Revista de Matemática: Teoría y Aplicaciones 2011 **18**(2) : 299–324
cimpa – ucr issn: 1409-2433

SENSOR FUSION USING ENTROPIC MEASURES OF DEPENDENCE

FUSIÓN SENSORIAL USANDO MEDIDAS ENTRÓPICAS DE DEPENDENCIA

Paul B. Deignan*

Received: 23 Feb 2010; Revised: 23 May 2011;
Accepted: 25 May 2011

---

*L-3 Communications\Integrated Systems, PO Box 6056 Greenville, TX 75403-6056, U.S.A. E-Mail: paul.b.deignan@l-3com.com

Abstract

As opposed to standard methods of association which rely on measures of central dispersion, entropic measures quantify multi-valued relations. This distinction is especially important when high fidelity models of the sensed phenomena do not exist. The properties of entropic measures are shown to fit within the Bayesian framework of hierarchical sensor fusion. A method of estimating probabilistic structure for categorical and continuous valued measurements that is unbiased for finite data collections is presented. Additionally, a branch and bound method for optimal sensor suite selection suitable for either target refinement or anomaly detection is described. Finally, the methodology is applied against a known data set used in a standard data mining competition that features both sparse categorical and continuous valued descriptors of a target. Excellent quantitative and computational results against this data set support the conclusion that the proposed methodology is promising for general purpose low level data fusion.

Keywords: Information theory; data association; fusion; estimation; entropy.

Resumen

Contrario a los métodos estándar de asociación que ligan medidas de dispersión central, las medidas de entropía cuantifican relaciones multivaluadas. Esta distinción es especialmente importante cuando no existen modelos de alta fidelidad de los fenómenos detectados. Se muesrta que las propiedades de las medidas de entropía calzan en la marco Bayesiano de sensores jerárquicos de fusión. Se presenta un método de estimación de la estructura probabilística para medidas categóricas y continuas, el cual es insesgado para colecciones finitas de datos. Adicionalmente, se describe un método de ramificación y acotamiento de selección óptima del sensor apropiado tanto para refinamiento del objetivo como para detección de anomalías. Finalmente, la metodología es aplicada sobre un conjunto conocido de datos usados en una competencia est[andar de minería de datos, que caracteriza tanto descriptores ralos categóricos como continuos de un objetivo. Excelentes resultados cuantitativos y computacionales con estos datos apoyan la conclusión de que la metodología propuesta es promisoria para propósitos generales con datos bajos niveles de fusión.

Palabras clave: Teoría de la información; datos de asociación; fusión; estimación; entropía.

Mathematics Subject Classification: 94A17.

1 Introduction

A method is proposed of low-level synchronous data fusion capable of operating on mixed type measured data so that the information content of each measurement from each data source can be objectively quantified relative to a distribution of measurements of a target phenomenon. Fusion is the attempt to answer a human-centric question from multiple sources of data. The closest analogue to fusion is system identification. However, unlike system identification, fusion is not strictly a method for determining characteristics in the associations of data. Often there is no known system to characterize, but simply a stream of data that might or might not be relevant to a question at hand. For example, fusion for anomaly detection combines aspects of system identification, e.g. "What is normal?" with aspects of data mining, e.g. "How is this data self-consistent?" and the human-centric question, "What is abnormal?" Within the context of fusion, questions can be very specific, e.g. "Is an underground weapons production facility at this location?" or ill-posed so that fused information serves only to guide the questioner between what is presently knowable and unknowable. If the question is well-posed, fusion should also quantify the confidence of the answer.

There are two areas in which the knowledge-generation process is bottlenecked by technological shortcomings. One is in the *vertical* generation of knowledge from sensed measurements. The other is in the *horizontal* synthesis of data and information between machines and data stores. Taken as a whole, this semi-automated process of creating knowledge from sensors and data stores is called *information fusion*. When referring particularly to sensed data, it is called *sensor fusion* and when referring only to the generation of knowledge from data stores, it is called *data fusion*. The methodology introduced is in the context of operations on data stores but may be expanded towards the near real time processing of streaming data using conventional reservoir sampling techniques.

The questions answered by information fusion are application dependent. However, by the Data Processing Inequality (DPI) which states that data processing cannot increase the information in the data relevant to its source, the quality of answers and associated uncertainty are best when the fusion process itself is application independent. Therefore, the information-theoretic principal underlying an ideal fusion methodology application is that it should incorporate as much relevant information as possible while systematically distilling that information towards the answer of the most important potential questions. If the potential questions

are not know a priori, the fusion process should be constrained only as much as the application itself. As it may be guessed, the systemization of information fusion is still an open question to various degrees dependent on the application, potential questions, point to which the data is processed, and the particular definition of "information".

Data fusion differs from data mining primarily in the respect that there is a causal or contemporaneous relation in the descriptors of the data fusion process to the targeted phenomena. The present interest in fusion as a specialty distinct from data mining and system identification is owed primarily to the fact that there is a core constituency within the military remote sensing community working to automate information processing. The focus of the development of fusion as a special area of study has been heavily influenced by target tracking (e.g. [1-2]) and hierarchal methods to aid command and control (e.g. [3-7]). Target tracking estimates the geospatial position of entities over time. Additional information may then be associated with the entity estimates from which the identity of the entities and an estimate of the situation may be inferred. This conventional method of fusion relies on the detection and modeling of an entity in a sequential and potentially suboptimal association of data to a target or situation estimate. Note that the composition of functional estimates may be multi-valued so that function approximation methods of data mining (e.g. [8]) may be overly restrictive.

Data fusion presumes the existence of a common frame of reference for data aggregation. If we consider the sensed phenomenon to be a Riemannian manifold, then the frame of reference is topologically equivalent to a hypercube with dimensions corresponding to distinct data descriptors possibly over a common geospatial basis (Fig. 1). If the dimensions corresponding to a certain data descriptor are projected to the real line, it is possible that a one-one relation may not exist between the embedded manifold and its projection, i.e. there may not be a functional relationship between measurements of the phenomenon and the coordinates of the hypercube. For this reason inter-dimensional data association statistics derived from measures of central tendency are not always sufficient to quantify entity-related information in the measured probability space. However, entropic measures do quantify multi-valued relationships and when estimated as proposed here, also allow entropy estimates to be identified with *information*.

Fusion is essentially a problem in data compression where measurements are synthesized into information useful to a decision maker. Most

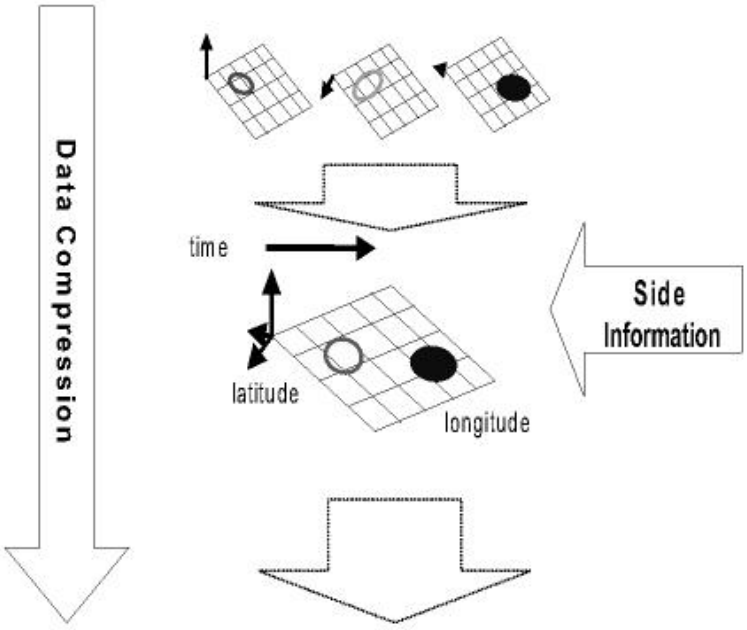

Figure 1: Initial levels of a sensor fusion architecture.

military fusion applications are formulated in terms of distances in time and space. Measurements of physical phenomenon such as imagery may be fused with non-physical data such as entity-type labels. This fused information may be then combined with side information such as known geographical features to yield a composite *picture* of a situation or event in a certain geographical region during a certain time period. Side information enters into the fusion process either as needed for the evaluation of a hypothesis or as might be necessary to elucidate features of the compressed data. For example, if the decision maker is concerned with the flow of materials across a geographical region, a map of the roads and waterways can be fused with the geospatial basis as side information. Additionally, if fusion resolves the data into a class of entities, side information relevant to the discrimination of class member may be incorporated. For example, if an entity is classified as a truck and the color of the truck is measured, the make and model of the truck by color is side information that may increase the confidence in the classification or, alternatively, to

further refine the classification by make and model. Side information is that information which is selectively applied a posteriori dependent on the hypothesis or the evolving information from the data fusion process. The process must be such that it presents an interface that admits the incorporation of side information dependent only on a priori knowledge of the data and the scope of knowledge discovery. Sequential, suboptimal, modeling fusion architectures do not generally admit a generic interface for side information incorporation.

In military parlance, the factors of consideration in the assessment of sources of measured data are *redundancy* and *mix*. Redundancy is the ensemble averaging of measurements within a common measurement space. By the central limit theorem, if the measurements are independent and identically distributed (iid) in a probability space, the estimate of the phenomenon should increase in confidence as a Gaussian distribution with decreasing variance. Redundancy in iid samples may be achieved with multiple measurements from one source over time during which the phenomenon is statistically stationary or by synchronous measurements of multiple independent data sources. Mix is the association of measurements between data dimensions. For example, an event can be measured in the blue or yellow visible light spectral dimensions of a multi-sepectral imaging sensor and those measurements associated with a common event through registration of the event measurements in a common geographical basis. The mix of the blue or yellow dimensions might then be associated as independent views of the green fused entity shown in Fig. 1. Mix is the property of the data fusion architecture of the strength of associations of data among independent dimensions.

Fusion, like mathematics, is founded in set theory. Also, like mathematics, we seek to unify and systematize the analysis. To this end we view dimensions as a set that dictate a relation among its members. Independent elements of a set are understood to exist in separate dimensions until a probabilistic relation is shown to exist between them. Probability is a method of associating beliefs with estimates (i.e. a method of making associations) while logic is a system of operating on sets with relations of probability one. Fusion then is nothing more than a hierarchy of probabilistic data association in which dimensionality is systematically reduced by projections and transformations as the probabilistic structure in the data is discovered. Mix is the process of making associations between dimensions while redundancy is a process in which associations are made within dimensions.

Interest in information-theoretic techniques of fusion has markedly increased over the past several years. Hero et al. [9] use Fisher's information measure as an optimization criterion for the scheduling of sensors in a sensor management algorithm. Hero notes that since the information-theoretic measures are model-independent and otherwise general, they decouple the risk\reward optimization from the collection task in the sensor management algorithm design. Hero cautions that the use of Fisher information is only justifiable when the underlying posterior distribution is smooth. Varshney [10] similarly advocates information-theoretic measures as general cost functions for sensor management algorithms. Varshney frames the problem of distributed sensor detection as a communication channel between sensor readings and the binary detection condition. The objective function is expressed as the Shannon's mutual information of the channel. Mahler [11] mentions central entropy and cross-entropy as a measure of statistical dispersion, but relies primarily on models for a radar-centric sensor fusion strategy of developing *hard* and *soft* mixture model clusters. Hard clusters are separable whereas soft clusters are not. On the other hand, Schuck et al. [12] go beyond the problem of detecting and locating the existence of a target and concentrate on its identification. In this context, a single sensor such as radar may be used to measure simultaneously several independent attributes of a target. Schuck therefore relies to a greater extent than Mahler in the use of information-theoretic measures for hypothesis discrimination since the identification problem is not only soft, but also multi-relational. However, it should be noted that Schuck limits himself to one-dimensional entropy measurements and considers the contributions of independent data descriptors only after the assignment of targets probabilities are made based on that single data descriptor alone. Thus, Schuck does not fuse measurements, but rather fuses hypotheses using information-theoretic measures.

The methodology for low level sensor fusion presented in this paper is fundamentally distinct from the aforementioned works in that estimates of entropic measures are formed directly from measurements rather than through a priori probability distributions or models of processes. The wider scope of the methodology necessitates the solution of several unique problems. First, the selection of the measure of information must be shown to be appropriate to the problem. We show by first principles that Shannon's mutual information is theoretically sufficient and computationally practical to adopt as the measure of multi-valued dependence for data associations. Second, the estimation of the measure must be unambigu-

ous. Here we advance the state-of-the-art by giving in closed form an evaluation of the maximal entropic estimate from any size data set on a uniform partition from which knowledge of the optimal informational content of a data descriptor can be assessed. Third, we provide a tractable method of selecting an informational-optimal set of data descriptors to support the construction of a fused model for a statistically quantifiable target signature. Since data fusion is essentially the same as a large class of data mining problems, the proposed methodology should be useful in fields apart from fusion.

This paper is organized as follows: A brief background is given on information-theoretic measure estimation in Section 2. In the absence of a priori information, estimates of the measures are calculated over uniform partitions of the measurement space. However, since the bias of the measures is dependent on the distribution of the data descriptors as well as the partition cardinality, Section 3 presents a method for determining the information-optimal partition for a particular data descriptor dimension. In Section 4, a branch and bound algorithm for the selection of mutual information-optimal data descriptors is presented. Due to the fact that most all data fusion strategies are suboptimal and model-dependent, it is very difficult to directly assess the relative merits of one methodology in comparison to others. Therefore, in Section 5 the proposed methodology is applied to a competition data mining problem that serves as an analogue to the general data fusion problem of descriptor selection for target identification. We conclude with a brief summary.

2 Entropic measures

The purpose of this section is to make an argument from first principles for the adoption of Shannon's entropy and mutual information for the quantification of information. There are two other entropic measures that are often used: Rènyi and Fisher. Rènyi's entropy is a parametric generalization of Shannon's entropy while Fisher's entropy is a parameterization specific to a model of a probability distribution. Both measures have found use in information-theoretic fusion (see e.g. [13] and [14]), yet the use of both assumes a priori information. Since the target distribution is not necessarily fully embedded in the measurement space, we instead develop Shannon's entropic measures as sufficient statistics.

2.1 Entropy

The Shannon entropy of a continuous random variable, Y, with the probability density function p is

$$H(Y) = -\int p(y) \log p(y)\ dy$$

which may be estimated on an indexed, uniform partition of m subsets by

$$\widehat{H}(Y) = \log N - \frac{1}{N} \sum_{i=1}^{m} n_i \log n_i \tag{1}$$

where $N = \sum_{i=1}^{m} n_i$ and $n_i \in \mathbb{R}^+, N \in \mathbb{Z}^+$. Note that the estimation of the entropy of a continuous random variable is identical to the calculation of a discrete random variable having m values and where n_i/N may be either the relative frequency or the normalization by the integral of all values applied to the i^{th} categorical value.

Three properties are sufficient to define the discrete entropy functional as a measure of uncertainty [15]:

1. $\widehat{H}(p_1, \ldots, p_m)$ is defined and continuous $\forall\{p_1, \ldots, p_m | 0 \leq p_i \leq 1, \sum p_i = 1\}$.
2. $\widehat{H}(\frac{1}{N}, \ldots, \frac{1}{N}) < \widehat{H}(\frac{1}{N+1}, \ldots, \frac{1}{N+1})$.
3. $\widehat{H}(\frac{1}{N}, \ldots, \frac{1}{N}) = \frac{1}{N} \sum_{i=1}^{m} n_i \widehat{H}(\frac{1}{n_i}, \ldots, \frac{1}{n_i}) + \widehat{H}(\frac{n_1}{N}, \ldots, \frac{n_m}{N})$.

The first property states that the functional is continuous over the domain of events. The second requires the functional to be monotonically increasing with increasing uncertainty, i.e. if a partition is refined, the uncertainty of an event occurring within this refined partition increases in proportion to the degree of the refinement. The third property enforces an alternative form of Bayes' Law in terms of uncertainty, i.e. the uncertainty associated with any particular event of a uniform partition is the uncertainty of that event in respect to others within the subset plus the uncertainty of that subset in respect to the set. Conditional uncertainty is composed by addition whereas probabilities are composed by multiplication in Bayes' Law. With this understanding, we can consider that the functional is Bayes-invariant to the structure of the partition of a fixed number of subsets. If the random variable is categorical or discreet, it is natural to think of the subsets of a partition as equivalence classes.

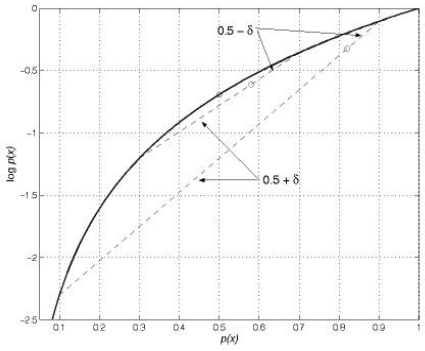

Figure 2: Geometry of the entropic functional.

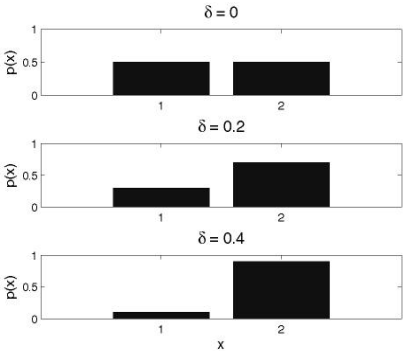

Figure 3: Proportional division of probabilities.

The logarithmic kernel of the entropy functional takes the multiplicative conditional probability composition property of Bayes' Law on the set of discrete class probabilities into an additive group of entropy through a group isomorphism from $(\mathbb{R}^+, \times)$ to $(\mathbb{R}, +)$. The transformation is in respect to the partition so that the equivalence relation (bin width for histogram probability estimates—a kernel in the statistical sense) is the implicit transform argument. Once the contributions of the measurement classes are transformed to an additive group structure, it is possible to decompose multiplicative products by addition through the distributivity of multiplication over addition and thereby associate all relative contributions of a total sum with individual component elements. This is nothing more than the superposition principle for the function of multiplication by a constant. So by (1), the contribution by the occurrence of events to the total uncertainty is apportioned linearly. Thus, the third property of uncertainty is satisfied by the group isomorphism property of the logarithm. Note that the base of the logarithm is not important. In base e the units are "nats".

Let $N = 1$ so that the negative value of the entropic functional is a simple linear combination of logarithmic kernels weighted by their respective arguments. Examine the logarithmic kernel from a geometric-entropic perspective (Fig. 2) for the discrete distributions of Fig. 3. The negative values of entropy are also the ordinates of the red circles dividing the respective dashed lines terminating on the logarithmic graph at points: $\{(.5 + \delta), \log(.5 + \delta)\}$ and $\{(.5 - \delta), \log(.5 - \delta)\}$ with segment lengths proportional to $(.5 - \delta)$ and $(.5 + \delta)$ respectively. This geometric repre-

sentation of the entropic functional is a consequence of the linearity of (1) and the group isomorphism property of the logarithm. Fig. 2 also shows that the greatest decrease in total entropy is achieved by the further resolution of events of the greatest relative certainty. It is important to note that a data fusion scheme operating by greedy entropy minimization to increase the certainty of probable events would also tend not to detect or refine ambiguous probabilistic structure. This flaw is prevalent in naïve optimization algorithms.

2.2 Mutual information

Shannon's mutual information relates the entropic content of sets of random variables to other sets through the intersection of common events in joint probability spaces thereby giving an information-theoretic quantification of data association. If restricted to finite probability spaces, normalized measures of mutual information satisfy all Rènyi postulates for measures of dependence [16]. The mutual information between random variables $I(X;Y)$ is a symmetric measure of probabilistic dependence, i.e.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

and is zero if and only if one variable is independent of the other, e.g., $H(Y) = H(Y|X)$ or $H(X) = H(X|Y)$ [17].

Mutual information can be estimated over partitions of X and Y as

$$\widehat{I}(X;Y) = \log N + \frac{1}{N} \sum_X \sum_Y n_{x,y} (\log n_{x,y} - \log n_x - \log n_y).$$

3 Estimation of entropic measures

The entropy of a set is dependent on the manner in which it is partitioned as well as the probabilistic structure of the distribution. Probabilistic structure is that structure of a distribution from which relationships may be inferred. For example, uniform distributions have minimal probabilistic structure whereas singletons have maximal probabilistic structure. The entropy of a distribution is a measure of probabilistic structure and hence also a measure of the bounds on the confidence in any relation that might be predicated on the probabilistic structure. The probabilistic structure of the discretized random variables may be unambiguously quantified once

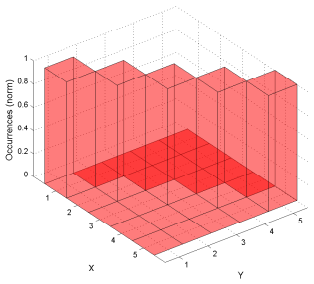

Figure 4: One-to-one relation: $H(X,Y) = H(X) = H(Y) = I(X;Y) = 1.61$ nats.

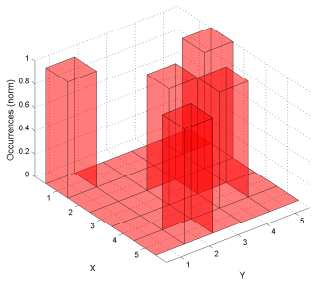

Figure 5: One-to-one relation: $H(X,Y) = H(X) = H(Y) = I(X;Y) = 1.61$ nats.

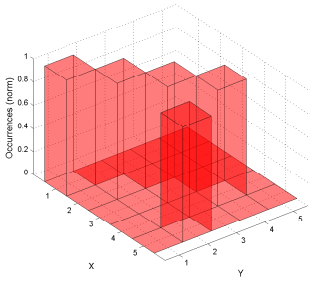

Figure 6: Multi-valued relation: $H(X,Y) = H(X) = 1.61$ nats; $H(Y) = I(X;Y) = 1.33$ nats.

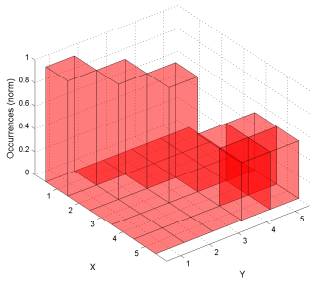

Figure 7: Multi-valued relation: $H(X,Y) = 1.89$ nats; $H(X) = H(Y) = 1.61$ nats; $I(X;Y) = 1.33$ nats.

the partition is established and the partition entropy is evaluated. Partition entropy is maximal if the partition is uniform. Therefore, since we are interested in the underlying probabilistic structure of a distribution and not on the scaffolding of its estimation, in the absence of a priori knowledge, a uniform partition between extremal values of the random variable will be adopted here for all entropic estimates.

The random variables returned by sensors as well as a priori environmental information may be either continuous valued or categorical. While the interpretation of the results of the dependency analysis of situational information is contingent on the meaning of the data being fused, the data itself can always be discretized and normalized into a probabilistic framework. Normalized variational structure is equivalent to probabilis-

tic structure and can be assessed consistently amongst dependent sensor dimensions once the zero probabilistic (variational) structure partition entropy of those dimensions is discounted. Since the contribution of partition entropy to the total uncertainty of the estimate is additive, the isolation of probabilistic structure by entropic measures is simply a matter of subtracting the partition entropy from the total uncertainty estimate. Note that by determining the partition of optimal probabilistic structure for each sensor dimension separately for a set of measurements it is implicitly assumed that the measurements are representative of a population of interest. When the probabilistic structure of a data descriptor is meaningful only in respect to another exogenous variable, the probabilistic structure of that variable is first determined and the mutual information with that variable becomes the entropic measure of interest.

3.1 Categorical valued random variables

Given an unlimited stream of categorical data, the choice of grouping of the classes for the entropic estimate is a compromise between resolution, timeliness, and computing resources. If data is finite, the choice is optimized by an exhaustive search through all combinations of categorical groupings for maximal probabilistic structure at some level of statistical significance apart from the estimated partition entropy of a uniform distribution of the same order as the partition. The entropic estimates of a uniform distribution are calculated by substitution of the probability estimates of a binomial distribution of $j \leq N$ occurrences of m classes,

$$\begin{aligned} p(j) &= f(j|N, \frac{1}{m}) \\ &= \left(\begin{array}{c} N \\ j \end{array} \right) m^{-j}(1-1/m)^{(N-j)}; (j = 0, 1, 2, \ldots, N) \end{aligned}$$

into the entropic (or mutual information) estimate,

$$\bar{H}_m^N = -m \sum_{j=1}^{N} \left(\frac{j \cdot p(j)}{N} \right) \log \left(\frac{j \cdot p(j)}{N} \right).$$

The value of the partition entropy of the estimated zero probabilistic structure offset may be pre-computed.

If the categories are unordered, the given classes together with the indeterminate class are taken as the partition. Unless the categories are known to be statistically independent from one another, combinations of

categories should be tested to determine the partition of maximal probabilistic structure. For example, if forming a map between X and Y that the region $\{(X,Y) : X \in [4,5], Y \in [4,5]\}$ should be consolidated into a single class.

The number of combinations of categories for each partition of order, m, is $\left(\begin{array}{c} N \\ m \end{array} \right)$ where N is the number of potential class labels. If an exhaustive evaluation is infeasible, categories may be suboptimally aggregated by Kullback-Leibler linkage lengths.

3.2 Continuous valued random variables

If the data is ordered and in the absence of a priori partition information, the probabilistic structure can be directly estimated as the uniform binning interval that maximizes the difference of the entropic estimate and its corresponding zero probabilistic structure partition entropy. A common method to approximate continuous-valued functions is by the enforcement of a regularity condition through the addition of a penalty function in an objective function. Using this approach, the optimal bin width is determined through

$$\arg\max_m \left(\bar{H}_m^N(x) - \widehat{H}_m^N(x) - \lambda G\right)$$

where G is the penalty function, m is the number of uniform bins of the entropic estimate (alternatively, the variable of interest in the mutual information estimate), and $\lambda \in \mathbb{R}^+$ is the penalty factor. If the data are uniform samples, then following [19], G may be chosen as a bias-corrected estimator of the finite difference approximation,

$$G = m^2 \left(\sum_{i=1}^{m-1} \left(\frac{n_{i+1} - n_i}{N} \right)^2 - \frac{2}{N} \right)$$

where n_i is the number of occurrences found in consecutively indexed bins. The penalty function is the sum of squares of the differences in estimated probabilities of adjacent bins so smooth functions are penalized with large values of λ. Since smoothness is estimated relative to process measurements, if the temporal extent of the optimization is much greater than the local neighborhood of sampling interval, then the range of optimal bin width should be readily discernible by large deviations in a unidirectional line search of $\bar{H}_m^N(x) - \widehat{H}_m^N(x) - \lambda G$ from $\lambda = (0, \bar{H}_m^N(x))$.

4 Data descriptor set selection

> There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. These are things we do not know we don't know. [20]
>
> *U.S. Secretary of Defense Donald Rumsfeld.*

To date, anomaly detection and target identification have largely been either treated separately or have been framed distinctly (see e.g. [21]). However, in this section we present a unified framework for the information-theoretic analysis of both aspects of data fusion. Under this framework uncertainty is quantified by class and managed by data descriptor dimension. The object of anomaly detection within this statistical context is to find significant probabilistic structure within the measurement space where none is expected. Conversely, the task of target identification is to associate the discovered probabilistic structure with that of known targets. In practice, the order of discovery is often reversed, i.e. that which cannot be matched with confidence to a known target signature model is categorized as a false alarm or anomaly.

The traditional methods of data fusion rely on "known" models where model mismatch is considered to be error rather than uncertainty. The great weakness of these naïve approaches is that they fail to assess uncertainty as a property of the fusion process independent of potentially erroneous target assignments and thus confuse what is knowable from what is unknowable. In this section we propose an information-theoretic data fusion framework where uncertainty is explicitly quantified by the data. In this framework targets are identified within a subspace of the anomaly detection measure space. So, when a target is found, it is possible to assess not only the confidence of the identification, but also the confidence of detection.

4.1 Anomaly detection (knowable unknowns)

In order to find probabilistic structure that is anomalous, a space of maximal probabilistic structure is found within the data space formed by descriptors above a specified level of statistical significance. That is, anomalies occur in data spaces X_Ω that maximize mix while minimizing

redundancy between data descriptors,

$$H_0 = \arg\max_{X_\Omega} \left(\bar{H}^N_{m_\Omega}(X_\Omega) - \widehat{H}^N_{m_\Omega}(X_\Omega) \right)$$

$\bar{H}^N_{m_\Omega}(X_\Omega)$ for any $X_\Omega \subset X_\alpha$ where X_α is the feasible set of data descriptors, is the zero probabilistic structure partition entropy estimate on the partition m_Ω inherited from X_α. Note that H_0 is monotonic with dimensionality,

$$H_0(X_1) \leq H_0(X_1, X_2) \leq H_0(X_1, X_2, X_3) \ldots, \ \forall (X_1, X_2, X_3, \ldots, X_n) \in X_\alpha.$$

Also, since

$$H_0(X_1, X_2, \ldots, X_n) \leq H_0(X_1, X_2, \ldots, X_{n-1}) + H_0(X_n)$$

it is possible to bound entropic estimates of a certain dimensionality over a given space without the need to compute the estimate over all combinations of dimensions.

4.2 Target identification (knowable knowns)

Given a subspace of maximal probabilistic structure X_Ω within the feasible measurement space, X_α the objective of target identification is to match geospatially localized regions of probabilistic structure X_{Ω_n} to known target signatures, Y. Spatiotemporal dimensions need not be within the optimal data space, X_Ω. However, it is necessary to be able to register the data space with the physical space in order that the identifications have relevance in a military context. Of course, nonmilitary applications of data fusion are not always constrained by the need to associate events with physical coordinates so we will proceed in the discussion with the understanding that target identification is augmented by geospatial side information as necessary.

With the preceding caveat, the information-theoretic optimal set of data descriptors is determined by the maximization of the bias corrected mutual information estimate between the optimal entropic set of data descriptors and the target, i.e.

$$I_0 = \arg\max_{X_\omega} \left(\bar{I}^N_{m_\Omega}(X_\omega; Y) - \widehat{I}^N_{m_\Omega}(X_\omega; Y) \right)$$

where $X_\omega \subseteq X_\Omega$. As before, X_ω inherits the partition structure of X_Ω. The effect of this optimization is to maximize redundancy while minimizing mix for any dimensionality less than $n = \dim(X_\Omega)$. The optimization

is a combinatorial problem in integer programming. Thus, the determination of the optimal subspace for target identification follows the same algorithmic rules as the determination of the optimal space for anomaly detection.

As in the case of entropy, the monotonicity of the estimated mutual information follows from the convexity of the natural logarithm over any interval on $\mathbb{R}^+$, i.e.,

$$I_0(X_1;Y) \leq I_0(X_1,X_2;Y) \leq I_0(X_1,X_2,X_3;Y) \leq \ldots$$

for any $\{X_1,X_2,X_3,\ldots,X_n\} \in X_\omega$. Also, since

$$I_0(X_1,X_2,\ldots,X_n;Y) \leq I_0(X_1,X_2,\ldots,X_{n-1};Y) + I_0(X_n;Y)$$

mutual information estimates within X_Ω may be conservatively bounded by the sum of pairwise estimates.

4.3 Statistical noise estimation (the unknowable)

Naturally, events that occur outside of the feasible measurement space, X_α, cannot be detected. Likewise targets which are not characterized in Y cannot be identified. Additionally, we are able to quantify the uncertainty within these spaces that is also unknowable. This is the zero information partition entropy, $\bar{H}^N_{m_\Omega}(X_\Omega)$ and mutual information, $\bar{H}^N_{m_\Omega}(X_\omega;Y)$. These values exist on the interval $[0,\log N]$ and are also dependent on the cardinality of the partition which in turn is chosen for maximal probabilistic structure of the corresponding random variable. Assuming additive Gaussian white noise uniformly distributed over the interval of measurement, the associated uncertainty is $H_\delta = m\log p_\delta$ where m is the cardinality of the partition and p_δ is the expected measurement error.

4.4 Entropic optimization by dimension algorithm

The globally optimal combination of entropic measures for either anomaly detection or target identification can be found following the branch and bound algorithm described in [22] with correction. The algorithm enumerates integer combinations via a spanning tree (Fig. 8), the branches of which are evaluated in order starting from the top node. At any node in the search of the spanning tree, subordinate branches are evaluated and ordered by decreasing entropic measure. If the bounding inequality is not satisfied for the current nodal maximal estimate, the algorithm backtracks

and selects the next unexplored subordinate branch. The critical determinate of the efficiency of the algorithm is the selection of an efficient bound. Since entropies correspond to probabilities on sets [23], to the first degree of intersection, the bound on a branch of dimensionality k in a problem of dimensionality n such that $(k < n)$ is conservatively given by

$$H_0(X^{(n)}) \leq H_0(X^{(k)}) + \max_{\alpha} \sum_{i=1}^{n-k} \left\{ H_0\left(X^{(k)}\right) - H_0\left(X^{(k)} \backslash X_{\alpha i}\right) \right\}$$

where $X^{(n)}$ is the data descriptor set at level n and α is the feasible set indices under consideration for inclusion to the $k+1$-dimensional set. This bound is computationally efficient as it does not require the calculation of combinations of entropies unless that portion of the branch has already been evaluated. Since the cardinality of the partitions increase exponentially with the dimensionality of the descriptor space while N remains fixed, a dimensional stopping criterion need not be calculated a priori, but rather can be conservatively estimated by an evaluation of the zero probabilistic structure partition entropy over the minimal product of feasible classes for the various dimensionalities. Therefore, while the spanning set might conceivably be n-dimensional, it is unlikely that high dimensional sets would be feasible. Note also that descriptors are not feasible if the optimal entropy is less than the specified measurement error entropy.

The corresponding bound for mutual information is

$$I_0(X^{(n)}; Y) \leq I_0(X^{(k)}; Y) + \max_{\Omega} \sum_{i=1}^{n-k} \left\{ I_0\left(X^{(k)}; Y\right) - I_0\left(X^{(k)} \backslash X_{\Omega i}; Y\right) \right\}.$$

This bound was found to be a computationally efficient compromise between the combinatorics of the local rejection of branches by the bound and the global increase in computation of interactions by dimensionality [24].

5 The 1998 KDD cup competition

It is often extremely difficult or impossible to assess the efficiency of a proposed data fusion technique due to the propensity for published results to be influenced by opaque and inaccessible incorporations of a priori information. Consider the problem of data fusion for target identification. An optimal descriptor set for target identification maximizes the useful

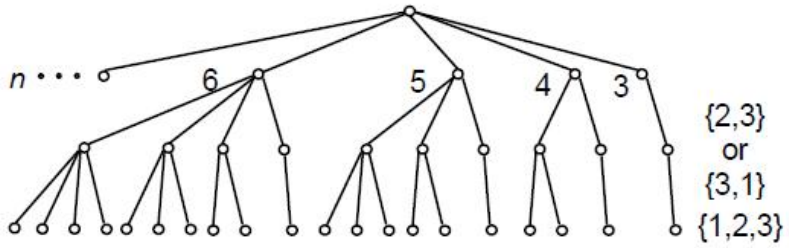

Figure 8: Branch and bound tree (the feasible set of descriptor indices are shown at the level of the estimate).

probabilistic structure of the measured data in an environment that both discriminates a target from the background clutter and distinguishes it as a target of interest. If the optimal target signature is unknown a priori, then the search for the optimal descriptor set is a problem in system identification. If the *ground truth* descriptors of a target are available along with sensor measurements, then the statistical problem of descriptor set selection for target identification is the same as that of data mining for descriptor fields of a target field.

The 1998 Knowledge Discovery and Data Mining (KDD) was chosen as a basis of comparison of the potential efficiency of the proposed use of entropic measures since the problem set closely approximates the data fusion challenges of mixed variables, sparsity of measurements, and ambiguity in the relation between descriptor data and characteristic target signatures. However, unlike the general data mining problem where multiple descriptor sets can be formulated dependent on the existence and local character of the data, operational fusion architectures are generally constrained to use a fixed set of descriptors since these descriptors often correspond to actual sensor measurements. Descriptor set selection was therefore constrained to only the determination of the globally optimal set for all data and thereby not allowing a possible modeling advantage by data segmentation.

The KDD competition is conducted annually and attracts the efforts of analysts in the areas of statistical data mining, sensor fusion, and system identification world-wide from both industry and academia. Other more recent KDD competitions focused on model learning, association rule dis-

covery, performance criteria, semantics, and relational analysis. The 1998 KDD competition attracted 57 participants, 21 of which returned results with slightly greater than half analyzed by production software tools with an average price of $27,761. The objective of the competition was to maximize the net receipts of a direct mailing to 96,367 individuals of a validation set given a $0.68 cost of solicitation. The total data set of 191,779 records was divided randomly into training and validation sets with 479 record descriptors and one target field consisting of the contributions received. There were no contributions less than the cost of solicitation so the target of association was simplified to be a binary value corresponding to whether or not the individual responded.

Since the signature of the target to be identified exists only within the same data set as the descriptors, the optimal detection space is taken as the entire feasible set. Also, since we are not concerned with anomaly detection in the comparison to competition results, this step of data fusion was omitted and the partitions for the descriptors were determined by optimal mutual information relative to the optimal target entropic partition. The composition of categorical classes within descriptors was determined by exhaustive combinatorial optimization of mutual information with the target up to and including cardinality four. For five and greater classes, composition of categorical classes was determined by the aggregation of Kullback-Leibler linkage lengths. The cutoff grouping was determined by inspection for purposes of the examination of the methodology. This is a shortcut that does not compromise the qualitative conclusions as the same method might be automated to give similarly good results at little computational cost.

An examination of the zero probabilistic structure over the entire record length reveals that for all data descriptors, there is no appreciable probabilistic structure for a number of partitions greater than the product of the first five primes. At a maximal 2310 partitions, the statistically significant mutual information at a measurement confidence of 99% yields measurement error entropy of 0.0020 nats under the assumption of additive Gaussian white noise. By this token, 50 data fields had significant probabilistic structure in respect to the target field. These descriptors formed the feasible set for target identification.

The model is taken to be the mean value of the target within each class of a hypercube formed by the model descriptor sets. The number of degrees of freedom of a partitioned hypercube is the number of hyperclasses. Therefore, while maximizing the statistical agreement between the infor-

mation of a descriptor set and the information of a target signature for a particular dataset, it is also necessary to balance the number of degrees of freedom of the model so that it does not overfit the data. Since the number of degrees of freedom is a product of the class order of each dimension, it is sensible to give priority to the minimization of the number of dimensions of the model rather than the minimization of classes for a particular descriptor set. The feasible dimensionality of the descriptor set is conservatively upper-bounded by n such that $\min_\Omega |classes_{opt}|^{(n)} \geq \max_\Omega |classes_{opt}|$. Here $n = 12$ since $\min_\Omega |classes_{opt}| = 2$ and $2^{12} \geq \prod_{i=1}^{M} p_i$, where p_i is a prime and M is determined to be five.

It would be extremely computationally expensive to calculate entropic measures on all combinations of fifty sensors up to and including dimensionality twelve. Therefore, the search for the optimal descriptor set for target identification is broken into two steps. In the first step, the branch and bound algorithm is exercised on all 50 feasible descriptor sets up to dimensionality four. Sets of dimensionality two are also calculated by the same algorithm. Next, the mutually exclusive optimal descriptor sets of dimensionality four are composed up to dimensionality twelve. These sets are further refined by the augmentation of mutually exclusive descriptors of dimensionality two and subsequently one about descriptor sets that demonstrate a potential for model improvement according a ratio of merit to degree of freedom. The results of these iterations are shown in Table 1.

The probabilistic structure of the target signature over the learning set, Profit'_{Lrn} , is the measureable value of the exogenous portion of the target data descriptor, i.e. the profit of a solicitation above the *ground clutter* or in this case, above the maximum of the *mail all* and *mail none* profit options. The Bayesian posterior sensor information is the product of the conditional and prior target signature estimates and may be normalized by the degrees of freedom to yield the portion of the target signature that might be reasonably found by a random sampling of the population.

Indices	$\frac{MI \cdot \text{Profit}'_{Lrn}}{C}$	$\frac{MI}{C}$	C	Profit_{Lrn}(\$)	Profit_{Val}(\$)
[10, 13, 48, 49]	0.1775	1.35×10^{004}	16	11,871	11,385
[4, 9, 10, 11, 13, 23, 48, 49]	0.0119	4.33×10^{006}	768	13,310	12,029
[2, 4, 9, 10, 11, 12, 13, 15, 18, 23, 48, 49]	0.0022	3.10×10^{007}	27,648	17,778	9542
[4, 9, 10, 11, 12, 13, 15, 23, 48, 49]	0.0042	1.30×10^{006}	3072	13,793	11,352
[2, 4, 9, 10, 11, 13, 18, 23, 48, 49]	0.0060	9.77×10^{007}	6912	16,664	10,471
[4, 9, 10, 13, 48, 49]	0.0585	3.76×10^{005}	64	12,113	11,826
[10, 11, 13, 23, 48, 49]	0.0349	1.49×10^{005}	192	12,896	12,318
[4, 10, 11, 13, 48, 49]	0.0585	3.69×10^{005}	64	12,141	11,805
[9, 10, 11, 13, 48, 49]	0.0552	3.69×10^{005}	64	12,052	11,753
[4, 10, 13, 23, 48, 49]	0.0351	1.50×10^{005}	192	12,901	12,348
[9, 10, 13, 23, 48, 49]	0.0351	1.49×10^{005}	192	12,914	11,894
[4, 10, 11, 13, 23, 48, 49]	0.0201	7.96×10^{006}	384	13,078	12,347
[9, 10, 11, 13, 23, 48, 49]	0.0203	7.99×10^{006}	384	13,102	11,985
[4, 9, 10, 13, 23, 48, 49]	0.0207	8.07×10^{006}	384	13,126	11,943

Table 1: Descriptor set selection.

Indices: Descriptor Label	Data Type	Unique Data Class Groups
4: PVASTATE	categorical	(1) (2,3)
10: HOMEOWNR	categorical	(1,3) (2)
13: CHILD12	categorical	(1) (2,3,4)
23: RFA_2	categorical	(1,6,7,9,13) (2) (3,4,12) (5,8) (10,14) (11)
48: RFA_2F	categorical	(1) (2,3,4)
49: RFA_2A	categorical	(1,4) (2,3)

Table 2: Selected model descriptors.

There is no claim that there should be a linear relationship between number of degrees of freedom and model fitness; rather the claim is that a meritorious model that does not overfit should have a greater criterion of fit than that of overfit models. This is indeed what is found by as shown in Table 1 where C is the number of classes and MI is the estimated mutual information between the target and descriptor fields. The first three entries of Table 1 are formed by assembling the most informative disjoint sets of four sensors clearly show that the set of dimensionality twelve overfits. Expanding on the set of dimensionality eight by the most informative partition of two of the last set of four also reveals overfitting in both cases. Expanding on the first set of four by all partitions of two sensors gives two characteristic combinations of 64 and 192 classes, neither of which significantly overfit the data. Of these two sets, the set highlighted more fully characterizes the target signature. Examining all combination of dimensionality seven of this set, it is clear that further expansion of the set overfits. Comparison with competitive results affirms that data descriptors [4, 10, 13, 23, 48, 49] (Table 2) is characteristic of the recoverable extent of the target signature as annotated by *miTool* in Fig. 9.

6 Conclusions

This paper introduces a principled data fusion methodology by entropic measures. Entropic measures are shown to quantify multi-relational dependencies, an ability which is essential for the consistent quantification

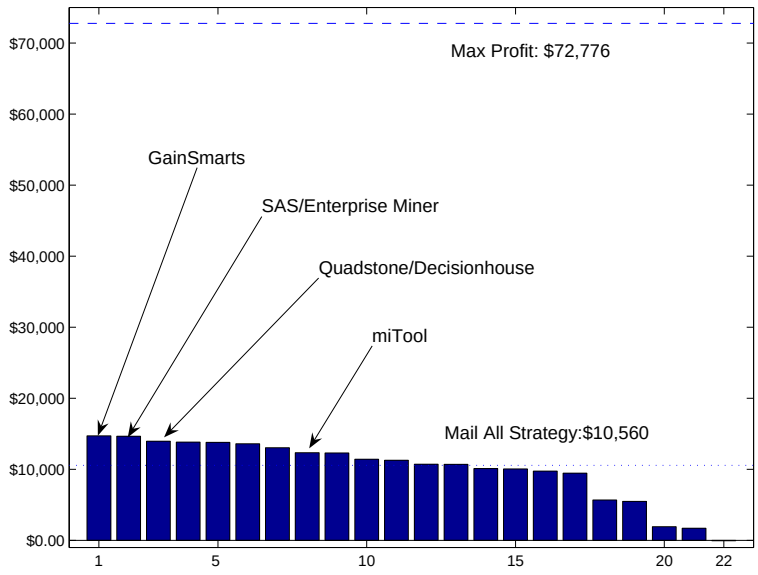

Figure 9: 1998 KDD Cup results.

of information throughout the various levels of a fusion architecture. The partitioning of the entropic estimates is discussed in detail and optimal partitioning algorithms are presented for both continuous and categorical cases. Finally, a branch and bound method of optimal data descriptor set selection is given and a demonstration on a standard competition database shows that the information-theoretic technique shows good results for target identification.

References

[1] Bar-Shalom, Y.; Li, X.; Eason, R.; Kirubarajan, T. (2001) *Estimation with Applications to Tracking and Navigation*. Wiley, New York.

[2] Blackman, S.; Popoli, R. (1999) *Design and Analysis of Modern Tracking Systems*. Artech House, Boston.

[3] Bossé, É.; Roy, J.; Wark, S. (2007) *Concepts, Models, and Tools for Information Fusion*. Artech House, Boston.

[4] Klein, L.A. (2004) *Sensor and Data Fusion: A Tool for Information Assessment and Decision Making*. SPIE Press, Bellingham, WA.

[5] Hall, D.L.; McMullen, S.A.H. (2004) *Mathematical Techniques in Multisensor Fusion, 2 ed.* Artech House, Boston.

[6] Antony, R.T. (1995) *Principles of Data Fusion Automation*. Artech House, Boston.

[7] Liggins, M.E.; Hall, D.L.; Llinas, J. (2009) *Handbook of Multisensor Data Fusion: Theory and Practice, 2 ed.* CRC Press, New York.

[8] Hastie, T.; Tibhirani, R.; Friedman, J. (2009) *The Elements of Statistical Learning, 2 ed.* Springer, New York.

[9] Hero, A.O.; Kreucher, C.M.; Blatt, D. (2008) "Information theoretic approaches to sensor management", in: Hero, Castanon, Cochran & Kastella (Eds.) *Foundations and Applications of Sensor Management*, Springer, New York: 33–57.

[10] Varshney, P.K. (1997) *Distributed Detection and Data Fusion*. Springer, New York.

[11] Mahler, R.P.S. (2007) *Statistical Multisource-Multi-Target Information Fusion*. Artech House, Boston.

[12] Schuck, T.M.; Hunter, B.; Wilson, D.D. (2009) "Developing information fusion methods for combat identification", in: M.E. Liggins, D.L. Hall & J. Llinas (Eds.) *Handbook of Multisensory Data Fusion: Theory and Practice*, 2 ed. CRC Press, New York.

[13] Kreucher, C.; Kastella, K.; Hero, A.O. (2005) "Sensor management using an active sensing approach", *Sig. Proc.* **85**(3): 607–624.

[14] Aughenbaugh, J.M.; LaCour, B.R. (2008) "Metric selection for information theoretic sensor management", *11th International Conference on Information Fusion*.

[15] Roman, S. (1992) *Coding and Information Theory*. Springer, New York.

[16] Bell, C.B. (1962) "Mutual information and maximal correlation as measures of dependence", *Ann. Math. Stat.* **33**: 587–595.

[17] Cover, T.M.; Thomas, J.A. (1991) *Elements of Information Theory.* Wiley, New York.

[18] Hall, P.; Morton, S.C. (1993) "On the estimation of entropy", *Ann. Inst. of Stat. Math.* **45**(1): 69–88.

[19] Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

[20] Defense.gov News Transcript: DoD News Briefing - Secretary Rumsfeld and Gen. Myers, United States Department of Defense (defense.gov), February 12, 2002.

[21] Simonin, C.; LeCadre, J.; Dambreville, F. (2007) "The cross-entropy method for solving a variety of hierarchial search problems", *10th International Conference on Information Fusion.*

[22] Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition.* Academic Press, London.

[23] Yeung, R.W. (1991) "A new outlook on Shannon's information measures", *IEEE Trans. Info. Theory* **37**(3): 466–474.

[24] Deignan, P.B.; Franchek, M.A.; Meckl, P.H. (2002) "Efficient information-theoretic model input selection", *45th Midwest Symp. Circuits and Systems* **1**: I-635–8.