International Journal of Digital Information and Wireless Communications (IJDIWC) 5(2): 124-140
The Society of Digital Information and Wireless Communications, 2015 (ISSN: 2225-658X)

Heuristic Methods to Identify Fuzzy Measures Applied to Choquet Integral Classification of Breast Cancer Data

Ken Adams
Sino-British College,
University of Science and Technology
1195 Fuxing Middle Road, Shanghai China, PRC
ken.adams@sbc-USST.edu.cn

ABSTRACT

When the search space for optimisation or classification problems gets large heuristics and soft computing are often used to find the needed parameters. The Choquet integral methods needs one parameter for each subset of the condition attributes because these sets form a lattice on which the fuzzy measure needed for integration is defined. In this paper the frequency of how many data from each class pass through each set in the lattice provides information that is then used to construct the fuzzy measure. Several heuristics based on this idea are tested upon the well-known Wisconsin breast cancer data set. The heuristics prove to be as successful, or more so, than a range of other methods quoted in the literature when 10-fold cross validation is performed. In order for the Choquet to separate classes after integration a boundary value has to be calculated. Remarkably, it is found that switching the way to view success from number of records correctly classified to the distance of the wrongly classified record from the boundary leads to a hill climbing approach that can reclassify the data with a 100% success rate.

Keywords: Data mining, Optimisation, Genetic Algorithm, GA, Choquet Integral, Heuristics, Classification, Fuzzy measure, Wisconsin Breast Cancer, Hill Climbing, Big Data, Steady State GA, Real-Valued GA.

1 INTRODUCTION

A data set, with one decision attribute, consists of a finite number of condition attributes (variables) $X = \{x_1, x_2, ,, , x_n\}$ and a condition attribute y that can have a number of values. A record is one observation of all the attributes. The value of an attribute x_i over all records is considered to be a function of x_i and it is customary to write $f_{j,i} = f_j(x_i)$ as the j^{th} observation of the i^{th} attribute, [1]. The classification occurs when information from existing records is used to predict the value of the decision attribute for a new record when only its condition attributes are known.

With many variables to explore, search algorithms such as genetic algorithms (GA) are often used. The idea of genetic algorithms was invented by Holland [2] Looking at nature Holland considered Darwin's theory of natural selection, i.e. that the fittest members of the species survived to breed more often and thus produced more offspring. As offspring inherit their characteristics from their parents, fit members of a population would have a better chance of passing their characteristics on to the next generation than unfit members. Over

generations the species would evolve as the average fitness of its members rises. It is also possible that some random change might appear in a member of the next generation. If this random change is of benefit then that individual will be successful and the random change will be passed on to yet another generation.

Heuristics are often used to direct the search in a possibly useful direction that may achieve a result of high fitness by the GA [3, 4]. A heuristic can be considered as some simple and fast technique which, although may not find the optimum will often find a good place to start. One or more of the members of the population are heuristically enhanced before the search begins. Blind heuristics, for example bit-flipping [5, 6] can be useful but use no information gathered from the data. In Adams [7] Chapter 6 heuristics based on the autocorrelation and on the discrete Fourier transform were developed. They were used to optimise the number of coefficients of multiple-valued polynomials. Results were shown to be significantly better than running the genetic algorithm without heuristics.

Heuristics methods can be a successful technique in their own right. In [8] a novel heuristic (activity of the variables) was used to partition a decision table into sub-tables and generate a simpler set of rules than the original. The proposed activity heuristic is shown to produce better results than a previously-published rough set analysis of the same data; and shows comparable results to the well-known ID3 information theoretic approach.

The Choquet integral is a well-known classification method. It requires that a coefficient, called a measure be allocated to every subset of X the condition attributes. In this paper heuristics are developed to find a good set of coefficients. Genetic algorithms are a common tool for finding the large number of coefficients needed to find the measures needed for Choquet classification [1], [9]. It is anticipated that fresh ideas on heuristics will help future authors with GA and other searches.

The data used in this experiment is the well-known Wisconsin Breast Cancer data set [10]. Results reported in this paper are comparable to others reported for the same data set in the literature.

Sections to follow are: §2 Describes the discrete Choquet integral, §3 Provides an illustrative example of how the integral may have use in a common situation of assessing employee performance. Also the example makes a comparison with results obtained by using a simple weighted average method. §4 The Wisconsin breast cancer data set is described and how the Choquet integral is calculated on this data set is demonstrated, §5 Rationale behind the heuristics developed are discussed, §6 Heuristics Developed are described in detail, §7 The results obtained by the heuristics described in this paper are compared to other reports in the literature, §8 Describes a hill climbing technique that examines small increments of the measure, and how a new way to optimise, by minimising the distance of all the wrong classifications to the cut-off point helps make this successful, §9

describes the use of a genetic algorithm for the four line segment method (that method is first described describes in section 6.5), finally §10 Concludes the paper.

2 DISCRIPTION OF THE DISCRETE CHOQUET INTEGRAL

The Choquet integral was proposed by Choquet [11] to study capacities in the field of economics. The discrete Choquet integral is used as an aggregation technique and has been developed by many authors for classification purposes, including [12] and [13]. The integral takes the form:

$$(C)\int f\, d\mu = \sum_{i=1}^{n}\left[f\left(x_i^*\right) - f\left(x_{i-1}^*\right)\right]\cdot \mu\left(x_i^*, x_{i+1}^*, \ldots x_n^*\right) \quad (1)$$

Where the variables x_i^* are a permutation of the x_i into ascending order of their value. Here μ is a measure or weighting associated with each subset.

A three variable example of how the integral is calculated will now be given. Suppose $X = \{x_1, x_2, x_3\}$, then there is a lattice of subsets [see Figure 1]. For three variables there are eight subsets in the lattice. The lattice is partially ordered by inclusion. A weighting μ called a measure is allocated to each subset. In this example the weights are as follows:

$\mu(\{\phi\}) = 0$, $\mu(\{x_1\}) = 0.4$, $\mu(\{x_2\}) = 0.1$, $\mu(\{x_3\}) = 0.2$, $\mu(\{x_1, x_2\}) = 0.4$, $\mu(\{x_1, x_3\}) = 0.5$, $\mu(\{x_2, x_3\}) = 0.6$, $\mu(\{x_1, x_2, x_3\}) = 1$.

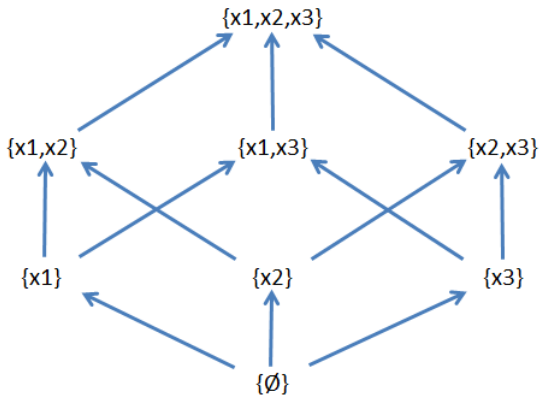

Figure 1 Lattice of Subsets

Now suppose that a record has the following values $f(x_1) = 4$, $f(x_2) = 8$, $f(x_3) = 2$. Firstly the variables are re-arranged in ascending order of size of their values, x_3, x_1, x_2. Then starting from the top at the universal set X, a path is made through the lattice. At each step the variable with the next lowest value drops out. $\{x_1, x_2, x_3\} \rightarrow \{x_1, x_2\} \rightarrow \{x_2\} \rightarrow \{\phi\}$. The measure for each subset in the path is multiplied by the difference between the lowest value of the variables in the subset and the value of the variable that has just dropped out. Then all the products are added up. The calculations are: $(2-0)\times 1 + (4-2)\times 0.4 + (8-4)\times 0.1 = 3.2$.

Aggregation is a way to replace all the numbers $x_1, x_2, , , x_n$ by just one number (a sort of averaging). Then if the aggregated figure falls below a certain cut-off the record is classified in one way and if above it is classified in another. The use of a simple weighted average takes into account the contribution each condition attribute makes to the information available to estimate

the decision attribute. However, the Choquet Integral also aims to take account of the interaction between variables, and has many applications such as for optimisation [1], to shift work [14], for multiple regression [15], and for decision rules [16].

3 AN ILLUSTRATIVE EXAMPLE OF MEASURES AND THE CHOQUET INTEGRAL

This example will explain the difference between a lattice bases approach and simple weighted average. Imagine that a manager of a small business wishes to have a system to assess her workers to see whether or not they should be retrained and in what order they should be retrained. She considers that the three attributes that are most important are: initiative, pride in your work, and listening. From her experience she knows that these quantities interact so she rates them individually and in combination. Each item and each subset of items is allocated the measures below, where $\mu(Y)$ is the measure of set Y. To simplify notation P will represent pride, L is listening and I is initiative. Then the measures that she uses are as follows:

$\mu(\{\theta\}) = 0$, $\mu(\{P\}) = 0.2$, $\mu(\{L\}) = 0.3$, $\mu(\{I\}) = 0.5$, $\mu(\{P, L\}) = 0.5$, $\mu(\{P, I\}) = 0.8$, $\mu(\{L, I\}) = 0.6$, $\mu(X) = \mu(\{P, L, I\}) = 1$, The various measures are illustrated in the lattice diagram below.

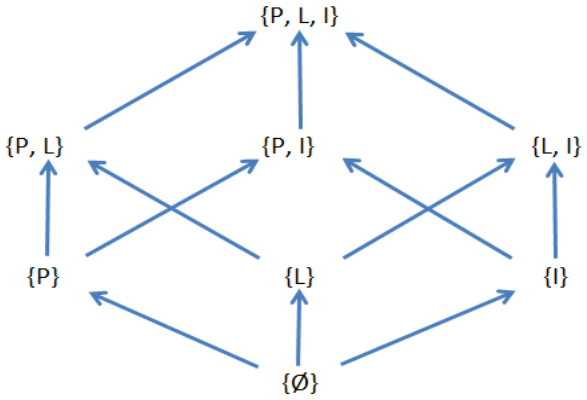

Figure 2 Measures Assessment of Employees

Some things to note are: The measure of the empty set θ is zero, with no qualities to assess the measure has to be zero. Generally the measures for singleton sets do not have to sum to one as it would do for a weighted mean. The manager believes that considered on its own, initiative is the most important quality that a worker can possess and she weighs this at 0.5. The weighting of a singleton set, a set with only one item is called a density. The ability to listen to instructions is rated a little behind initiative with a weighting of 0.3, and the least important when considered on its own is pride in your work only 0.2.

Then the manager has to consider how the attributes interact with each other. She believes, from her experience, that pride and listening do not interact much at all so she weights the set {P, I} as just the addition of the measures for {P} and {L} in other words they are additive. The attributes of pride and initiative {P, I} is believed to work well together and reinforce each other so the manager rates their combination quite highly at 0.8. They are said to be super-additive. However she believes that the combination of

listening and initiative can sometimes not combine so well, and sometimes detract from each other, so their joint measure is less than the sum of the two separately. They are said to be sub-additive.

Employees are assessed on a Likert integer scale from zero to 5 with the meaning of each value as follows:

Table 1: Likert Scale for Employees

Language Meaning	Value
Not capable	0
Can just manage	1
Adequate	2
OK or average	3
Good	4
Excellent	5

Using these criteria she rates each employee as follows:

Table 2: Rating Employees

Employee	Pride	Listening	Initiative
A	1	5	3
B	3	5	1
C	3	1	5
D	5	3	1

In this example all four employees score, a one, a three and a five across the attributes however they make these values at different places; that enables a discussion on how the Choquet integral differentiates between them.

The calculation of each integral follows the same steps as those explained in the previous section. They are given here;

The order of the assessment scores for A is:

P< I <L with integral $(1)\times 1+(3-1)\times 0.6+(5-3)\times 0.3=2.8$,

Then for B is I< P <L with integral $(1)\times 1+(3-1)\times 0.5+(5-3)\times 0.3=2.6$,

for C is L< P <I with integral $(1)\times 1+(3-1)\times 0.8+(5-3)\times 0.5=3.6$,

and finally for D is P< L <I with integral $(1)\times 1+(3-1)\times 0.5+(5-3)\times 0.2=2.4$.

Employee C obtains the highest score of 3.6. His strongest strength is initiative which is weighted as the most important single attribute by the employer. The combination of his top two strengths, initiative and pride, is also rated highly, in fact it is rated the top out of all secondary interactions. A secondary interaction is an interaction between two attributes.

The next strongest is employee A with 2.8 his top strength of listening is not rated so highly as the strength of initiative; nor are the combination of this top two strengths, listening and initiative rated as highly as the top two for employee A.

The third strongest is B with 2.6. His top strength is only rated in the middle of the three attributes and the combination of his best two is the least highly rated of the secondary interactions.

Finally the weakest employee is D with 2.4. His top strength of pride in his work is the lowest weighted as an individual factor and the combination of his best two qualities is also the lowest rated in combination. So he should be the first to get retrained.

With this example a simple averaging of the three scorers would put everyone equal at 3, so they are not distinguishable. A weighted average could be calculated using the densities as weights. For example for employee A, his pride score is weighted at 0.2, so that adds 0.2×1 to his score. The average for A is $0.2\times1+0.3\times5+0.5\times3=3.2$. The weighted averages turn out to be a follows: A has 3.2, B has 2.6, C has 3.4 and D has 2.4. The weighted average for A is higher than the number obtained when then the Choquet integral is used (2.8). This is because the combination of listening and initiative is sub-additive, less that the two attributes separately. The weighted average for employee C is lower than that calculated by the Choquet integral because the Choquet integral favors the combination of pride and initiative which are C two strongest qualities.

4. WISCONSON BREAST CANCER DATA SET

The Wisconsin Breast Cancer data set is available at [10]. It consists of 699 records of patients who were examined for breast cancer and had various measurements taken. As sixteen of the records have a missing value only the 683 complete records are used here, of these 239 had cancer and 444 had not. The data was collected from January 1989 until November 1991 by Dr. WIlliam H. Wolberg (physician) at the University of Wisconsin Hospitals Madison, Wisconsin, USA [17, 18].

The scale of each measurement is an integer in the range one to 10, and the attribute information for the data is given below. For this paper the first condition attribute x_1 is Clump Thickness and the last condition attribute x_9 is Mitoses. The decision attribute y is Class. In order to integrate with the authors existing software, the decision class was relabelled as $y=1$ meaning benign and $y=2$ meaning malignant.

Attribute Information:

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 for benign, 4 for malignant)

There follows an example of classification using the Choquet Integral on a data record from the Wisconsin data set [Note: It is unnecessary to calculate differences when adjacent attributes in ascending order have the same value]. The nine condition attributes are labelled from to and their values recorded in Table 3:

Table 3: Example Record

Att.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
F(x)	6	8	8	1	3	4	3	7	1

After sorting in ascending order the table becomes:

Table 4: Sorted

Att.	x_4	x_9	x_5	x_7	x_6	x_1	x_8	x_2	x_3
F(x)	1	1	3	3	4	6	7	8	8

Table 5: Calculation

Att. x_1 to x_9	**Diff.**	μ	**Diff× μ**
111 111 111	1	1.00000	1.000000
111 011 110	2	0.670496	1.340992
111 001 010	1	0.656327	0.656327
111 000 010	2	0.656327	1.312654
011 000 010	1	0.300222	0.302222
011 000 000	1	0.302222	0.302222
000 000 000			
		Total	**4.914417**

As you can see the value of the integral is heavily dependent upon the measures assigned to each subset. There are 512 subsets of the nine condition variables of the Wisconsin data set. It is difficult to allocate optimum values to all of these coefficients thus search techniques like genetic algorithms are often used for this and for similar problems.

5. RATIONALE

The data itself provides information on how new data may be processed. If a subset is traversed by a new record and all the evidence from the previous data says that any patient whose records passed through that subset has cancer then there is strong evidence to suggest that this patient may have cancer. Likewise, if all information from prior records showed no cancer, it would be expected that the patient's condition was benign. The frequency considered here to be important is the frequency of class two (malignant). Looking at the whole data set of 683 records there are 239 which are malignant. That is an overall probability of 239/683. If a patient passes through a subset which is close to this global ratio then there is little evidence either way. If far from the global probability then passing through the subset may well provide a strong indication of the patient's condition. Thus expected frequency is the tool used in this paper.

The other key idea is that instead of looking for values for all 512 measures (coefficients) and searching through a vast number of combinations of settings. A small set of parameters is instead examined. Suppose a black box needs to receive k parameters in order to operate and its job is to generate a much larger set of m parameters, needed for optimisation. Then it is easier to search the k parameters than the m directly. This is useful to get an approximate solution that can then be passed through some further optimisation procedure. It is believed that this idea can be

applied more generally to many other optimisation problems.

6 HEURISTICS DEVELOPED

The ideas behind the various heuristics are explained in the sub-sections following. The heuristics design a function and if this function is graphed then the horizontal axis (x) is the frequency of class two records passing through that particular subset in the lattice; and the vertical axis (y) is in the range $[-1, 1]$. However when the measure is finally made the ranger is rescaled into [0,1]. These methods were first reported by the author in [19].

6.1 Mean

This simply calculates the mean of the nine condition attributes and then chooses the best cut off point. Using the Choquet integral the mean is the result of calculations when the measure of a subset M is $|M|/9$ where $|M|$ is the cardinality of M [15].

6.2 Class Frequency Formula

For each subset in the lattice, a frequency count is made of the number of records that take a path through the subset that are of either class one (no cancer) and class two (malignant). The grand totals across all of the data set are also used. If both frequencies are zero then the measure is assigned to be zero. When at least one of the frequencies in non-zero the following ratio is calculated:

$$\frac{Class2\ freq \times TotalOne - Class1\ freq \times TotalTwo}{Class2\ freq \times TotalOne + Class1\ freq \times TotalTwo} \quad (2)$$

Here *TotalTwo* is the total number of records with classification two (malignant) in the data set; $Class2\ freq$ is the number of records classified as two that pass through this particular subset.

This ratio has the following properties:

If Class2 is zero the ratio becomes -1.

If Class1 is zero the ratio becomes 1.

The expected number of Class2 (cancer patients) that pass through the subset when the grand totals are used for calculation are:

$$(Class1\ freq + Class2\ freq) \times \frac{TotalTwo}{TotalOne + TotalTwo} \quad (3)$$

When the frequency of class two (cancer) is exactly that which is expected the ratio calculates to zero. The measure is calculated from the ratio in the following manner:

measure= 0.5*(ratio +1) this re-scales so that the range is the interval [0, 1].

6.3 Simple Straight line Method

The measure is calculated using a straight line that passes through the origin and the point (Class1freq + Class2freq, 1). Here, the horizontal axis is the class two frequency, so if every record is class two then the value *one* will be calculated. If there are no class two records, then the measure will be zero.

6.4 Two Straight line Segments

This method is designed that if the frequency of class two is the same as expected, then the y-coordinate will be 0; If all are class two, then the y-coordinate is 1; and if no class two, it is −1.

6.5 Four Straight line Segments

This method is similar to the two straight line method and is designed that if the frequency of class two is the same as expected, then the y-coordinate will be 0; If all are class two, then the y-coordinate is 1; and if no class two, it is −1.

However instead of just one straight line between the expected value and the sum of the frequencies two straight lines are used. The mid-point is used as an additional x-coordinate (frequency) and the y-coordinate can be set arbitrarily. The two lines will intersect at these coordinates. A search process can be used to find a suitable y-coordinate.

Similarly the mid-point between the origin and the expected value can be used to generate another pair of lines. The motivation is that frequencies close to the expected value do not provide strong evidence so this provides a way of making their measure even smaller than it would be if a single straight line was used. In total there are four line segments. Obviously more than four line segments may be used.

Here is an example where six records belonging to class one and fourteen belonging to class two, passes through a certain subset. The expected frequency for class two is then $20\times\frac{239}{683}=6.9985\approx 7$. The upper mid-point is $0.5\times(7+20)=13.5$ and this is to have a y-value of 0.2. Therefore, there is a change of line segments at the coordinates (13.5, 0.2). Similarly the lower mid-point is $0.5\times(0+7)=3.5$ and this is to have a measure of -0.3. The graph, Figure 3, shows the example function calculations with all four line segments

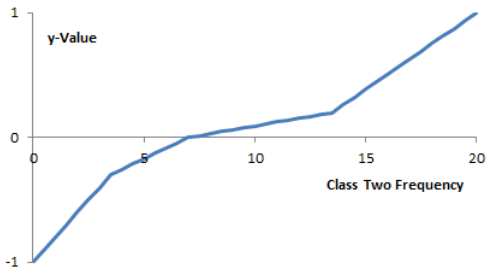

Figure 3: Example with Four Line Segments

7 RESULTS & COMPARISIONS

A reclassification of the entire data set was made. (Reclassification was performed on an artificial data set in [20] as a proof of concept of the Choquet integral classifier.) The reclassification results for this paper are presented in Table 4.

The formula method produced a reclassification rate of 97.22%. The structure of the formula has a logic to it. That if the class two frequency is what would be expected from the global frequencies then it calculates to zero; but if the frequency of class two is zero the formula returns -1, and if instead the frequency of class one is zero it calculates to +1. It may be worth considering using a bias where an extra parameter is added to the expected frequency

before calculation so that a number higher than (or lower than) the expected frequency is actually the zero. Such a parameter can be global or can even be optimized for each subset.

Table 6

Method	Reclassification
Mean	97.36%
Formula	97.22%
One Segment	97.51%
Two Segments	97.22%
Four Segments	97.80%

The mean method did well considering its simplicity with a reclassification rate of 97.36%. This result and the closeness of all the various results reported in the literature would suggest that the Wisconsin data set does not stress the classifications techniques hard enough to produce strong differences.

Three papers [21, 22, 23] used only the 683 complete records. In the papers the first 400 records were used for training and the last 283 for testing. The resulting testing accuracy's were: 98.10%, 97.50% and 98.10% respectively.

A comparison of the efficiency of five different classification methods on the Wisconsin Breast Cancer data set was reported in [24]. The method used was 10-fold cross validation and a t-test. The Weka (Waikato Environment for Knowledge Analysis) tool [25], which is a downloadable platform housing a variety of classification algorithms, was used for the calculations. The five methods were: Bayesian Network, Naïve Bayes, Multi-layer Neural Network, the ADTree decision tree, and the J4.8 decision tree. The resulting efficiencies for each classifier were: 97.20%, 96.11%, 95.58%, 95.49%, and 94.92% respectively.

Ten-fold cross validation was performed on four of the methods of this paper results are in Table 5.

Table 7

Method	10-fold cross validation
Mean	97.51%
Formula	96.20%
One Segment	95.71%
Two Segments	96.20%
Four Segments	96.62

As some of the test data would pass through subsets of the lattice that were not used in the training data, when both the frequencies of class one and of class two were zero the measure was set to $|M|/9$. Methods such as those used here are probably more suited for large benchmarks were lots of information regarding frequencies can be calculated.

The mean did slightly better than the best of the five methods reported in [24]. The two line segment method and the Formula method outperformed the lowest three reported namely: Multi-layer Neural Network, ADTree and J4.8 tree. The one line segment method tailed at the end.

The heuristics demonstrated here have shown to be comparable to and often better than the more complicated ideas of neural network and decision trees. This fundamentally different and

fresh approach should now be explored further on a wider set of benchmarks.

8 HILL CLIMBING

In the experiments reported in this section two ways to assess the fitness of a measure function are considered. It is similar to the bit-flipping method of [5] and [6] except it is suitable for continuous variables rather than binary. The first way is simply by counting the number of correctly classified records. The second way is to sum of the absolute distances of between the cut-off point and each wrongly classified record. If the value of the integral for a record with class one is above the cut off value it will be miss-classified, so how far it is above the cut-off is calculated. Similarly for all those records that are class two but have an integral below the cut off. Summing all these distances across all the miss-classified records provides an assessment of fitness. Clearly should the sum of these distances drop to zero then all records must be correctly classified. Over any iteration of the hill climbing method described here the cut-off point can also change as well as the distance.

If counting the number of correct classifications is the only consideration, then when two measures have the same count they are indistinguishable. However they can both have the same count but the distances away from the cut-off for one may be less than the other. In other words one of them may have more near misses. The area of the search space which has the most near misses may be the better place to begin exploring next. With this strategy the distance sum decrease sometimes ironically the fitness may also go down; but as all the miss-classified points are closer to the cut-off there is a good chance that after a few more iterations the fitness will go up again and indeed may surpass the original level from which it dropped.

The hill climbing heuristic will now be described. At the beginning a distance ‘delta’ is specified. Then the process loops through all the subsets. On each subset, three values of the measures are tested: the original value, the original plus delta, and the original minus delta. The value among these three that is the fittest will be retained. Then the process moves on to the next subset. The order of processing subsets is just the natural order used to loop through the variables in binary order, the most significant bit first. A complete iteration is when each subset has had the chance to hill climb once. Should the process pass over two consecutive iterations for which the distance has not changed, then the value of delta is dropped by a multiple of itself.

In the experiments performed the initial value of delta was set to 0.5 and this was changed in the following manner delta= 0.75*delta when no change in distance occurred after any two complete iterations. So each time it was changed it lost 25% of its original value. In this work the range of all the measure function was restricted to the interval [0, 1]. If the process calculated a measure less than zero that was re-set to zero; and if a measure would be more than one then that was reset to one.

To see how the two ways of measuring fitness are related one-hundred measure functions were randomly generated and tested on all the data set. It was found that distance to cutoff was negatively correlated to counting correct coefficients with a value of Pearson's r equal to −0.885.

For the first experiment ten measure functions were randomly generated and each was stored so that it could be used again. For each measure function the number of correctly classified records was recorded after ten iterations of hill climbing. When each subset was tested the value of the measure with the highest number of correct classifications was retained. Then the measure was re-set to its original values and results after ten iterations using the distance as fitness were performed. Now the lowest distance was retained. Results are presented in tables in this section.

Table 8: Correct Classifications and Distance

Run Number	**Non-Enhanced**	**Correctly Classify**	**Distance**
1	575	671	680
2	631	670	679
3	632	672	680
4	580	672	681
5	638	672	682
6	640	668	679
7	626	668	676
8	682	670	682
9	635	671	679
10	569	675	678
Means	**620.8**	**670.9**	**679.6**

In Table 8 the non-enhanced column records the original fitness of each measure function. The correct count column records the results using number correct classifications and the distance column uses distance as the fitness. When processing iterations of the distance method it is the highest number of correct classifications found thus far that is recorded. This permits a direct comparison with the count correct classifications method.

I can be seen from this table that the distance method was better than the correct count method in all of the ten runs in the experiment. Indeed in most runs the distance method gets close to the global optimum whereas the correct count method seems to get stuck in a local optimum.

Table 9: 100 Iterations of Correct classification count and Distance

Run Number	**Correctly Classified**		**Distance**	
	After 100 itts.	**First seen**	**Up to 100 itts**	**First seen**
1	671	3	680	7
2	670	6	682	4
3	672	2	683	21
4	672	6	683	17
5	672	1	683	27
6	668	3	683	27
7	668	3	682	55
8	670	2	683	12
9	671	4	683	29
10	675	7	683	25

Then the process was run again this time for a maximum of 100 iterations. Results are in Table 9. It is also recorded at which iteration the final result first appeared. So for example run 10 of counting correct classifications finished on 675 and this was first found at iteration 7. However run ten of the distance method found all 683 and this was first seen at the 25th iteration.

It can be seen that the count correct method obtains its best result after only a few iterations but then never seems to make any further improvement. The count correct method never found the global optimum of 683 even after 100 iterations. The distance method found the global optimum in seven out of the ten occasions. Also in the other three runs it came close to the optimum. Clearly on this data the distance method looks like a good approach on its own right and not just a heuristic.

It is interesting that the number of correct classifications improves anyway and can reach the optimum even though a different measurement, distance, is actually being optimized. To investigate further Table 10 displays the figures for run ten of the testing process up to the first ten iterations. Thus changes in fitness with changes in distance can be compared. It can also be seen how delta drops as the iterations progress.

This is an example of where the continuous distance metric can be better than the discrete count of correct classifications. You will observe that the distance is always coming down. After the first iteration the distance has dropped considerably from 87.86 to 18.84 (2 d. p.) and the number of correct classifications has risen considerably, by eighty-six.

In the next iteration the distance has dropped considerably again to 3.26 but now the number of correct classifications has risen by a mere 20 to 674. In run number six the number of correct classifications dropped from 674 to 673 however this was soon overcome and the fitness went up again by run number seven to a fitness value of 676 and then continued to rise. The lower distance may have meant that the records misclassified had remained close to the cut-off and further exploring this new region eventually pulled them in again.

Table 10: Ten Iterations of a Hill Climb

Iteration Number	Correctly Classified	Distance	delta
0	568	87.863823	
1	654	18.839607	0.500000
2	674	3.263489	0.500000
3	674	3.263489	0.375000
4	674	3.251822	0.375000
5	674	3.251822	0.281250
6	673	3.251822	0.281250
7	676	1.278676	0.281250
8	677	1.182086	0.281250
9	677	1.182086	0.281250
10	678	1.004053	0.281250

9 A GENETIC ALGORITHM FOR THE FOUR LINE SEGMENT METHOD

Genetic algorithms (GA) are a widely used optimization method that has the advantage that detailed knowledge of the domain space is not

necessary for a GA search and so can be used in many diverse applications [26]. Here a GA will be applied to the four line segment method of estimating the fuzzy measure.

Section 6.5 described the four line segment method where the frequency of the number of records passing each subset in the lattice was used to set the y-coordinates of both elbows. An elbow is a set of coordinates (not on the horizontal axis) where two of the lines meet. In this section the same idea is repeated but this time the x-coordinate of the each elbow is allowed to vary as well.

The x-coordinates at the elbows are set by using number in the range zero to one. These numbers will be parts of a chromosome for the genetic algorithm to be described later in this section. The first elbow needs to be between zero and the expected value of class two, so the expected value of class two is multiplied by a random number in the range zero to one. The x-coordinate for the second elbow is calculated similarly so that it lies between the expected value of class one, and the value that is sum of both frequencies. Optimum values for the multiplying factor for the x-coordinates and the values of the y-coordinates are searched for using a Genetic Algorithm (GA). The structure of the GA will now be discussed.

The GA is based on a steady state design [27], [28]. A type of tournament selection is used. The tournament actually selects the fittest chromosome in the tournament and the least fit. The least fit is subjected to crossover with the fittest and then mutation follows. The new chromosome replaces the least fit chromosome in the population. The tournament size was set to one-third of the population size.

A chromosome consists of two floating point values that represent the y-coordinate values at each elbow and two floating point numbers that are used to calculate the x-coordinate. These four numbers are in the range [0,1].

Mutation And crossover are real valued. Mutation is carried out by the normal distribution method as described in [29]. The old value is mutated by adding another value. The other value is drawn randomly from a normal distribution with mean zero and a pre-assigned standard deviation. Each variable in the chromosome a random decision is made on mutation or not depending upon the pre-set mutation rate.

Crossover is unimodal normal distribution crossover (UNDX) as introduced in [30]. When two floating point numbers are to be crossed there mid-point and the absolute distance apart is calculated. The new value is the mid-point plus the addition of another value. The other value is formed by multiplying the absolute distance by a value drawn randomly from a normal distribution. As in mutation the normal distribution has a mean of zero and a pre-set standard deviation.

The crossover rate was 50% and the mutation rate 12%. Crossover and mutation used the same standard deviation value for their normal distributions. Each time the GA was allowed to run for 1000 generations. The figures quoted are derived from ten-fold cross validation.

Population sizes of 12 and of 48 were tested. For each test the standard deviation was set to 0.2 and then the test was repeated with a standard deviation of 0.4. Experiments included both ways of assessing fitness as discussed in section 8; counting the number of correctly classified records and summing absolute distances to the cut-off. As the GA is a random process each part of the test was repeated five times so that results could be averaged. For each of these five times was a separate ten-fold cross validation. The results are displayed in Tables 11 and 12 where 'pop' is the population and 'SD' means standard deviation.

Table 11: GA with Counting Correct

Settings	Mean of five
Pop 12, SD 0.2	96.30%
Pop 12, SD 0.4	96.37%
Pop 48, SD 0.2	96.55%
Pop 48, SD 0.4	96.55%

Table 12: GA with Distance to Cut-off

Settings	Mean of five
Pop 12, SD 0.2	96.67%
Pop 12, SD 0.4	96.61%
Pop 48, SD 0.2	96.75%
Pop 48, SD 0.4	96.58%

Actually the best single result achieved was 97.22% for a cross validation with a population of 12 and a standard deviation of 0.4 using distance to cut-off. The worst result was 96.05% for across validation with population 12 and also a standard deviation of 0.4 but with count correctly classified records as the way of assessing fitness.

It can be seen that all the results are roughly in the middle between 96% and 97% and similar to the figure of 96.62 found in section 6.5. It seems that generally varying the x-coordinate did not add more efficiency than the earlier method; indeed only two of the experiments achieved a higher result (see Table 12 with standard deviation 0.2). It can also be seen that every time, for the same population size and standard deviation, the distance to cut-off procedure out-performed counting correct classifications. This observation is similar to the hill climbing results of section 8 where similar comparisons were made. Of course, all possible settings of GA could not be tested but it has been shown that the method worked and could potentially be used for modeling with more than four line segments.

10 CONCLUSION

The idea of using frequency counts through the lattice has proved to be very successful and is often better than other methods. One drawback might be that information from the data may not be available for every set in the lattice thus a lot of data may be needed perhaps the technique is most suited to big data sets. However the simple heuristics developed in this paper can always be used to get other techniques a good starting point for further optimisation. The novel idea, that instead of looking at the number of correct classifications, calculate the total distance of all the wrongly classified records to the boundary value, was very successful when applied in parallel to a hill climbing technique that uses small increments and decrements to the fuzzy

measure. Maybe the reason for this is that, as the iterations progress the boundary cut-off value is also self-adjusting. Certainly these fresh approaches will complement other existing methods and deserve further development. Indeed to quote [26] "Comparing different classification methods applied to different data sets, one can find a very interesting fact that sometimes a simple approach outperforms the other methods including those which are very sophisticated".

REFERENCES

[1] Spilde, M., Wang, Z.: Solving nonlinear optimisation problems based on generalised Choquet integrals by using soft computing techniques:, Proc. IFSA 2005, pp. 450--454(2005).

[2] Holland, J. H., (1975),:Adaptation in Natural and Artificial Systems: University of Michigan Press, (1975).

[3] Goldberg, D.E.: Genetic Algorithms in Search, Optimisation and Machine Learning: Addision-Wesley Publishing Company Inc., (1989).

[4] Michalwicz, Z., Fogel, D.B.: How to Solve It - Modern Heuristics: Springer-Verlag, New York, (2000).

[5] Drechsler, R.: Evolutionary Algorithms for *VLSI CAD*: Kluwer Academic Publisher, (1998).

[6] Drechsler,R., Backer, B., and N. Drechsler, (2000) :Genetic algorithm for minimisation of fixed polarity Reed-Muller expressions: *IEE Proceedings Computers Digit. Techniques*, Vol. 147, No. 5, pp. 349--353 (2000).

[7] Adams, K.: Optimisation of Multiple-Valued Logic Polynomials by Polarities and Affine Transforms: Ph.D. Dissertation, University of Ulster, Faculty of Informatics, at Magee College Londonderry, (2007).

[8] Adams, K., Bell, D., Maguire, L.P., McGregor R. J.: Knowledge discovery from decision tables by the use of multiple-valued logic: Artificial Intelligence Review, Vol. 19, No. 2, 2002, pp. 153--176 (2003).

[9] Deng, X., Wang, Z.: Learning probability distributions of signed fuzzy measures by genetic algorithm and multipleregression: Proc. IFSA 2005, pp. 438--444 (2005).

[10] https://archive.ics.uci.edu/ml/datasets /Breast+Cancer+Wisconsin+ (Original).

[11] Choquet, G.:Theory of Capacities: Annales de l'Institut Fourier, 1953,Vol. 5, pp. 131--295, (in French) (1953).

[12] Yang, Z., Yang R., Leung K.: Nonlinear Integrals and their Applications in Data Mining: World Scientific, (2010).

[13] Marichal J.:Entropy of discrete Choquet capacities: IEEE Trans. Fuzzy Syst, Vol. 9, No 1, pp. 164--172 (2001).

[14] Demirel T., Taskan, E.:Multi-criteria evaluation of shifts and overtime strategies using Choquet integral: Proceedings of the World Congress on Engineering, vol. 2, WCE 2012, July 4--6, London U.K (2012).

[15] Hui J., Wang, Z.: Nonlinear Multiregression based on Choquet Integral for Data with both Numerical and Categorical Attributes: Proc. IFSA 2005, pp. 445--449 (2005).

[16] Wending, L., Rendek, J., Masksakis, P.: Selection of suitable decision rules using Choquet integral In: SSPR&SPR 2008, LNCS 5342,: N. D. Vitora Lobo et al. (Eds.): Springer-Verlag Berlin Heidelberg 2008, pp. 947--955 (2008).

[17] Mangasarian O.L., Wolberg W.H.: Cancer diagnosis via linear programming: SIAM News, Vol. 23, No. 5, pp. 1--18 (1990).

[18] William, W.H., Mangasarian, O.L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology: Proceedings of the National Academy of Sciences, U.S.A., Vol. 87, pp. 9193--9196 (1990).

[19] Adams. K.: Simple Heuristics for the Choquet Integral Classifier: Proc. 2nd Int. Conf. on Electrical, Electronics, Computer Engineering and their Applications (EECEA2015), pp.170-179.

[20] Xu, K., Wang, Z., Ke,Y.:Classification by nonlinear integral projections: IEEE Trans. On Fuzzy Systems, 2003, Vol. 1, No. 2, pp 187--201 (2003).

[21] Fogel, D.B., Wasson, E.C., Boughton E.M.: Evolving neural networks for detecting breast cancer: Cancer let. 1995; 96(1): pp. 49--53 (1995).

[22] Abbass, H.A., M. Towaey, M., Finn, G.D.: C-net: a method for generating non-deterministic and dynamic multivariate decision trees: Knowledge Inf. Syst. 2001, No 3: pp. 184--97 (2001).

[23] Abbass H.A.: An evolutionary artificial neural networks approach for breast cancer diagnosis: Artificial Intelligence in Medicine 2002, No. 25, pp. 265--281 (2002).

[24] Aloraina, A.: Different machine learning algorithms for breast cancer diagnosis: International Journal of Artificial Intelligence & Applications (IJAIA) 2012, Vol. 3, No. 6, 2012, pp. 21--30 (2012).

[25] Roberts, A. :Guide to Weka: http://www.andy-roberts.net/teaching/ai32/weka.pdf.

[26] Tabassum , M., Kuruvilla, M.: A Genetic Algorithm Analysis towards Optimization Solutions: Int. J. , Digital and Wireless Communications (IJDIWC), 4(1): pp 124--142, (2014).

[27] DeJong K., Sarma j.,: Generation Gaps Revised In : Whitley, D. (ed.) Foundations of Genetic

Algorithms 2. Morgan-Kaufmann , San Mateo (1993).

[28] Iejdel, B., Kazar, O.: Genetic Agent to Optimize the Automatic Generalization Process of Special Data: Int. J. , Digital and Wireless Communications (IJDIWC), 1(3): pp 693--701,(2011).

[29] Hitoshi, I., Nasimul, N.: New Frontiers in Evolutionary Algorithms: Theory and Applications: Imperial College Press (2012).

[30] Ono, I., Kobayashi, S.: A Real-Coded Genetic Algorithm for Function Optimization using Unimodal Normal Distribution Crossover In: Proc. 7th Int. Conf. on Genetic Algorithms, pp 246--253 (1997).

[31] Marcel, J., Marcel, J. Jr: Separation in Data Mining Based on Fractal Nature of Data: Int. J. , Digital and Wireless Communications (IJDIWC), 3(1): pp 44--60,(2013).