IEEE Trans. Fuzzy Systems. **8**, 796-799.

Fuzzy Analysis of Statistical Evidence

Yuan Yan Chen
Center for Army Analysis
Fort Belvoir, Va 22060-5230, U.S.A.

ABSTRACT

Fuzzy Analysis of Statistical Evidence (FASE) is utilizing the fuzzy set and the statistical theory for solving problems of pattern recognition and/or classification. Several features of FASE are similar to the human judgment. It can evaluate the weight of statistical evidence (information); it can update inference with new information; and it can incorporate missing data. Furthermore, since it can extract expert rules from data, it can also serve as a link to machine learning and expert systems.

1. Introduction

In this paper we develop a classification method using the fuzzy operators to aggregate attribute information (evidence), and it is named Fuzzy Analysis of Statistical Evidence (FASE). This method is closely related to the Bayesian classifiers. However, the classification is represented by the possibility and belief measures. By employing the possibility measure there is no need for the consideration of prior and there is more mathematical tools can be used. In machine learning when we infer from training sample to population, this is a process of inductive inference. As pointed out by Chen [3] it is more suitable to measure the inductive belief by the possibility measure than the probability measure. If two or more rival hypotheses (classes) are equally likely, then we do not have confidence or belief, which is true. Thus, belief is always a relative measure.

Machine learning algorithms are aiming for higher precision and faster computation. However, with the inconsistency of the data evidence, insufficient information provided by the attributes, and the fuzziness of the class boundary. We do not expect machine learning algorithm or even human expert to make the correct classification all the time. So it is important to have an uncertainty measure to represent our ignorance.

2. FASE Methodology and its Properties

Let C be the class variable and $A_1,\ldots, A_n$ be the attributes variables; and let Pos be the possibility measures. Based on the statistical inference developed in [3] we have

$$\text{Pos}\,(C \mid A_1,\ldots, A_n) = \Pr(A_1,\ldots, A_n \mid C) / \sup_C \Pr(A_1,\ldots, A_n \mid C), \qquad (1)$$

if the prior belief is uninformative.

Pos ($C \mid A_1,\ldots, A_n$) can be interpreted as the fuzzy membership that an instance belong to class C, and Bel ($C \mid A_1,\ldots, A_n$) = 1- Pos ($\bar{C} \mid A_1,\ldots, A_n$) is the belief measure or certainty factor (CF) that an instance belong to class C. The difference of (1) and the Bayes formula is simply the difference of normalization constant. In possibility measure the sup norm is 1, while in probability measure the additive norm (integration) is 1.

In machine learning, the number of attributes are usually very large, with limited number of training sample, the joint probability Pr ($A_1,\ldots, A_n \mid C$) can not be estimated directly from the data. This problem is similar to *the curse of dimensionality*. If estimate the conditional probability Pr ($A_i \mid C$) from each attribute separately, then we need a suitable operation to combine them together.

Next we give a definition of t-norm, which is often used for the conjunction of fuzzy sets.

Definition A fuzzy intersection/t-norm is a binary operation T: $[0,1] \times [0,1] \to [0,1]$, which is communicative, associative and satisfies the following conditions (cf. [4]).

(i) T (a, 1) = a, for all a.

(ii) T (a, b) ≤ T (c, d) whenever a ≤ c, b ≤ d. (2)

The following are examples of some t-norms that are frequently use in the literatures.

Minimum: M (a, b) = min (a, b)
Product: Π (a, b) = ab.
Bounded difference: W (a, b) = max (0, a + b -1).

And we have $W \le \Pi \le M$.

Based on different relationship of the attributes, we have different belief update rules. If A_1, A_2 are independent then we have (cf. Chen [2])

$$\text{Pos}(C \mid A_1, A_2) = \text{Pos}(C \mid A_1)\,\text{Pos}(C \mid A_2) / \sup_C \text{Pos}(C \mid A_1)\,\text{Pos}(C \mid A_2), \quad (3)$$

and if A_1, A_2 are completely dependent, i.e. Pr ($A_1 \mid A_2$) = 1 and Pr ($A_2 \mid A_2$) = 1, then we have

$$\text{Pos}(C \mid A_1, A_2) = \text{Pos}(C \mid A_1) \wedge \text{Pos}(C \mid A_2) / \sup_C \text{Pos}(C \mid A_1) \wedge \text{Pos}(C \mid A_2), \quad (4)$$

where $\wedge$ is a minim operation. This holds since Pos ($C \mid A_1, A_2$) = Pos ($C \mid A_1$) = Pos ($C \mid A_2$). Note that if A_1, A_2 are functions of each other, they are completely dependent; so the evidences are redundant.

In general the relations among the attributes are unknown, but, it seemed reasonable to employ a t-norm in between Π and M for belief update. For simplicity we restricted to the model that aggregate all attributes with a common t-norm ⊗ as follows

$$\text{Pos}(C \mid A_1,\ldots,A_n) = \otimes_{i=1,\ldots,n} \text{Pos}(C \mid A_i) \,/\sup_C \otimes_{i=1,\ldots,n} \text{Pos}(C \mid A_i). \quad (5)$$

If we choose $\otimes$ equal to the product Π, then (5) is equivalent to the *naïve Bayesian* classifier with uninformative prior.

As shown in [3] product rule implies adding the weights of evidence. If attributes are completely dependent by employing the product rule we are basically counting the same evidence twice.

The following are some characteristic properties of FASE.

(1) For any t-norm if attribute A_i is noninformative, i.e. Pos $(C = c_j \mid A_i) = 1$, $\forall j$, then

$$\text{Pos}\,(C \mid A_1,\ldots, A_n) = \text{Pos}\,(C \mid A_1,\ldots,A_{i-1}, A_{i+1}\, A_n). \quad (6)$$

This holds since T (a, 1) = a.

Equation (6) indicates that a noninformative attribute did not contribute any evidence for overall classification, and it happens when an instance a_i is missing or A_i is a constant. Similarly if A_i is a white noise then it provide little information for classification, since Pos $(C = c_j \mid A_i) \approx 1$, $\forall j$. Thus FASE is noise tolerant.

(2) For any t-norm if Pos $(C \mid A_i) = 0$ for some i, then

$$\text{Pos}\,(C \mid A_1,\ldots, A_n) = 0. \quad (7)$$

This holds since T (a, 0) = 0.

Equation (7) indicates that the process of belief update is by eliminating the less plausible classes/hypothesis, i.e. Pos $(C \mid A_i) \approx 0$, based on evidences. The ones that survive the process become truth.

(3) For any t-norm if Bel $(C = c_j \mid A_1) = a$, Bel $(C = c_k \mid A_2) = b$, $j \neq k$ and $b \leq a$, then

$$\text{Bel}\,(C = c_j \mid A_1, A_2) = (a - b) / (1 - b). \quad (8)$$

Since $(a - b) / (1 - b) \leq a$, equation (8) implies that if the evidences conflict, it will lower our confidence which class it belongs; however, the computation is the same no matter which t-norm is used.

The only situation where t-norm makes a difference is when we have Bel $(C = c_i \mid A_1) = a$, and Bel $(C = c_i \mid A_2) = b$, $0 < a, b \leq 1$. The t-norm will determine how much our confidence should increase.

Thus, if we employ different t-norms to combine attributes the computations are quite similar with each other. This also explains, even though the independence assumption of the *naïve Bayesian* classifier is very often violated, it still can perform well.

3. Computation of FASE

For continuous attributes we employ the kernel estimator for density estimation

$$p(x) = 1/nh \sum_i K\,((x - x_i)/h), \quad (9)$$

where K is chosen to be uniform for simplicity. For discrete attributes we use the maximum likelihood estimates. The estimated probabilities from each attribute are normalized into possibilities and then combined by a t-norm as in (5). We examine the following two families of t-norms, since these t-norms contain wide range of fuzzy operators. One is proposed by Frank [5] as follows

$$T_s(a, b) = \log_s (1+ (s^a - 1) (s^b - 1) / (s - 1)), \text{ for } 0 < s < \infty. \quad (10)$$

We have $T_s = M$, as $s \to 0$, $T_s = \Pi$, as $s \to 1$ and $T_s = W$, as $s \to \infty$.

The other is proposed by Schweizer & Sklar [8] as follows

$$T_p(a, b) = (\max (0, a^p + b^p - 1))^{1/p}, \text{ for } -\infty < p < \infty. \quad (11)$$

We have $T_p = M$, as $p \to -\infty$, $T_p = \Pi$, as $p \to 0$ and $T_p = W$, as $p \to 1$.

For binary classification FASE is equivalent to the likelihood ratio statistics. If we are interested in the disciminant power of each attribute, then Kullback's [7] information of *divergence* can be applied, which is given by

$$I(p_1, p_2) = \sum_x (p_1(x) - p_2(x)) \log (p_1(x)/p_2(x)). \quad (12)$$

FASE does not require consideration of the prior. However, if we multiply the prior, in term of possibility measure, to the likelihood, then it discounts the evidence of certain classes. So in a loose sense prior can also be considered as a kind of evidence.

4. Experimental Results

The data sets used in our experiments come from the UCI repository [1]. The computation is based on all records, disregarding it has missing values or not. A five-fold cross validation method [6] was used for perdition accuracy. We include those records with missing values in the training set since those non-missing values still provide information for model estimation. If an instance has missing values, which are assigned as null beliefs, its classification is based on lesser number of attributes. But, very often we do not need all the attributes to make the correct classification. Horse-colic data contains 30% missing values; it still can perform reasonably well.

Table1. Experimental results of the primary approaches discussed in this paper

Data set	t-norm parameter**		Π	M
1 australian	s =.75	85.0	84.7	81.8
2 breast*	s =.5	96.7	96.7	96.2
3 crx*	s = .1	85.5	84.9	83.9
4 DNA	s =.5	95.5	94.3	82.5
5 heart	s =.8	82.3	82.3	81.1
6 hepatitis*	p = -.1	85.4	85.3	84.7
7 horse-colic*	p = -3	80.7	79.0	80.2
8 inosphere	s =.7	88.5	88.5	83.8
9 iris	s =.5	93.3	93.3	93.3
10 soybean*	p = -1	90.1	89.8	87.7

11 waveform	s =.1	84.2	83.6	80.9
12 vote*	p = -8	94.9	90.3	95.2

*Data set with missing values.
** T-norm parameters that perform well for the data set. s- Frank parameter, p- Schweizer & Sklar parameter

T-norms weaker than the product are less interesting and do not perform as well, so we did not include them here. Min rule reflects the strongest evidence among the attributes. It does not perform well if we need to aggregate large number of independent attributes, such as the DNA data. However it performs the best if the attributes are strongly dependent on each other, such as the vote data.

Although in many situations FASE classifier did not show significant improvement over *naïve Bayesian* classifier, however, it can provide a better estimate for confidence measure. The confidence measures under *naïve Bayesian* classifier tend to be too close to 1 to be meaningful, since it over compensates the weight of evidences. Those confidence measures of FASE do provide useful information for classification. For example in the crx data, FASE classifier comes up to be about 85% accuracy. If we consider those instances with a higher confidence level, e.g. CF > .9, then we can achieve an accuracy over 95%.

5. Conclusion

In this paper we explore the classification problem from an evidential reasoning point of view. We only investigate a simple model of aggregating attributes information with a common t-norm. However, this approach might suffice in many situations, as shown from the experiments, a precise belief model for the attributes is not very important. The advantage of using the possibility measure over of the probability measure for classification is clearly demonstrated here. Since FASE is noise-tolerant and able to handle missing values with ease, it allows us to consider as many attributes as possible. This is important since many patterns become separable if we increase the dimensionality of data. The belief Bel (C | A) can be interpreted as " If A then C with certainty factor CF". Thus, by extracting the statistical pattern from the data and combine them with the fuzzy inference rule FASE can serve as a link to inductive reasoning (machine learning) and deductive reasoning (expert systems).

REFERENCES

1. Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases. [http://www.ics.uci.edu/~mlearn/MLRepository.html].
2. Chen, Y. Y. (1993). Bernoulli trials: from a fuzzy measure point of view. *J. Math. Anal. Appl.* **175**, 392-404.
3. Chen, Y. Y. (1995). Statistical Inference based on the Possibility and Belief Measures *Trans. Amer. Math. Soc*. **347**, 1855-1863.
4. Klir, G. J. and Yuan, B. (1995). *Fuzzy set and Fuzzy Logic: Theory and Application*. Prentice Hall.
5. Frank, M. J. (1979). On the simultaneous associativity of F (x, y) and x + y – F (x, y). *Aequationes Math*., **19,** 194-226.

6. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference for Artificial Intelligence.* Morgan Kaufmann, San Francisco. 1137-1143.
7. Kullback S. (1968). *Information Theory and Statistics*. Dover, New York.
8. Schweizer, B. and Sklar, A. (1963). Associative functions and abstract semi-groups. *Publ. Math. Debrecen* **10**, 69-81.