Simple Heuristics for the Choquet Integral Classifier

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

Abstract

The Choquet Integral is a successful classification method. However, like other methods, when applied to large data sets where many coefficients have to be optimised, search methods such as genetic algorithms (GA) are used. In this paper, heuristics are developed that can achieve reasonably accurate results by using information gained from the data set. This enables optimisation using a smaller set of coefficients than those needed in the original search space and this may provide a good starting place for a GA search and other optimisation methods. For the purposes of this research, the data used is the Wisconsin Breast Cancer data set, and it is envisaged that the methods described here will generalise onwards to many other optimisation techniques and data sets.

Key Words: Data mining, Optimisation, Genetic Algorithm, GA, Choquet Integral, Heuristics, Classification, Wisconsin Breast Cancer.

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 Wisconsin Breast Cancer Data Set, §3 Description of the discrete Choquet integral §4 Rationale (of the heuristics) §5 Heuristics Developed §6 Results and Comparisons.

2. 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 [11, 12]. 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)

3 DISCRIPTION OF THE DISCRETE CHOQUET INTEGRAL

The Choquet integral was proposed by Choquet [13] 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 [14] and [15]. The integral takes the form:

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

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

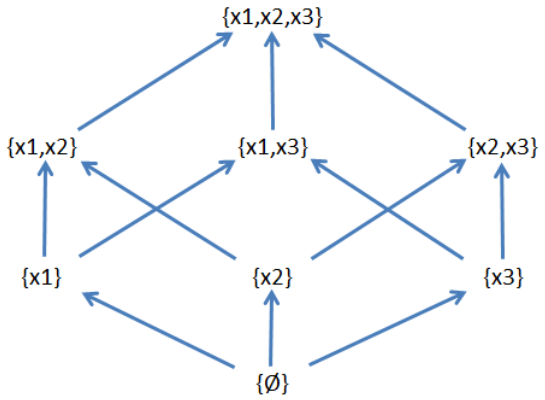

Figure 1 Lattice of Sub-sets

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 [16], for multiple regression [17], and for decision rules [18].

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 x_1 to x_9 and their values recorded in Table 1:

Table 1: 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 2: 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 3: Calculation

Att. x_1 to x_9	**Diff.**	μ	**Diff× μ**
111 111 111	1	1.00000	1.000000

111 110	011	2	0.670496	1.340992
111 010	001	1	0.656327	0.656327
111 010	000	2	0.656327	1.312654
011 010	000	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.

4. 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.

5 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].

5.1 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].

5.2 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 of the measure is the interval [0, 1].

5.3 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.

5.4 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.

5.5 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 shows the example function calculations with all four line segments

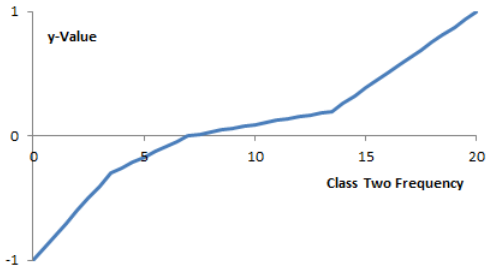

Figure 2: Example with Four Line Segments

6 RESULTS & COMPARISIONS

A reclassification of the entire data set was made. (Reclassification was performed on an artificial data set in [19] as a proof of concept of the Choquet integral classifier.) The reclassification results for this paper are presented in Table 4.

Table 4

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

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.

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 [20, 21, 22] 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 [23]. The method used was 10-fold cross validation and a t-test. The Weka (Waikato Environment for Knowledge Analysis) tool [24], 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 5

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 [23]. 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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

[11] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Vol. 23, Number 5, September 1990, pp. 1-18.

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

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

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

[15] J. Marichal, "Entropy of discrete Choquet capacities", IEEE Trans. Fuzzy Syst, 2001, Vol. 9, No 1, pp. 164–172.

[16] T. Demirel and E. Taskan, "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.

[17] J. Hui, and Z. Wang, "Nonlinear multiregression based on Choquet integral for data with both numerical and categorical attributes", Proc. IFSA 2005, pp. 445-449.

[18] L. Wending, J. Rendek and P. Masksakis, "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.

[19] K. Xu., Z. Wang, Y. Ke, "Calssification by nonlinear integral projections", IEEE Trans. On Fuzzy Systems, 2003, Vol. 1, No. 2, pp 187-201.

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

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

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

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

[24] A. Roberts, Guide to Weka, [on line]. UK. Available from: http://www.andy-roberts.net/teaching/ai32/weka.pdf.