Nonlinear Integrals and Their Applications in Information Fusion and Data Mining

Zhenyuan Wang

Department of Mathematics
University of Nebraska at Omaha, USA

Extended abstract:

Let $X = \{x_1, x_2, \cdots, x_n\}$ be the set of some attributes in a given database and $f : X \to (-\infty, \infty)$ be a record (or, an observation) of these attributes. The most common aggregation tool in information fusion is the weighted sum

$$y = a_1 f(x_1) + a_2 f(x_2) + \cdots + a_n f(x_n)$$

where real numbers $a_1, a_2, \cdots, a_n$ are weights. This form can be expressed as a classical Lebesgue integral of function *f* on discrete space *X* with respect to an additive measure μ, which is defined on the power set of *X* (denoted by $\mathscr{P}(X)$), determined by $\mu(\{x_i\}) = a_i$, $i = 1, 2, \cdots, n$,

$$y = \int f \, d\mu .$$

Based on such a linear aggregation tool, the linear multiregression model is established as

$$\begin{aligned} y &= a_0 + a_1 f(x_1) + a_2 f(x_2) + \cdots + a_n f(x_n) + N(0, \sigma^2) \\ &= a_0 + \int f \, d\mu + N(0, \sigma^2) , \end{aligned}$$

where the regression coefficients are weights $a_1, a_2, \cdots, a_n$ (the values of measure μ at singletons) and constant a_0, and $N(0, \sigma^2)$ is a normally distributed random perturbation with expected value 0 and variance σ^2. Once a complete data set

x_1	x_2	…	x_n	y
f_{11}	f_{12}	…	f_{1n}	y_1
f_{21}	f_{22}	…	f_{2n}	y_2
⋮				
f_{l1}	f_{l2}	…	f_{ln}	y_l

with the data set sixe *l* not smaller than $n+1$ is given, where the row

$$f_{j1} \quad f_{j2} \quad \cdots \quad f_{jn} \quad y_j$$

is the *j*-th observation of attributes $x_1, x_2, ..., x_n$ and *y*, $j = 1,2, ..., l$, the regression coefficients can be estimated through an algebraic method. Similarly, the linear aggregation tool is also used to form the linear classification models, where the classifying boundary can be expressed as a contour (it is an $(n-1)$-dimensional hyper plane) of the function consisting of the Lebesgue integral, $\int f \, d\mu = c$, with a proper constant *c*. To use either linear multiregression or linear classification model, we need a basic assumption that there is no interaction among considered attributes towards the aggregating target such that the global contribution from attributes is just the simple sum of their individual contributions.

However, there do be such an interaction in most rear-world problems. It is totally different from the co-relationship discussed in probability theory and statistics. In fact, a system being nonlinear, one of the major causes is the existence of the above-mentioned interaction, When the interaction cannot be ignored, it may be described by a nonadditive set function $\mu: \mathscr{F}(X) \to (-\infty, \infty)$ with $\mu(\varnothing) = 0$. As two very special cases, the subadditivity of μ describes an inhibitory interaction among attributes towards the target, while the superadditivity of μ describes a promotive interaction. Thus, the integrals with respect to a nonadditive set function are nonlinear with respect to the integrand generally. The nonlinear integrals are generalizations of the classical Lebesgue integral, that is, they coincide with the Lebesgue integral when the involved set function is additive occasionally. There are several deferent types of nonlinear integrals. The most common type of nonlinear integrals is the Choquet integral $(\mathrm{C})\int f\, d\mu$ defined by

$$(\mathrm{C})\int f\, d\mu = \int_{-\infty}^{0} [\mu(F_\alpha) - \mu(X)]d\alpha + \int_{0}^{\infty} \mu(F_\alpha)d\alpha,$$

where $F_\alpha = \{x \mid f(x) \geq \alpha\}$ for $\alpha \in (-\infty, \infty)$, if two Riemann's integrals in the right-hand side are not both infinite with different signs.

Using the Choquet integral as the aggregation tool, a new nonlinear multiregression model can be established as

$$y = a_0 + \int (a + bf)\, d\mu + N(0, \sigma^2),$$

where $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$, are used to balance the various dimensions of attributes. They should satisfy constraints $\min\limits_{1 \leq i \leq n} a_i = 0$ and $\max\limits_{1 \leq i \leq n} |b_i| = 1$. In this nonlinear model, the regression coefficients are a_0, $a_1, a_2, \cdots, a_n$, $b_1, b_2, \cdots, b_n$, and the values of μ at every set in $\mathscr{F}(X)$ except the empty set. The values of these coefficients can be numerically determined through a soft computing technique, such as a genetic algorithm or the pseudo gradient search, once a proper data set is available. Such a nonlinear multiregression model is a generalization of the linear multiregression. It can mine the information on the interaction among predictive attributes towards the objective attribute.

The Choquet integral can also be used in classification problems. In this case, the classifying boundary is a contour $(\mathrm{C})\int (a + bf)\, d\mu = c$. It is a broken $(n-1)$-dimensional hyper plane with $(n-1)!$ pieces in the *n*-dimensional sample space. Furthermore, if we use a nonlinear polynomial of *f* to replace the linear form *a*+*bf* as the integrand in the Choquet integral, these pieces may be $(n-1)$-dimensional hyper curved surfaces. The unknown parameters in the model can be optimally determined under the criterion of minimizing either the misclassification rate or some kind of total distance once a learning data set is available.

There are many deformations and generalizations of the above basic models in data mining. By dealing with a number of artificial and real world data sets, these new nonlinear models have been shown practicable, effective, and powerful. The details of the relevant theory and applications can be found in book "Nonlinear Integrals and Their Applications in Data Mining" (by Z. Wang, R. Yang, and K.-S. Leung, World Scientific, 2010).

Keywords: Nonadditive set functions, nonlinear integrals, soft computing, information fusion, data mining.