of systems of linear inequalities with applicationds in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14:326–334, 1965.

[3] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Contols, Signals, and Systems*, 2:303–314, 1989.

[4] B. V. Dasarathy. *Decision Fusion.* IEEE Computer Society Press, Los Alamitos, California, 1994.

[5] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[6] R. Dudley. *École d'Été de Probabilités de St. Flour 1982*, volume 1097 of *Lecture Notes in Mathematics*, chapter A Course on Empirical Processes, pages 2–142. Springer-Verlag, New York, 1984.

[7] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[8] V. Kurkova. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5:501–506, 1992.

[9] L. Ljung. *System Identification.* Prentice Hall, Engelwood Cliffs, NJ, 1987.

[10] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41(3):677–687, 1995.

[11] D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, New York, 1984.

[12] N. S. V. Rao. Fusion methods for multiple sensor systems with unknown error densities. *Journal of Franklin Institute*, 331B(5):509–530, 1994.

[13] N. S. V. Rao. Fusion rule estimation in multiple sensor systems using training. In H. Bunke, T. Kanade, and H. Noltemeier, editors, *Modelling and Planning for Sensor Based Intelligent Robot Systems*, pages 179–190. World Scientific Pub., 1995.

[14] N. S. V. Rao. Distributed decision fusion using empirical estimation. *IEEE Transactions on Aerospace and Electronic Sysetms*, 1996. to appear.

[15] N. S. V. Rao. Fusion methods in multiple sensor systems using feedforward neural networks. *Intelligent Automation and Soft Computing*, 1996. submitted.

[16] N. S. V. Rao. Nadaraya-Watson estimator for sensor fusion. *Optical Engineering*, 36(3), 1997.

[17] N. S. V. Rao and S. S. Iyengar. Distributed decision fusion under unknown distributions. *Optical Engineering*, 35(3):617–624, 1996.

[18] N. S. V. Rao and E. M. Oblow. Majority and location-based fusers for PAC concept learners. *IEEE Trans. on Syst., Man and Cybernetics*, 24(5):i713–727, 1994.

[19] N. S. V. Rao and E. M. Oblow. N-learners problem: System of PAC learners. In *Computational Learning Theory and Natural Learning Systems, Vol IV: Making Learning Practical*, pages 189–210. MIT Press, 1997.

[20] N. S. V. Rao, E. M. Oblow, C. W. Glover, and G. E. Liepins. N-learners problem: Fusion of concepts. *IEEE Transactions on Systems, Man and Cybernetics*, 24(2):319–327, 1994.

[21] N. S. V. Rao, V. Protopopescu, R. C. Mann, E. M. Oblow, and S. S. Iyengar. Learning algorithms for feedforward networks based on finite samples. *IEEE Trans. on Neural Networks*, 7(4):926–940, 1996.

[22] V. Roychowdhury, K. Siu, and A. Orlitsky, editors. *Theoretical Advances in Neural Computation and Learning.* Kluwer Academmic Pub., 1994.

[23] J. Shawe-Taylor. Sample sizes for sigmoidal neural networks. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 258–264, 1995.

[24] J. Sima. Back-propagation is not efficient. *Neural Networks*, 9(6):1017–1023, 1996.

[25] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[26] V. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York, 1982.

[27] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

[28] P. K. Varshney. *Distributed Detection and Data Fusion.* Springer-Verlag, 1996.

[29] S. A. Vavasis. *Nonlinear Optimization.* Oxford University Press, New York, 1991.