Classification Based on Lower Integral and Extreme Learning Machine

Aixia Chen(✉), Huimin Feng, and Zhen Guo

College of Mathematics and Computer Science, Hebei University, Baoding 071002, China
aixia_chen@163.com

Abstract. It is known that the non-linear integral has been generally used as an aggregation operator in classification problems, because it represents the potential interaction of a group of attributes. The lower integral is a type of non-linear integral with respect to non-additive set functions, which represents the minimum potential of efficiency for a group of attributes with interaction. Through solving a linear programming problem, the value of lower integral could be calculated. When we consider the lower integral as a classifier, the difficult step is the learning of the non-additive set function, which is used in lower integral. Then, the Extreme Learning Machine technique is applied to solve the problem and the ELM lower integral classifier is proposed in this paper. The implementations and performances of ELM lower integral classifier and single lower integral classifier are compared by experiments with six data sets.

Keywords: Non-linear integral · Lower integral · Extreme Learning Machine · Possibility distribution

1 Introduction

Non-additive set functions have described well the potential interaction of a group of attributes. Then, several types of non-linear integrals with respect to non-additive set functions have been defined. In [1], Sugeno generally defined the concept of fuzzy measure and Sugeno integral. In [2], the Choquet integral, which is the extension of the Lebesgue integral, was proposed. And the upper and lower integrals, which are two extreme specified indeterminate integrals, were given by Wang et al. in [3].

Many researchers have attempted to use non-linear integral as an aggregation operator in multi-attribute classification problems [4],[5],[6],[7],[8], and the results are very inspiring. In this kind of approaches, the decisions of different classifiers are fused into a final classification result by a non-linear integral with respect to a non-additive set function, which expresses the weights and the interactions of each classifier for a given class. In the nonlinear classification, Sugeno fuzzy integral and Choquet fuzzy integral have been used in [5],[6],[7]. The upper integral has been used in [13],[14] with the nonlinear classification, and the performance of upper integral classifier is competitive.

X. Wang et al. (Eds.): ICMLC 2014, CCIS 481, pp. 43–51, 2014.
DOI: 10.1007/978-3-662-45652-1_5

In this paper, we attempt to introduce ELM into the lower integral classifier. The lower integral is an extreme decomposition by which the least integration values can be obtained [3]. We know that the decision of non-additive set function is the difficult problem in fuzzy integral classification. As the efficiency and performance of Extreme Learning Machine, we will introduce ELM into the lower integral classifier. In our ELM lower integral classifier, the non-additive set functions are randomly generated and the huge task of learning non-additive set functions is avoided. However, due to the existence of weights β_j, we expect that ELM lower integral classifier has better performance than the single lower integral classifier.

In the following we briefly present the basic firstly. And then the ELM lower integral is proposed. In the end, the comparison of ELM lower integral classifier and single lower integral classifier are provided by experimenting with some data sets.

2 Fuzzy Measure and Integral

Because the feature spaces which we deal with are usually finite, the definitions of fuzzy measure and integrals will be presented in the restrictive case of finite spaces.

Fuzzy measures have been introduced by Sugeno [1].

Definition 1 Assume that X is a non-empty finite set and $\wp$ is the power set of X. The fuzzy measure μ defined on the measurable space $(X,\wp)$ is a set function $\mu:\wp\to[0,1]$, which verifies the following axioms:

$$\mu(\Phi)=0,\mu(X)=1 \tag{1}$$

$$A\subseteq B\Rightarrow\mu(A)\leq\mu(B) \tag{2}$$

$(X,\wp,\mu)$ is said to be a fuzzy measure space.

Definition 2 [11] Assume that $(X,\wp,\mu)$ is a fuzzy measure space, and $X=\{x_1,x_2,\cdots,x_n\}$. Assume that f is a measurable function from X to [0, 1], and without loss of generality, $0\leq f(x_1)\leq f(x_2)\leq\cdots\leq f(x_n)\leq 1$, and $A_i=\{x_i,x_{i+1},\cdots,x_n\}$. The Sugeno integral and the Choquet integral of f with respect to the measure μ are defined as respectively

$$(S)\int fd\mu=\bigvee_{i=1}^{n}(f(x_i)\wedge\mu(A_i)) \tag{3}$$

$$(C)\int fd\mu=\sum_{i=1}^{n}(f(x_i)-f(x_{i-1}))\mu(A_i) \tag{4}$$

where $f(x_0)=0$.

Definition 3 [3] Assume that $X=\{x_1,x_2,\cdots,x_n\}$, and $\wp$ is the power set of X. In this case, any function defined on X is measurable. The lower integral and the upper integral of f with respect to the set function μ can be defined as follows

$$(L)\int fd\mu=\inf\left\{\sum_{j=1}^{2^n-1}\lambda_j\bullet\mu(E_j)\,\middle|\, f=\sum_{j=1}^{2^n-1}\lambda_j\bullet\chi_{E_j},\lambda_j\geq 0\right\} \tag{5}$$

$$(U)\int fd\mu=\sup\left\{\sum_{j=1}^{2^n-1}\lambda_j\bullet\mu(E_j)\,\middle|\, f=\sum_{j=1}^{2^n-1}\lambda_j\bullet\chi_{E_j},\lambda_j\geq 0\right\} \tag{6}$$

where some λ_j may be zero and $E_j, j=1,2,\cdots 2^n-1$ are subsets of X arranged in this way: the binary expression of j, $(j)_2=b_n^{(j)}b_{n-1}^{(j)}\cdots b_1^{(j)}$, is determined by

$$b_i^{(j)}=\begin{cases}1 & if\ x_i\in E_j\\ 0 & if\ x_i\notin E_j\end{cases}\qquad i=1,2,\cdots,n. \tag{7}$$

where, $E_1=\{x_1\},E_2=\{x_2\},E_3=\{x_1,x_2\},E_4=\{x_3\}$, $E_5=\{x_1,x_3\}$, $E_6=\{x_2,x_3\}$, $E_7=\{x_1,x_2,x_3\}$, $\cdots$.

The evaluation of the upper and lower integral is essentially a linear programming problem, when the integrand f and the set function μ are known.

Table 1. The values of set function μ in example 1

Set	Value of μ
Φ	0
$\{x_1\}$	5
$\{x_2\}$	6
$\{x_1,x_2\}$	14
$\{x_3\}$	8
$\{x_1,x_3\}$	7
$\{x_2,x_3\}$	16
$\{x_1,x_2,x_3\}$	18

Example 1 [3] There are three workers x_1, x_2, x_3 working for $f(x_1)=10$ days, $f(x_2)=15$ days and $f(x_3)=7$ days, respectively, to manufacture a kind of products. Their efficiencies of working alone are 5, 6 and 8 products per day, respectively. Their joint efficiencies are not the simple sum of the corresponding efficiencies given above, but are listed in table 1.

It is equivalent to solve this following linear programming problem.

$$\begin{aligned} \min \quad & 5a_1 + 6a_2 + 14a_3 + 8a_4 + 7a_5 + 16a_6 + 18a_7 \\ s.t. \quad & a_1 + a_3 + a_5 + a_7 = 10 \\ & a_2 + a_3 + a_6 + a_7 = 15 \\ & a_4 + a_5 + a_6 + a_7 = 7 \\ & a_j \geq 0, j = 1,2,...,7 \end{aligned} \tag{8}$$

By running the program of the lower integral, we obtain

$$(L)\int f d\mu = 154 \tag{9}$$

with $a_1 = 3, a_2 = 15, a_3 = 0, a_4 = 0, a_5 = 7, a_6 = 0, a_7 = 0$. It is the minimum value of the number of products made by these workers. The corresponding working schedule s x_1 and x_3 work together for 7 days; then x_1 works alone for 3 days and x_2 works alone for 15 days.

As the length of the paper is limited, the properties of fuzzy integral are present in references [1],[2],[3],[4],[11].

3 Classification by Fuzzy Integral

3.1 Possibility Theory

Possibility theory, which is an extension of the theory of fuzzy sets and fuzzy logic, was proposed by L.A.Zadeh in 1978 [8]. It is an uncertainty theory and substitution of probability theory, which is used to deal with the incomplete information. Possibility theory has been used in a number of fields, such as interval analysis, database querying, data analysis, etc.

Definition 4 [9] Let X is a variable, which takes values in the universe of discourse U, and a value of X denoted by u. Informally, a possibility distribution Π_X is a fuzzy relation in U, which acts as an elastic constraint on the values that may be assumed by X, thus, if π_X is the membership function of Π_X, we have

$$Poss\{X = u\} = \pi_X(u), \quad u \in U \tag{10}$$

where the left-hand member denotes the possibility that X may take the value u and $\pi_X(u)$ is the grade of membership of u in Π_X .When it is used to characterize Π_X, the function $\pi_X : U \to [0,1]$ is referred to as a possibility distribution function.

Example 2 Let X is the age of a chairman. Assume that X is a real-valued variable and $55 \leq X \leq 70$. Then, the possibility distribution of X is the uniform distribution defined by

$$\pi_X(u) = \begin{cases} 1 & u \in [55,70] \\ 0 & \text{elsewhere} \end{cases} \tag{11}$$

3.2 Extreme Learning Machine

In the following we will briefly present Single hidden layer feedforward networks (SLFNs)[12].

For l arbitrary distinct sample (x_i, t_i), where $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T \in R^n$, and $t_i = (t_{i1}, t_{i2}, \cdots, t_{im})^T \in R^m$, standard single hidden layer feedforward networks with N hidden nodes and activation function $g(x)$ are mathematically modeled as

$$\sum_{j=1}^{N} \beta_j g(W_j X_i + b_j) = o_i \tag{12}$$

where W_j is the weight connecting the jth hidden node and the input nodes, $\beta_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jm})^T$ is the weight vector connecting the jth hidden node and the output nodes. b_j is the threshold of the jth hidden node. W_j and b_j are randomly generated, and β_j need to be learned.

Theorem 1. Any continuous target function $f(x)$ can be approximated by SLFNs with adjustable hidden nodes. In other words, given any small positive value ε, for SLFNs with enough number of hidden nodes (L) we have

$$\left| f_L(x) - f(x) \right| < \varepsilon \tag{13}$$

3.3 Learning Process

Let us consider the learning process of ELM lower integral classifiers now.

Assume that there are l_j samples $X_1^j, \cdots, X_{l_j}^j$ in class C_j, and similarly for all classes $C_1, \cdots C_m$. We denote $l = \sum_{j=1}^{m} l_j$ as the total number of samples and use indices i, j, k to denote respectively a feature, a class and a sample.

Let X is an unknown sample. The function $\Phi(C_j)$ is said to be the discriminant function, which is described by the possibility distribution $\pi(C_j \mid X)$. Similarly, the function $\phi_i(C_j)$, is called the partial matching degree of X to class C_j with the attribute x_i, which is described by the possibility distribution $\pi(C_j \mid x_i)$. Using Cox's axioms for defining conditional measures, it is known that

$$\pi(C_j \mid x_i) = \pi(x_i \mid C_j) \quad \forall i, j \tag{14}$$

So, we should assign all $\pi(x_i \mid C_j)$ at first.

Learning of the possibility distributions

We will learn a known $\pi(x_i \mid C_j)$ as follows. And we use all the samples in class C_j to construct a "possibilistic histogram". At first, a classical histogram with h boxes

$p_1, \cdots p_h$, here $p_r = n_r / l_j$, with n_r the number of samples in box r , will be constructed from the samples, and the tightest possibility distribution $\pi_1, \cdots \pi_h$ having the same shape as the histogram will be searched. Without loss of generality, assuming that $p_1 \geq \cdots \geq p_h$, this is attained by $\pi_r = \sum_{s=r}^{h} p_s$. At last, we obtain the continuous shape of $\pi(x_i \mid C_j)$ by a linear interpolation of the values π_r .

Learning of lower integral networks

As we know, the determination of the set functions is the difficult problem of nonlinear integral classifiers. In this paper, the Extreme Learning Machine is applied in lower integral classifier to solve the problem. The scheme of ELM lower integral classifier could be briefly described as follows. D is the given training data, and T is the testing data.

(1) For each feature in class C_j samples, we determine the frequency histogram. With continuous feature i , determine h boxes and the corresponding frequencies p_i . With nominal feature i , consider each value of feature i as a box and the corresponding frequencies p_i .

(2)Rearrange the p_i , and determine the possibility distribution π_j^i of each feature. When the feature is continuous, the possibility distribution will be obtained by the linear interpolation.

(3) For l arbitrary distinct sample (x_i, t_i) , where $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T \in R^n$, and $t_i = (t_{i1}, t_{i2}, \cdots, t_{im})^T \in R^m$, standard single hidden layer feedforward networks with N hidden nodes and activation function $g(x)$ are mathematically modeled as

$$\sum_{j=1}^{N} \beta_j g\left((L)\int f(x_i) d\mu_j\right) = o_i \tag{15}$$

where μ_j is the set function connecting the ith hidden node and the input nodes, $\beta_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jm})^T$ is the weight vector connecting the ith hidden node and the output nodes.

It is equivalent to minimize the cost function

$$E = \sum_{i=1}^{l} \left\| \sum_{j=1}^{N} \beta_j g\left((L)\int f(x_i) d\mu_j\right) - t_i \right\| \tag{16}$$

where, $\|\bullet\|$ denotes the norm of the vector.

In the end, one ELM lower integral network is produced for each class. In the method, the set functions are randomly generated and the huge task of learning set functions is avoided. However, due to the existence of weights β_j , the lower integral with the set functions can also show itself effectively and smoothly.

(4) Test the classifier on some data sets.

4 Test on Real Data

In order to investigate how well ELM lower integral classifier works, we conduct an experimental study on some UCI machine learning databases, which are extensively used in testing the performance of different kinds of classifiers. The information about data sets used in our experiments is listed in Table 2.

Table 2. Data used in experiment

Data Set	Number of examples	Number of classes	Number of attributes
Iris	150	3	5
Pima	768	2	9
Wine	178	3	14
Hayes	132	3	6
Ecoli	336	8	8
Tic-tac-toe	958	2	9

10-fold cross validation for 20 times worked at each data set in our experiments. Firstly, we construct the possibility histogram for each feature in the samples of class C_j. The histogram is a graphical data analysis method, which has summarized the distributional information of a variable. If a feature is continuous, the feature is divided into equal sized h boxes (the value of h between 7~15 is appropriate. If the size of samples of class C_j is not large enough, the value is lower). And $p_1, \cdots p_h$ are the frequencies in each box. If the feature is nominal, each feature value is considered as a box. Without loss of generality, assuming that $p_1 \geq \cdots \geq p_h$ and $\pi_r = \sum_{s=r}^{h} p_s$, then $\pi(x_i \mid C_j)$ is given by a linear interpolation of the values π_r.

The comparison of the accuracy between single lower integral and ELM lower integral are shown in table 3.

Then, the ELM lower integral classifier is trained, in which the set functions $\mu_j (j = 1, 2, \cdots, N)$ are randomly generated and weights $\beta_j (j = 1, 2, \cdots, N)$ are learned to minimize the cost function E. Due to the existence of weights β_j, the ELM lower integral classifier can also show itself effectively and smoothly.

Table 3. Comparison of the accuracy between single lower integral and ELM lower integral

Data Set	Lower Integral		ELM Lower Integral	
	Mean	Std Dev	Mean	Std Dev
Iris	0.9517	0.0088	0.9701	0.0079
Pima	0.7562	0.0122	0.7808	0.0094
Wine	0.9517	0.0095	0.9687	0.0071
Hayes	0.6420	0.0243	0.6671	0.0117
Ecoli	0.7495	0.0156	0.7711	0.0110
Tic-tac-toe	0.6501	0.0107	0.6861	0.0079

From Table 3, it can be seen that the ELM lower integral classifier works well for both nominal and continuous attributes. And a comparison of the accuracy between single lower integral and ELM lower integral is present. It is obvious that the performance of ELM lower integral classifier precede the single lower integral classifier. We know the decision of non-additive set function is the difficult problem of fuzzy integral classifiers. In our ELM lower integral classifier, the non-additive set functions are randomly generated. We can see that the computational complexity will be exponentially reduced. However, due to the existence of weights β_j, the ELM lower integral can also show itself effectively and smoothly.

5 Conclusion

In order to effectively use the information of each attribute, and motivated by the effectiveness of ELM, this paper we proposed ELM lower integral classifier. In this approach, the non-additive set functions are randomly generated and the huge task of learning set functions is avoided. So the learning speed of ELM lower integral classifier is very fast. As ELM lower integral classifier takes the weights of importance of individual attributes into account, it is able to model interaction between attributes in a flexible way. From the experimental results, we can see that the performance of ELM lower integral classifier is better than the single lower integral classifier. This paper has demonstrated that the effectiveness of the ELM lower integral classifier, but the relationship between the number of the boxes for continuous feature and the classification performance will be our future investigation.

Acknowledgements. This paper was supported by Project of Hebei Provincial Department of Education (No. Z2012101, No. QN20131055) and the Natural Science Foundation of Hebei University (No. 2010Q26).

References

1. Sugeno, M.: Theory of fuzzy integrals and its applications. Doct. Thesis. Tokyo Institute of Technology (1974)
2. Sugeno, M., Murofushi, T.: Choquet integral as an integral form for a general class of fuzzy measures. In: 2nd IFSA Congress, Tokyo, pp. 408–411 (1987)
3. Wang, Z., Li, W., Lee, K.H., Leung, K.S.: Lower integrals and upper integrals with respect to non-additive set functions. Fuzzy Sets and Systems **159**, 646–660 (2008)
4. Grabisch, M., Sugeno, M.: Multi-attribute classification using fuzzy integral. In: 1st IEEE Int. Conf. on Fuzzy Systems, San Diego, pp. 47–54 (March 8-12, 1992)
5. Tahani, H., Keller, J.M.: Information fusion in computer vision using fuzzy integral. IEEE Trans. SMC **20**(3), 733–741 (1990)
6. Grabisch, M., Nicolas, J.M.: Classification by fuzzy integral. Performance and Tests: Fuzzy Sets and Systems **65**, 255–271 (1994)
7. Grabisch, M.: The application of fuzzy integrals in multicriteria decision making. European Journalof Operational Research **89**, 445–456 (1996)

8. Grabisch, M.: Fuzzy integral in multicriteria decision making. Fuzzy Sets and Systems (Special Issue on 5th IFSA Congress) **69**, 279–298. (1995)
9. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1**, 3–28 (1978)
10. Dubois, D., Prade, H.: Possibility Theory. Plenum Press, New York (1988)
11. Wang, Z., Klir, G.J.: Fuzzy measure theory. New York, Plenum Press (1992)
12. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: Theory and applications. Neurocomputing **70**, 489–501 (2006)
13. Wang, X., Chen, A., Feng, H.: Upper integral network with extreme learning mechanism. Neurocomputing **74**, 2520–2525 (2011)
14. Chen, A., Liang, Z., Feng, H.: Classification Based on Upper Integral. In: Proceedings of 2011 International Conference on Machine Learning and Cybernetics 2, pp. 835–840 (2011)