Classification Data Mining with Hybrid Fuzzy Logic Aggregation

JOHN F. SANFORD
Information Systems
Philadelphia University
Philadelphia, PA 19144-5497
USA
e-mail: SanfordJ@PhilaU.edu

LES M. SZTANDERA
Computer Information Systems
Philadelphia University
Philadelphia, PA 19144-5497
USA
e-mail: SztanderaL@PhilaU.edu

Abstract - Fuzzy logic is applied to the category discrimination problem related to identification of mammary lesions as benign or malignant. Results of other similar studies are reviewed. The current analysis expands the fuzzy logic approach by using the normal distribution function as set membership functions and using a genetic algorithm to optimize performance with the training partition. The approach is applicable to problems having arbitrarily large number of parameters. Two different data sets are examined. Data is portioned into a training set and validation set and each set is segregated into benign and malignant records. Values of mean and standard deviation are initially computed from the associated attributes and are different for the benign and malignant records. In one training method the standard deviations are adjusted to minimize overall error. In a second method a bias adjusts the importance of each membership function. Defuzzification is accomplished in three ways: modified averaging and OR process; comparison of multiplied fuzzy set values; and comparison of the multiplied squared set values. Results are compared with results obtained through statistical logistic regression.

Key Words: fuzzy data-analysis discrimination statistical analysis screening

1. Introduction

1.1 General

Since its introduction 1965 by Lotfi Zadeh [1] fuzzy logic has found extensive application in control systems from small appliances to cameras and even heavy equipment. Although control systems are its forte, not surprisingly it has found its way into most if not all areas of data mining as a seriously competing analytic tool. To this repertoire we would add numerous information search and utilization techniques that have spawned even a fuzzy database query language developed by Takihashi in 1995 [2].

Fuzzy logic shows itself to great advantage when it can simplify an otherwise extremely difficult algebraic formulation for a control system without apparent loss of effectiveness. In fact the fuzzy approach may be more effective. For example in the automatic control of a camera the input parameters are light intensity, distance to object, motion of the object relative to the camera, and capability of the recording media. Of these the only constant parameter is the capability of recording media. Other parameters may exhibit variations over a wide range. Formulation of a crisp (exact) control system is quite difficult. The fuzzy approach classifies the input parameters into fuzzy sets. For example, these sets might be “dim light”, “medium light”, “bright light”, “nearby”, “some distance”, and “remote distance”. The task then is to determine the percent of membership that the actual parameter has in each associated set. For example light may have 80% membership in “medium”, 10 % membership in “bright” and zero membership in “dim”. The setting of camera speed and aperture then is based on linguistic rules. One such rule might be, “If bright light and remote distance and zero relative motion then aperture setting is 16 and speed is 1/100. Another might be if medium light and remote distance and zero relative motion then aperture setting is 11and speed is 1/100

Linguistic rules are logic statements. They include intersection and union. Typically these are resolved through a set of rules as follows.
A AND B where A and B are limited to the range
(0, 1), becomes equivalent to min (A, B) and
A OR B, where A and B are limited to the range
(0, 1), becomes equivalent to max (A, B).

The capability of classifying patterns for subsequent decision making processes is one of the most fundamental characteristics of business intelligence and data mining. As a field of study, classification has been evolving since the early 1950s, closely following the emergence and evolution of computer technology and classification techniques. As previously mentioned fuzzy logic has been applied to this type of problem.

1.2 Other papers in the area.

There is a perfusion of articles on the topic of fuzzy data discrimination. We mention a few that deal with medical diagnosis which is also the direction of the current work.

Cios J K, et al. [3] reported on Fuzzy logic used for classification of coronary stenosis from planar thallium-201 scintigraphs. The object was diagnosis of stenosis which might be occurring in any of the three major arteries, left anterior descending (LAD), right coronary artery (RCA), the circumflex artery (CCX). The scintigraphy consisted of three views, left ventricle-anterior (ANT), left lateral (LAT), and anterior oblique (LAO). These data are themselves fuzzy with considerable overlap in perfusion patterns that might be created by stenosis. Some other approaches to solving this problem include: fuzzy sets with probability assigned membership functions described by Cios al in 1991[4], machine learning algorithms described by Cios al in 1994 [5] and Cios et al in 1994 [6], and an expert system approach described by Cios et al in 1990 [7].

Cios et al in 1994 [3] used a data set containing 64 patients 46 with stenosis and 18 without. They used trapezoidal membership functions for each of the three scintigraphy views. They evaluated three known learning algorithms, ALFS, CLILP2, and EXP. They formulated 13 different linguistic statements to provide crisp results. The notation used in formulating the if-then statements relates to the medical parameters described in the paper but one statement is presented here as a demonstration of defuzzification approach. “IF LAT4 has perfusion defect value in the range of [0,4], [15,34],[45,54],[85,1001] with (0.2) AND LAT8 has perfusion defect value in the range of [5,24],[35,44],[75,100] with (0.4) AND LAT5 has perfusion defect value in the range of [0,4] with (0.4) THEN STENOSIS is in CCX.

They reported on the results of the three different algorithms for defuzzification. Accuracy for classifying normal arteries ranged from 96% to 100%. However accuracy for correctly classifying stenosis ranged from 85% to 93% for two algorithms. A third had lower accuracy.

Malek J, et al [7] discussed an automated breast cancer diagnosis system utilizing fuzzy aggregation. They proposed a fuzzy instrument for hardware implementation of the automated system using a CMOS integrated circuit. They reported on active research conducted with 200 patients to prove the concept and finally proposed a concept design for the CMOS integrated circuit.

For the active research, cells were extracted from the patients’ lesions and fixed on microscope slides. For each patient 19 attributes are analyzed. The GVF-Snake model was used to identify the contour of a typical nucleus within the object. Texture-analysis-based feature extraction involved wavelet transforms. Statistical attributes included mean, variance, entropy, and normalized Shannon entropy. The process is mathematically intensive. The fuzzy clustering C-means algorithm as is available in MATLAB was use to assign fuzzy set memberships. Each new pattern was assigned a degree of membership of malignant class or benign class. The sum of all membership functions at any point equals 100%

They reported 97% correct classification of benign cases and 93% correct classification of malignant cases.

Bagher-Ebadian H, et al [8] in 2004 used an automated neural network and fuzzy clustering for diagnosis of coronary artery disease. They report two studies, one with 58 subjects and one with 115. Data was obtained from planar images in three projections, anterior, left anterior oblique, and left lateral using a collimator. Images were taken at rest and again under stress. Background noise was suppressed by segmenting myocardium from the background using the fuzzy clustering Picard iteration algorithm. Euclidean distance was used as the distance metric.

Each myocardium was partitioned into four sections and the center of activity found for each section. Sampling radii were taken every 2 degrees to project the sectors on an x-y plane. There were three transformed projections divided into 4 regions for “at rest” and the same for “stress”. Thus there were two vectors of dimension12. The mean and variance row vectors were defined as feature vectors. The number of features was reduced through a selection process based on maximum separation in multidimensional feature space

The artificial neural network was trained and then evaluated. The true values, normal or abnormal, were obtained from coronary angiographies’ taken within three months of the imaging. Accuracy in prediction positive results varied from 77% to 86% and prediction of negative results from 63% to 66%.

In 2010 Licata [10] presented a most interesting demonstration of fuzzy logic as applied to the more general diagnostic process. This is different from the previous citations in that it does not deal with data analysis to identify one of two possible outcomes as a “test” for a specific disease. Rather, he suggested that the application of fuzzy set theory in place of the physician’s usual reliance solely on probabilistic logic can improve the general diagnosis process which attempts to identify the cause of observed symptoms.

He developed the fuzzy set theory process through an example application to one case study. Specific symptoms in this case are: dyspnea at rest (a), oedema (b), tachyarrhythmias (c), epatomegally (d), ascites (e), pleural effusion (f).

Possible diagnoses would be hepatic cirrhosis (t), nephrosic syndrome (u), pneumonia (w), myocardial ischemia (x) congestive heart failure (y), worsening of the supraventricular arrhythmias (z).

A typical If-then rule would have the form, “If a and b and c and d and e and f then t or w or x or y or

Severity of these symptoms a, b, c, etc. are typically linguistic and fuzzy such as tachyarrhythmias is “slow”, “moderate”, or “fast”.

Numerous laboratory tests provide more data for assignment of fuzzy membership functions. Linguistic observations and laboratory results are assigned set memberships between 0 and 1 through the use of charts that have been prepared.

Defuzzification involves a considerable number of predicate logic statements. These are evaluated using generalized modus pones and sometimes more directly. This likely involves a good deal more skill in logic than most physicians would poses. Licata’s work points out that the fuzzy logic approach is in lieu of the probabilistic approach generally employed. Licata suggests that the fuzzy set theory solution is more exacting and probably superior to the usual process where individual decisions in the chain leading to a conclusion are made on the basis of probabilistic estimates. This paper is of course not able to present accuracy figures since there is only one case considered. It is recounted here as an indication of the pervasive influence of fuzzy logic on the medical discipline.

1.3 Different approach

The work we present in the current paper offers another variation. It employs the normal distribution as set membership functions. These functions are individually adjusted by using weight and varying each standard deviation. Optimization is obtained through application of a genetic algorithm. Linguistic rules are not employed and hence the fundamental process is applicable to data sets containing many parameters and to classification into more than two categories. The process is applied to several different data sets and results are compared with the more common statistical regression analysis.

2. Analytic approach

Step 1 of the process involves assigning data to memberships in the fuzzy sets.

2.1 Triangular membership function

Numerous membership functions are recognized for creating the fuzzy sets [11] These include triangle, rectangle, parabola, trapezoid, and even projections of one operating characteristic onto another plane. If, for example, an isosceles triangle is used then the height would generally be considered equal to 1 and the apex is located at the parameter's average value over the totality of past data. A particular data record will generally have a value for this parameter that is not at the average. Membership will be determined by the height of the line AC in Figure 1. This height has a simple relationship shown in equations 1 and 2

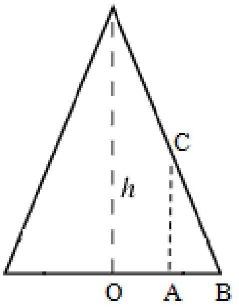

Figure 1

$$(1)\, AC = h \times \frac{AB}{OB} = 1 \times \frac{AB}{OB} = \frac{OB - OA}{OB} = 1 - \frac{OA}{OB}$$

$$\text{Membership} \quad \begin{cases} = 1 - \left|\frac{OA}{OB}\right| & \text{For } |OA| \le |OB| \\ = 0 & \text{Otherwise} \end{cases}$$

It should be noted that the value of O need not be an exact match to the average value of the parameter for all existing data records. In some cases, particularly in cases involving dynamic control systems, the value of O may be designated in accordance with engineering analysis or other considerations.

A comparison of performance with triangular membership function and the exponential function was reported in an earlier paper [14]. In this case some adjustment was made to triangular base and exponential width in an effort to find the better function. The exponential function performed better in that instance but triangular, exponential (or sine wave), and trapezoidal are all common.

2.2 Trapezoidal membership functions

Suppose that there are n generalized fuzzy numbers $A_1, A_2, \ldots, An$, with trapezoidal membership functions $A_i = (c_i, a_i, b_i, d_i, w_i)$. The trapezoidal membership function of a generalized fuzzy number, A_i, is then given by:

$$(3) \qquad \mu_{Ai}(x) = \begin{array}{l} a_i \le x \le b_i \\ b_i \le x \le c_i \\ c_i \le x \le d_i \end{array} \begin{cases} (x\text{-}ai)/(bi\text{-}ai) \\ 1 \\ (d_i\text{-}x)/(d_i\text{-}c_i) \end{cases}$$

2.3 Exponential membership function

Exponential membership functions and sign-wave functions have also been employed. In view of the fact that all parameters in the data sets under study here deal with physical attributes that are most likely normally distributed, the normal distribution suggested itself as a membership function for the current analysis. This distribution is exponential and probability is measured as between zero and one. The membership function has the happy property that no parameter will actually have zero membership but will have a diminishingly small membership as its location diverges far from the mean. The probability of being at the mean is ½ and not 1 so for no other reason than esthetics we divided our membership percent by ½ in order to produce membership of 1 for parameters actually at the mean.

The mean for each membership function is the mean of that parameter over existing data. Spread of the membership function is achieved by adjusting the standard deviation. Before training this value is set at the standard deviation of that parameter over existing data. And membership percentage is determined by the probability of the tail. Equation 3 describes this.

$$(4)\ \text{Membership} \begin{cases} = P(X)/.5 & \text{for } X \le \mu \\ = (1 - P(X))/.5 & \text{for } X > \mu \end{cases}$$

Thus for a given parameter such as "thickness"
Standard deviation = σ
Mean = μ
Measured value = x
If x > μ then membership = (1- P(x))/0.5
If x < μ then membership = P(x)/0.5

2.4 General Aggregation Operations

Aggregation operations are operations by which several fuzzy sets are combined into a single set. In general, any aggregation operation is defined by a mapping [14]

(5) $h:[0,1]^n \rightarrow [0,1]$

for n >=2. When applied to n fuzzy sets $A_1, A_2, ...A_n$, defined on X, h produces an aggregate fuzzy set A by operating on the membership grades, $\mu(x)$ of each x ε X in the aggregated sets. Thus, we have:

(6) $\mu_A(x)=h(\mu_{A1}(x), \mu_{A2}(x),\ldots \mu_{An}(x))$ for each x εX:
The following requirements express the essence of the notion of aggregation, after Klir and Yuan (Klir and Yuan, 1995):
Axiom 1. h (0,0, ...,0) = 0 and h (1, ... 1) = 1.

Axiom 2. For any pair of numbers a_i, b_i, where a_i ε [0,1], b_i ε [0, 1], if $a_i \geq b_i$ for all i then $h(a_i) \geq h(b_i)$, that is, h is monotonic non-decreasing in all its arguments.

Two additional axioms are also employed to characterize aggregation operations.

Axiom 3. h is a continuous function, that is it guarantees that an infinitesimal variation in any argument of h does not produce a noticeable change in the aggregate.

Axiom 4. h is symmetric function in all its arguments, that is $h(a_i) = h(a_p)$ for any permutation, p, of the arguments, that is the aggregated sets are equally important.

Fuzzy unions and intersections can be qualified as aggregation operations on only two arguments, their property of associativity, provides a mechanism for extending their definition to any number of arguments. Also, in the case of bi-symmetry and strict monotonicity, the extensions are straightforward. Thus, generalized fuzzy unions and intersections can be viewed as special aggregation operations that are symmetric, usually continuous, and are required to satisfy conjunctive and disjunctive attitudes. As a result, generalized fuzzy unions and intersections can produce only aggregates that are subject to the following restrictions [12]

(7) $\max(a, b) \leq \mu(a, b) \leq \mu_{\max}(a, b)$

(8) and $i_{\min}(a, b) \leq i(a, b) \leq \min(a, b)$
where u(a,b) and i(a, b) is the fuzzy union, and fuzzy intersection, respectively.

Generalized Dombi's operations [13] possess properties of fuzzy unions and intersections, which will be utilized in this work. They are defined as follows:

Dombi's Fuzzy Union:

(9)
$$\frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{-\lambda}+\left(\frac{1}{b}-1\right)^{-\lambda}\right]^{-\frac{1}{\lambda}}}$$

where λ is a parameter by which different unions are distinguished, and λ ε (0,∞)

Dombi's Fuzzy Intersection

(10)
$$\frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{\lambda}+\left(\frac{1}{b}-1\right)^{\lambda}\right]^{\frac{1}{\lambda}}}$$

where λ is a parameter by which different intersections are distinguished, λ ε (0,∞).

From the above inequalities, one can see that generalized fuzzy unions and intersections do not produce any aggregates that generate values between min and max. Hence the general rules previously stated.
A AND B where A and B are limited to the range (0, 1), becomes equivalent to min (A, B) and
A OR B, where A and B are limited to the range (0, 1), becomes equivalent to max (A, B).

Aggregates that are not restricted in this way, however, are allowed by Axioms 1 through 4. Operations that produce them are called averaging operations. There are several classes of averaging operations are a compromise between mm and mm One such class, which covers the entire interval between the min and max operations consists of generalized means. This class of operations was used in fuzzy prediction, defined after (Klir and Yuan, 1995) as:

$$(11)\, h_a(a_1, a_2, \ldots a_n) = \left(\frac{a_1^{\alpha} + a_2^{\alpha} + \ldots + a_n^{\alpha}}{n} \right)^{\frac{1}{\alpha}}$$

where α is a parameter by which different means are distinguished, and α ε R (α is not equal to 0). In our application, we opted for α = 4, based on our experience. Function ha clearly satisfies all axioms of aggregation operations and, consequently, it represents a parameterized class of continuous and symmetric aggregation operations.

A variant of this approach is used in the following analysis. The approach uses weighting values for each alpha and these weighting value may be adjusted for optimum performance. A further variant that is investigated involves multiplication rather than averaging. In that case the individual weighting factors are of no consequence. Linguistic statements are usually associated with defuzzification. However in a classification problem, when there are many parameters of equal importance, it is difficult to formulate such linguistic statements.

2.5 Fuzzy Sets and Data Mining

The typical analytic approach to data mining employs random selection to partition the data into two groups (sometimes 3). The first group is used to “train” the data mining algorithm. For example if regression is employed, the training partition is used to set the parameters used in the regression formula. The usual measures for performance are used for this purpose, residual error, p-values of coefficients, and so on. When two partitions are use they are usually based on a 60/40 or 70/30 ratio. These are generally accepted ratios but it not really possible to demonstrate a theoretical optimum and if the data set is small more may be required for training.

Once the data mining model has been selected it is evaluated against the second partition, usually called the validation partition, to see how it performs. If the data has been divided into three partitions then different models may be evaluated against the second data set and small further adjustments may be made to the most promising model before it is evaluated against the third partition of data. Only two partitions were employed for the analysis reported here.

Data in the first, “training” partition are separated into two groups. One containing date of known benign instances and the other containing data of the known malignant instances. For each set the mean and standard deviation is computed for each of the attributes. These values are then use to set the membership function for each attribute as described in equation (4) above.

When the model is “run” the data from each record is evaluated for membership in each of the two groups of fuzzy sets. The results are aggregated and compared with the decision for benign or malignant being made on the basis of which aggregation has the larger value. Data from both the first, “training”, partition and the second, “validation”, partition are evaluated by the model. Usually the training data performs better as might be expected.

2.6 Methods employed here

Several aggregation methods have been compared. There is not clear analytical president for some of the selections. Establishing a data mining model is in some ways an empirical process. That is why the training partition and the validation partition methodology is employed.

The averaging method would be defined as follows. In this case the individual membership functions are not weighted but the aggregate is.

(12) $\mu_A(x) = \sum_{k=1}^{n} \mu_{Ak}(x_k)$ additive

(13) $\mu_A(x) = \sum_{k=1}^{n} b_k \mu_{Ak}(x_k)$ additive with individual b_k weight adjustment.

μ_A is the membership function for class A.

μ_{A1} is the membership function for class A, parameter number 1

μ_{A2} is the membership function for class A, parameter number 2.

x_1 is the value of parameter # 1 associated with object x.

x_2 is the value of parameter # 2 associated with object x.

Equations (12) and (13) would be repeated for μ_B where A is associated malignancy and B with benign.

The multiplicative aggregates are defined as:

(14) $\mu_A(x) = \prod_{k=1}^{n} \mu_{Ak}(x_k)$ multiplicative

(15) $\mu_A(x) = \prod_{k=1}^{n} (\mu_{Ak}(x_k))^2$ multiply squares

In each instance the individual $\mu_{Ak}(x_k)$ membership functions may be adjusted during the training phase to achieve minimum classification error.

2.7 Training

Training is the distinctive feature of the somewhat heuristic model presented here. During the training stage the standard deviation used in each membership function can be adjusted to improve overall performance of the "training" partition data. Increasing individual standard deviation has the effect of increasing the width of that membership exponential function. The function now departs from the true statistical probability function, but the choice of the normal probability function for the exponential was arbitrary and not usually used in fuzzy logic anyway. The data represents physical attributes and such attributes might be assumed to have normal distributions. However the data was obtained from human observations that are quantized and hence not normal.

The unaided model performed quite well. However performance was improved by adjusting the value used for standard deviation of each individual parameter associated with each class.

The additive method of combining individual parameter memberships and the product method performed comparably in practice. It would seem that using the squares of membership values would emphasize parameters where there is "good" membership. However empirical work suggests that there was no significant difference.

An adjusting coefficient may be used when adding membership values. This allows a small amount of improvement in some instances.

With some data mining models over training is possible. In that case the error percentage is very good for the training partition but markedly degraded for the validation set. This situation did not present itself.

2.8 Training Optimization

There are many parameters and each one has an adjustable standard deviation. Adjusting by eye is time consuming. A genetic algorithm was used in an attempt to optimize this process. The genetic algorithm used was a Microsoft Excel add-in called xl bit (http://www.xlpert.com) [15]

3. Results with Data Set #1

2.1 Statistical Method

Classification data mining is often accomplished on a statistical basis with a logistic regression model. In order to compare performance of the fuzzy model logistic regression was applied to the same data. The vehicle used was a Microsoft Excel add-in called XLMiner [16].

We report results obtained with several data sets here. Data set number 1 is "Mammographic Mass Data" made public by Schulz-Wendtland [17]. The mammographic Mass database consisted of the following parameters.

1. BI-RADS assessment: 1 to 5 (ordinal)
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binominal)

Missing Attribute Values:

- BI-RADS assessment:	2
- Age:	5
- Shape:	31
- Margin:	48
- Density:	76
- Severity:	0

The predicted classification is either benign or malignant. Cases with missing variables were removed. Summary of results are in the table A. For these calculations the gain of each individual fuzzy set was left constant during the aggregation (averaging) process but individual standard deviations were adjusted so as to narrow or increase the width of the membership function. Percentages shown are error

TABLE A	Benign Errors	Cancer Errors	Overall Error
logistic regression (Training)	15.95% 41/257	15.61% 37 / 237	15.79% 78/494
logistic regression (Validation)	17.96% 30/167	15.95% 26/163	16.97% 56/330
Fuzzy Logic Model sum of membership values (Training)	19.84% 51/257	16.88% 40/237	18.4% 91/494
Fuzzy Logic sum of membership values (Validation)	16.17% 27/167	17.79% 29/163	17.00% 56/330
Fuzzy Logic Model using membership product (Training)	19.46 50/257	16.03 38/237	17.8 88/494
Fuzzy Logic Model using membership product (Validation)	17.37 29/167	15.95 26/163	16.7 55/330
Membership product & genetic algorithm to adjust Sigma (Training)	17.9% 46/257	11.39% 27/237	14.8% 73/494
Membership product & genetic algorithm to adjust sigma (Validation)	20.96% 35/167	12.27% 20/163	16.7% 55/330
using sum of memberships & genetic algorithm to adjust Sigma (Training)	15.18% 39/257	15.61% 37/237	15.4% 76/494
using sum of memberships & genetic algorithm to adjust sigma (Validation)	15.57% 26/167	16.56% 27/163	16.1 53/330

percentages. The ratios, such as 7/270, show error count and total records in that classification.

Validation scores are the important ones from the point of view of data analysis. Also there is not a great deal of variation among results. This seems to match results from some other researchers in this area.

Training and validation scores are shown. As anticipated, training scores are slightly better.
Using the genetic algorithm to optimize the membership functions contributed a slight improvement when the decision rule was based on the sum of membership values and made no improvement in overall accuracy when the product of membership values was used. In this latter instance it altered the probabilities for failures in detecting malignancy by increasing them with a consequent decrease in failures to classify as benign.

The genetic algorithm, like other training adjustments, worked to minimize the overall error in the training process. The hope is that this will project into improved performance with the validation partition data. It is true that some adjustment of the sigma values in the individual membership functions can be used to shift errors from one type to another while still maintaining overall error rate near constant.

In lieu of adjusting the standard deviations a weighting factor for each attribute can be adjusted. Of course, this is only possible when the aggregation function is a summation. Thus:

$$\mu_A(x) = \sum_{k=1}^{n} b_k \mu_{Ak}(x_k) \qquad (16)$$

The coefficient b_k applies to the attribute k. Results in this simple case are shown in Table B. And they are not markedly different from results previously obtained

TABLE B	Benign Errors	Cancer Errors	Overall Error
sum of memberships & genetic algorithm to adjust weight (Training)	13.23% 34/257	20.25% 48/237	16.6% 82/494
sum of memberships & genetic algorithm to adjust weights (Validation)	11.98% 20/167	20.25% 33/163	16.1 53/330

At least in some instances weighting the membership functions is as useful as adjusting the individual shapes of the membership functions. It is not obvious that this would be the case. It does say that simple dynamic weighting adjustment during the training phase may be as useful as anything else.

Results also suggest that the statistical logistic regression approach is about as valuable as the fuzzy logic approach. This is not the case in control systems where fuzzy logic performs exceedingly well. This paper reports on an empirical investigation with a limited amount of data. It offers no proof that statistical methods will always perform as well or less well as fuzzy methods. Not all researchers have reported a comparison so one hesitates to make generalized observations.

4. Results Data Set #2

Data set #2 was Breast Cancer data made public by Mangasarian & Wolberg [18] [19]. This data set was also briefly reported in our previous paper. The methodology has now been greatly expanded with adjustment of individual membership functions and use of the genetic algorithm. Data set parameters are as follows:

1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 – 10
11. Class	2= benign; 4 = malignant

TABLE C	Benign Errors	Cancer Errors	Overall Error
logistic regression (Training)	1.11% 3/270	2.88 4/139	1.71% 7/409
logistic regression (Validation)	4.02% 7/174	3.06% 3/98	3.68% 10/272
using sum of memberships & genetic algorithm to adjust Sigma (Training)	4.07% 11/270	13.67% 19/139	7.3% 30/409
using sum of memberships & genetic algorithm to adjust sigma (Validation)	1.15% 2/174	10.2% 10/98	4.4% 12/272

The predicted classification is either benign or malignant. Standard deviation of membership functions were adjusted using the genetic algorithm. Sets were aggregated as an averaging process similar to equation (11) and selection made on the basis of which was the larger. Summary of results are in the table C.

Again results from the fuzzy approach are very good but not better that statistical logistical regression. They are in some respect comparable. Results for data set #2 were considerably better than those for data set #1 possibly because data set #2 had more attributes from which to establish set memberships. All results were really comparable with results of other researchers as reported earlier in this paper.

The Fitness chart developed by the genetic algorithm software for data set #2 is presented as Figure 2. It shows that the genetic algorithm has approached asymptotically a horizontal line and further iterations may not produce improvement unless there are local minimums. Considering the nature of the problem, local minimums are unlikely.

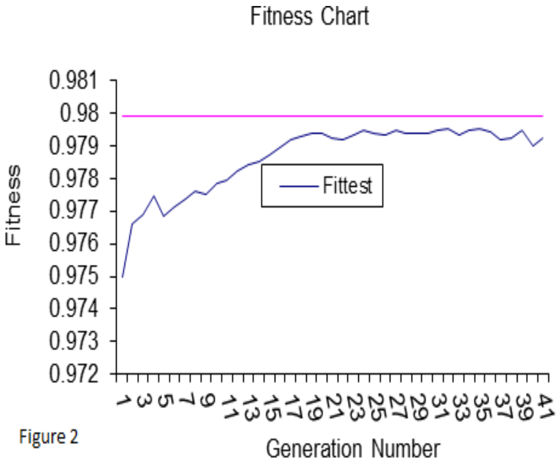

Figure 2

5. Solution Efficacy

There is naturally a question as to why these analytic attempts are not more successful in discriminating among the two classes. In most medical and probably most biological entities various attributes will be highly correlated. Figure 3 present the correlation matrix for data set #2. Here it can be seen that a high degree of correlation exists. In general this makes statistical regression less accurate or questionable at least. This limitation will certainly be present in all attempts at discrimination of biological entities until or unless orthogonal characteristics can be found.

	Thi	Uni	sh	ad	siz	nu	Ch	Nu	Mit	cla
Thick	1.0									
Uniform	0.6	1.0								
shape	0.7	0.9	1.0							
adhesion	0.5	0.7	0.7	1.0						
size	0.5	0.8	0.7	0.6	1.0					
nuclei	0.6	0.7	0.7	0.7	0.6	1.0				
Chromatin	0.6	0.8	0.7	0.7	0.6	0.7	1.0			
Nucleoli	0.5	0.7	0.7	0.6	0.6	0.6	0.7	1.0		
Mitoses	0.3	0.5	0.4	0.4	0.5	0.3	0.4	0.4	1.0	
class	0.7	0.8	0.8	0.7	0.7	0.8	0.8	0.7	0.4	1.0

Figure 3 Correlation of Data Set #3

6. Observations

Analysis shows that the parameters of data sets are strongly correlated. This is not surprising since they usually represent physical characteristics of some class of objects.

The multiple linear regression model assumes independent variables. Collinearity may make it impossible to see the contributions of each variable to the result. It can lead to incorrect answers when using regression to fit higher order polynomials. But this is not a problem here. In the heuristic approach the model is based on actual performance. No one attempts so say it would be valid for classes different from the ones with which the model is validated.

The fuzzy logic approach presented here is not limited to selection among two categories. Clearly the resolution through a logical OR operation can be applied to three or even more categories and presumably even more. [14] There is no practical limit provided there is sufficient data. Three classes requires that training data be divided into three sets; for classes would require 4 sets, etc. Statistical evaluation of results would require some reasonable number of records for each class.

7. Conclusions

It is clear that fuzzy logic employing a exponential membership function, weighted averaging, and logical OR operation is an effective tool in data analysis involving discrimination where many parameters are involved. It performs as well as results reported by other researchers looking at similar problems with different fuzzy methods and hybrid fuzzy methods. Because membership functions may be adjusted individually the usual method of dividing data into a training partition and a validation partition is appropriate. Rather extensive adjustment ("learning") during the training stage improves performance. This may include use of a genetic algorithm. Results observed here must be classified as anecdotal because only two data sets were used. However, from these observations it appears that fuzzy logic is comparable but does not outperform usual data analysis methods such as logistic regression. In instances where image analysis is a principal feature (as reported in other papers noted earlier) the situation does not lend itself to standard statistical regression methods.

References

[1] Zadeh, L.A. "Fuzzy sets", Information and Control 8 (3): 338–353 1965.

[2] Takahashi Y. (1995) "A fuzzy Query Language for Relational Databases", in Fuzziness in Database Management Systems, Physica Publisher, Heidelberg, 365-384.

[3] Cios K J, Goodenday L S, Sztandera L M ″Hybrid intelligence system for diagnosing coronary stenosis. Combining fuzzy generalized operators with decision rules generated by machine learning algorithms", Engineering in Medicine and Biology Magazine, IEEE, Vol. 13 Issue 5, 1994 pages 723-729

[4] Cios K J, Shin I, and Goodenday LS "Using Fuzzy Sets to Diagnose Coronary Artery Stenosis" , IEEE Computer Magazine Pages 57 -63, 1991

[5] Cios K J, Liu N (1994) "A machine Learning algoriothm Using Interger Linear Programming" KYBERNETES, MCB University Press,U.K. 1994

[6] Cios K J, Noraes I 91991) "An inductive Learning Algorithm" KYBERNETES, MCB University Press, U.K. 920:3), pages 19-30 1991

[7] Cios K J, Freasier R E, Goodenday L S, Andrews L T "An expert system for diagnosis of coronary artery stenosis based on 201TI scintigrams using the Dempster-Shafer theory of evidence", Oxford University Press, U.K. (6) No. 4 pages 333-342 1990

[8] Malek J, Sebri A, Mabrouk S, Torki K, Tourki R Journal of Signal Processing Systems Volume 55 Issue 1-3, pages 49-66, April 2009 Kluwer Academic Publishers

[9] Bagher-Ebadian H, Soltanian-Zadeh H, Setayeshi S, Smith S T "Neural network and fuzzy clustering approach for automatic diagnosis of coronary artery disease in nuclear medicine" IEEE Transactions on Nuclear Science Vol51 Issue 1 pages 184-192 (Feb 2004)

[10] Licata G "Employing fuzzy logic in the diagnosis of a clinical case" Health., Vol.2, No.3, pages 211-224 March 2010

[11] Pedrycz W. *Fuzzy Control and Fuzzy Systems* John Wiley & Sons, ISBN 0 86380 081 5 1989

[12] Klir G. J. and Yuan B. (1995), Fuzzy Sets and Fuzzy Logic – Theory and Applications, Prentice Hall, Englewood Cliffs.

[13] Sztandera L. M., Goodenday, L. S., and Cios K. J. (1996), A Neuro-Fuzzy Algorithm for Diagnosis of Coronary Artery Stenosis, Computers in Biology and Medicine Journal, 26(2), 97-107.

[14] Sanford J, Sztandera L (2010) "Classification Data Mining Using Fuzzy Logic", International Academy of Business and Public Administration Disciplines New Orleans Conference Proceedings October 2010, pages 120-127

[15] XLBIT Genetic algorithm for Microsoft Excel from http://www.xlpert.com based in Kuala Lumpur, Malaysia.

[16] XL-Miner is an Excel Add-in developed by Cytel Software Corporation and distributed by Resampling Stats Inc. http://www.resample.com

[17] Schulz-Wendtland 2007 Mammographic Mass Data. Original owner: Prof. Dr. Rüdiger Schulz-Wendtland, Institute of Radiology, Gynaecological Radiology, University Erlangen-Nuremberg, Universitätsstraße 21-23, 91054 Erlangen, Germany. **Relevant Paper:** M. Elter, R. Schulz-Wendtland and T. Wittenberg (2007), The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical Physics 34(11), pp. 4164-4172

[17] Mangasarian O. & Wolberg W. (1990) "Cancer diagnosis via linear programming", SIAM News, 23(5), 1-18.

[18] Mangasarian O & Wolberg W, 1990, "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[19] Wolberg W. & Mangasarian O. (1990) "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, pp. 9193-9196.