SCIMA 2003 - International Workshop
Soft Computing Techniques in Instrumentation, Measurement and Related Applications
Brigham Young University, Provo, Utah, USA, 17 May 2003

Objective Evaluation of Subjective Decisions

Mel Siegel[(contact author)1], Huadong Wu[2]
Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213 USA, phone: +1 412 268 8742
e-mail: {[1]mws, [2]whd}@cmu.edu

Abstract *- The Dempster-Shafer "theory of evidence" encompasses and extends the Bayes Theorem-based decision making machinery. Dempster-Shafer's innovation is the introduction of lower and upper bounds, designated "belief" and "plausibility", that are attached to probability estimates. The Dempster-Shafer algebra provides for propagation of and reasoning about these quantities according to an algebra whose outcome phenomenologically mimics human decision making in many contexts that are laden with quantitative uncertainty. The approach's decisions thus seem to be subjective, i.e., the product of a sentient mind, vs. objective, i.e., the mechanical outcome of an immutable algorithm. In this paper we address the "objective evaluation of subjective decisions" in particular with the Dempster-Shafer sort of "subjective" decision making algorithm in mind. As an initial baseline approach, we examine the "receiver operating characteristic" (ROC) graph. We regard this as a first step towards identifying in advance circumstances under which Dempster-Shafer-like approaches should and should not be expected to deliver results that pass the human "sanity test".*

Keywords - decision theory, Dempster-Shafer theory, context aware computing, human-computer interaction, sensor fusion, receiver operating characteristic

1. INTRODUCTION

While superficially the title phrase "objective evaluation of subjective decisions" appears to be oxymoronic, on reflection it is not. People make subjective decisions all the time: based on my subjective assessment of traffic density, visibility, my own and the drivers' states-of-mind, etc., I decide that at this moment it is safe for me to cross the street against the red light. Whether or not my subjective decisions in the field of jaywalking are good or bad can be evaluated objectively: if I do it frequently and I am never injured or arrested, it is objectively obvious that my subjective decision making algorithm is indeed effective.

The everyday-language definition of *objective* in this context is "uninfluenced by emotions or personal prejudices; based on observable phenomena; presented factually". The term is also used in formal medical diagnosis with the meaning "based on a symptom or condition perceived as a sign of disease by someone other than the person affected".

In comparison, the everyday-language definition of *subjective* is "proceeding from or taking place in a person's mind rather than the external world; particular to a given person", and the corresponding medical diagnosis is "designating a symptom or condition perceived by the patient and not by the examiner".

Extending these definitions to the computer science domain, it is natural to associate conventional computer decision-making algorithms with the term *objective*, and "soft computing" methodologies with attempts to synthesize *subjective* behavior via algorithms that exhibit human-like interpretation of perceptual data. We are particularly interested in learning how objectively to evaluate the performance of algorithms that exhibit – or that appear to exhibit – subjective decision-making behavior.

In this paper we adopt, as a baseline approach, the "receiver operating characteristic" (ROC) graph. The ROC curve was originally developed and applied in the target identification field; it is now best known in the medical community, where it is used for characterizing and understanding the utility diagnostic tests. It basically analyses the dependence of the numbers or fractions of {true positives, false positives, true negatives, false negatives} yielded by a particular test as a function of the decision threshold or "cut-point".

The setting for the experimental instantiation of our work is "context sensing for context-aware computing", particularly the sensor fusion challenges that arise in its pursuit. The goal of context aware computing – which, as a practical matter, is synonymous with "context aware human computer interaction" – is for computers subjectively to understand environmental context, and via this understanding, better to interpret noisy and ambiguous inputs. The noisy and ambiguous inputs received from humans communicating via human-human interaction modalities, e.g., explicit and implicit gestures, are of particular interest.

The challenges to this program include the requirements for adaptability to a sensor suite that changes continuously, e.g., due to drift, and abruptly from time-to-time, e.g., due to failure or substitution, sensor system performance that is quantitatively commensurate with human perception, and artificial

sensing modalities that map qualitatively into the human senses and perception mechanisms.

We have written previously [1][2][3] on the applicability of the Dempster-Shafer "theory of evidence" to the sensor fusion aspects of this problem within the decision-making architecture illustrated in Figure 1. The attraction of the Dempster-Shafer approach is primarily its built-in uncertainty management and inference mechanisms, which exhibit behaviors reminiscent of human "subjective" reasoning processes. The Dempster-Shafer approach was shown in [1] to be practical and effective for implementing a general sensor fusion system architecture that was applied therein explicitly to the fusion of video and audio sensors that separately and jointly find and track meeting participants' focus-of-attention.

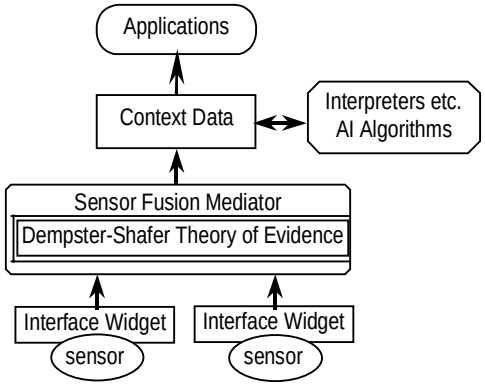

Figure 1. System architecture for sensor fusion of context-aware computing.

The conclusion of [2] was that, compared with previous *ad hoc* sensor fusion methods used in the focus-of-attention experiments, the Dempster-Shafer approach yields a definite but quantitatively only marginal improvement in accuracy. This was furthermore true again when the classical Dempster Shafer approach was extended to incorporate weights representative of sensor precision. However in both the classical and the extended case the Dempster-Shafer approach yields a significant improvement in robustness, e.g., against data packet loss or catastrophic sensor failure.

The conclusion of [3] is that accuracy and robustness are still further, but once again only marginally, improved via a further extended Dempster-Shafer approach that incorporates a memory mechanism – similar to Kalman filtering – to compensate for sensor drift and long-time-constant environmental variations.

The aim of the research program contemplated in this paper is to use existing and, as necessary, to develop new objective measures of decision system performance, especially in instances where the decision systems involve subjective reasoning mechanisms. At least two important application classes are foreseen:

(1) Confident resolution of the uncertainty of the reality of small apparent performance improvements as sensor fusion algorithms are made "more sophisticated". This is important because if apparent improvements can be proven to be real, albeit small, it would support continuing to work in the direction of the particular added "sophistication".

(2) Optimization of currently-arbitrary threshold or "cut-point" parameters within the decision-making machinery. Whereas in principle this optimization could be done via a neural network-like iterative training process, the actual computational expense is too high for this to be practical, especially if continuous adaptive behavior is required.

The rest of the paper is organized as follows. Section 2 reviews Dempster-Shafer theory, beginning with its background in Bayesian inference and carrying through the classical Dempster-Shafer theory and the extensions to it that we have introduced: sensor precision-based weighting and the temporal evolution of the weights. Section 3 reviews the "receiver operating characteristic" (ROC) curve for evaluating and understanding, from a variety of perspectives, the utility of classification tests that depend on a threshold or "cut-point" for dichotomization. Section 4 briefly discusses the issues involved in optimizing decision-making algorithms; it's intent is to tie together the preceding two sections on the Dempster-Shafer "theory of evidence" and the ROC curve of decision utility. Section 5 is a brief concluding section.

2. REVIEW OF DEMPSTER-SHAFER THEORY

2.1. Background: Bayesian Inference

Although it is sometimes regarded by the statistical sub-communities in the social sciences and economics more as an article of religious faith than a scientific fact, for physical scientists and engineers, Bayesian inference is hardly mysterious. It is illustrated by a simple example in which the data are:

P(M) = probability that an employee is a manager
P(~M) = probability that s/he is not a manager = 1 – P(M)
p(A|M) = probability per year that a manager will have an accident
p(A|~M) = probability per year that a non-manager will have an accident

We are told there was an accident; we want to know

P(M|A) = probability that the accident victim was a manager

Bayes Theorem says, rather intuitively, that

P(M|A) = P(M)•p(A|M) / (P(M)•p(A|M)+P(~M)•p(A|~M))

2.2. Dempster-Shafer Theory of Evidence

The Dempster-Shafer decision theory [4][5][6], or "theory of evidence" can be viewed as a generalization of Bayesian statistical inference. Its new feature is that it allows distributing support for a proposition (e.g., "person in meeting is employee A") to the union of propositions that include it (e.g., "person in meeting is likely either employee A or employee B"). In a Dempster-Shafer reasoning system, all the mutually exclusive possibilities are enumerated in a "frame-of-discernment", denoted Θ. For example, if we know that there is a person in a meeting room, and we want to recognize whether s/he is employee A, employee B, or somebody else, then the "frame of discernment" is:

$$\Theta = \{A,\ B,\ \{A,\ B\},\ \{\text{somebody else}\}\}$$

meaning s/he is certainly one of A, B, either A or B, or somebody else, i.e., neither A nor B.

Now suppose we have sensors that contribute additional information. Each sensor S_i contributes its observation by assigning its beliefs over Θ. This assignment function is called the "probability mass function" of S_i, denoted m_i. So, according to sensor S_i's perception, the probability that "the detected person is A" is indicated by a "confidence interval" whose lower bound is a "belief" and whose upper bound is a "plausibility":

$$[belief_i(A),\ plausibility_i(A)]$$

$belief_i(A)$ is quantified by all pieces of evidence E_k that support the proposition, e.g., that the person is A:

$$belief_i(A) = \sum_{E_k \subseteq A} m_i(E_k)$$

$plausibility_i(A)$ is quantified by all pieces of evidence E_k that do not rule out the proposition:

$$plausibility_i(A) = 1 - \sum_{E_k \cap A = \phi} m_i(E_k)$$

These definitions are reasonably intuitive if we interpret "plausibility" to mean "what we would believe if all the missing data that could support the proposition actually does turn out to support it".

For each proposition in Θ, e.g., A, Dempster-Shafer theory gives a rule for combining sensor S_i's observation m_i and sensor S_j's observation m_j:

$$(m_i \oplus m_j)(A) = \frac{\sum_{E_k \cap E_{k'} = A} m_i(E_k)\, m_j(E_{k'})}{1 - \sum_{E_k \cap E_{k'} = \phi} m_i(E_k)\, m_j(E_{k'})}$$

Obviously this rule can be chained straightforwardly.

By associating "belief" with the lower end of a probability range and "plausibility" with its upper end, the Dempster-Shafer approach manages to capture some features characteristic of the human perception-reasoning process. In contrast, the Bayesian approach provides no mechanism for dealing quantitatively with the ranges of "belief" and "plausibility" that humans characteristically attach to their estimates of likelihood.

2.2.1. Adding realistic sensors

The fundamental Dempster-Shafer combination rule implies that we trust sensors S_i and S_j equally. Misplaced trust can produce counterintuitive outcomes, e.g., if two observers agree that there is an arbitrarily small possibility of X, but they agree on no other possibility, Dempster-Shafer will say X is the only possible conclusion. Nor is this scenario far-fetched, as in many Dempster-Shafer applications the frame-of-discernment, and the numerical values of "belief" and "plausibility", are essentially educated guesses supplied by human experts. The human tendency to hedge a bet by assigning a small probability to an unlikely alternative conclusion expands the overall frame-of-discernment. It thus becomes easy for two experts' individual frames-of-discernment to share only one outcome, albeit one that both experts think is unlikely. The result is a catastrophe: the Dempster-Shafer algorithm decides that the small area of agreement is the only possible conclusion. Despite the usually intuitive behavior of the Dempster-Shafer algorithm, in this sort of case its conclusion is counter-intuitive.

But in sensor-based systems we should be able to do better, e.g., by quantitatively invoking technical knowledge about each sensor's *expected* performance, ground-truth knowledge about each sensor's current *actual* performance, and *historical* knowledge about the evolution of their performance, e.g., as the sensors age.

This sort of differential trust can be accounted for by a simple modification to the Dempster-Shafer formula in which the observations m_i are weighted by trust factors w_i derived from the corresponding expectations. These expectations might be based on, e.g., the sensor manufacturer's specifications, calibration experiments, or histories that capture a data stream of occasional ground-truth observations of the corresponding sensor S_i's performance. The weighting process is expressed formally by inserting the weights w_i as factors that multiply the probability mass functions, i.e., the observations m_i:

$$(m_i \oplus m_j)(A) = \frac{\sum_{E_k \cap E_{k'} = A} [w_i m_i(E_k) \cdot w_j\, m_j(E_{k'})]}{1 - \sum_{E_k \cap E_{k'} = \phi} [w_i m_i(E_k) \cdot w_j\, m_j(E_{k'})]}$$

When the weight factors w_i are functions of time, the approach is reminiscent of Kalman filtering. A simple practical implementation is to define

$$w(t) = \Sigma_{n=0 \text{ to } \infty}\, c(t - n\Delta t)\, r^n$$
$$c(t - n\,\Delta t) = n\text{-th\_previous\_correct ? 1 : 0}$$
$$r = \{0 \ldots 1\}$$

where the *remnace* (our coined term) r, range 0. to 1., is a parameter that controls how rapidly past performance is discounted.

3. RECEIVER OPERATING CHARACTERISTIC (ROC)

The receiver operating characteristic (ROC) graph was originally developed in the field of military target analysis to characterize ratios like signal to signal-plus-noise as a function of discriminator threshold. ROC curves have since been adopted and further developed primarily in the medical diagnostic test community.

Consider a sensor *S* that examines a person and delivers a numerical output n_S, that we believe increases monotonically with a classification of interest, e.g., if n_S is above a threshold value T_S then there is a high probability that the person under observation is a computer scientist. Increasing the threshold T_S increases the probability that an individual identified as a computer scientist really is a computer scientist and decreases the probability that an individual identified as a computer scientist is not really a computer scientist, and vice versa.

A separate test, regarded as "ground truth" or a "gold standard", provides performance characterization of the test as a function of the discriminator threshold in terms of four fractions: TP, the number (or fraction) of true positives, i.e., computer scientists for whom $n_S > T_s$, FP, the number (or fraction) of false positives, i.e., non-computer scientists for whom $n_S > T_s$, TN, the number (or fraction) of true negatives, i.e., non-computer scientists for whom $n_S < T_s$, and FN, the number (or fraction) of false negatives, i.e., computer scientists for whom $n_S < T_s$. This is summarized in Table 2.

sensor/class	in class	not in class
$n_S > T_S$	true positives	false positives
$n_S < T_S$	false negatives	true negatives

Table 2. True/False Positives/Negatives

The *sensitivity* of the classification is defined as the ratio TP/(TP+FN); it is the ratio of members of the class correctly identified by the test to the actual members of the class. The *specificity* of the classification is defined as the ratio TN/(FP+TN); it is the ratio of non-members of the class correctly identified by the test to the actual non-members of the class. Many other ratios of various combinations of TP, TN, FP, and FN are defined and named in the medical literature – positive/negative predictive value, positive/negative likelihood ratio, etc.. Sensitivity and specificity are the only ones we need now.

The ROC curve is the plot of sensitivity vs. (1-specificity) as the discrimination threshold is scanned through the output range of the sensor. A test is *reliable* if there is at least one threshold value for which there are no false positives and no false negatives. If this holds for all non-zero thresholds then the test is described as *ideal* or *perfect*: the ROC curve is then made up of the left and top sides of the unit square. A *useless* test has an ROC curve that is the diagonal of the unit square. Useful real tests have ROC curves that fall in between these extremes. A commonly used quality-index is the area under the ROC curve: unit area is ideal, area 0.5 is useless. These possibilities are illustrated in Figure 2.

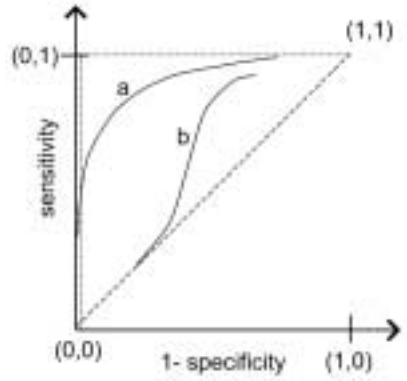

Figure 2: ROC curves. (a) *reliable*;
(b) typical; (dotted) *ideal*; (dashed) *useless*.

4. OPTIMIZING SUBJECTIVE DECISION-MAKING

As discussed in detail in [1],[2], and[3], a large set of focus-of-attention data have been collected in meetings of several participants. The data consist of video streams that have been analyzed by a gaze-direction algorithm, audio streams that have been analyzed by a voice origin algorithm, and *nominal* ground-truth that has been decided by human analysis of the video. The ground-truth is identified as *nominal* because it has a subjective component, e.g., two human observers do not always agree on which meeting participant is the focus-of-attention at any instant. The video and audio are objective at the raw data level, but must be considered subjective by the time they have been algorithmically abstracted to focus-of-attention.

Several sensor fusion algorithms have been used to combine the video and audio focus-of-attention reports: an ad hoc linear combination algorithm [1], a classical Dempster-Shafer algorithm [1], a Dempster-Shafer algorithm incorporating fixed sensor weights [2], and a Dempster-Shafer algorithm incorporating time-varying sensor weights that compensate for sensor drift [3]. The agreement between the output of the sensor fusion algorithm and the nominal ground truth shows a small but we believe real trend toward improvement with increasing sophistication of the Dempster-Shafer based sensor fusion algorithm.

However none of the algorithms has been formally optimized by variation of the decision-making parameters with a training data set. The discrimination parameters have rather been informally and arbitrarily set by the respective algorithm coders. ROC curve analysis will be undertaken to more precisely evaluate the relative performance of the different sensor fusion algorithms, with the intent of demonstrating conclusively whether or not the apparent improvement with increasing sensor fusion algorithm sophistication is indeed real and significant. Outcome of the anticipated analysis and optimization will be presented at the conference.

5. CONCLUSION

We have raised the question of "objective evaluation of subjective decisions". While at first-hearing the phrase seems to be self-contradictory, on thoughtful examination it is seen in fact to be possible, and recognized that in our everyday lives we do it all the time. The paper extends our consideration from the realm of everyday human decisions to the realm of decision-making algorithms that exhibit "subjective" behaviors.

We consider specifically the relative performance of a sequence of sensor fusion algorithms each of which combines in a qualitatively and quantitatively more sophisticated way the output of two "focus-of-attention sensors" in different perceptual modalities. Increased accuracy, i.e., agreement with ground truth generated by a human observer, seems to accompany increased sophistication, but the improvement is marginal, and perhaps of questionable statistical significance.

We consider improved tests to decide whether the apparent improvements are real, hence whether the corresponding evolutionary directions are worth pursuing. As a starting point, and as a baseline with which to compare future tests, we propose receiver operating characteristic (ROC) curve analysis. This approach will allow quantitative objective comparison of alternative sensor fusion algorithms. Furthermore, inasmuch as the ROC curve is a parametric plot in which the parameter is the tests cut-point, ROC curve analysis may provide a systematic mechanism for optimizing the currently arbitrary thresholds incorporated in the individual sensor output algorithms.

REFERENCES

[1] H. Wu, M. Siegel, and S. Ablay, "Sensor Fusion for Context Understanding", presented at IEEE International Measurement Technology Conference (IMTC'2002), Anchorage AK USA, 2002 May.

[2] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor Fusion Using Dempster-Shafer Theory", presented at IEEE International Measurement Technology Conference (IMTC'2002), Anchorage AK USA, 2002 May.

[3] H. Wu, M. Siegel, and S. Ablay, "Sensor Fusion Using Dempster-Shafer Theory II: Static Weighting and Kalman Filter-like Dynamic Weighting", to be presented at IEEE International Measurement Conference (IMTC'2003), Vail CO USA, 2003 May.

[4] Lawrence A. Klein, "Sensor and Data Fusion Concepts and Applications" (second edition), SPIE Optical Engineering Press, 1999, ISBN 0-8194-3231-8.

[5] Glenn Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[6] *Advances in the Dempster-Shafer Theory of Evidence*, edited by Ronald R. Yager, Janusz Kacprzyk, and Mario Fedrizzi. Wiley, 1993.

[7] S. Vida, AccuROC for Windows (Manual), 2001, Accumetric Corporation, http://www.accumetric.com.

[8] L. van Schalkwyk and J. van Schalkwyk, "The Magnificent ROC (Receiver Operating Characteristic curve)", http://www.anaesthetist.com/mnm/stats/roc/.

[9] K. H. Zou, "Receiver Operating Characteristic (ROC) Lit. Res.", http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html.