Expert Systems With Applications 44 (2016) 156–167

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Mathematical programming for piecewise linear regression analysis

Lingjian Yang [a], Songsong Liu [a], Sophia Tsoka [b], Lazaros G. Papageorgiou [a,*]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, UK
[b] Department of Informatics, School of Natural and Mahtematical Sciences, King's College London, Strand, London WC2R 2LS, UK

ARTICLE INFO

Keywords:
Regression analysis
Surrogate model
Piecewise linear function
Mathematical programming
Optimisation

ABSTRACT

In data mining, regression analysis is a computational tool that predicts continuous output variables from a number of independent input variables, by approximating their complex inner relationship. A large number of methods have been successfully proposed, based on various methodologies, including linear regression, support vector regression, neural network, piece-wise regression, etc. In terms of piece-wise regression, the existing methods in literature are usually restricted to problems of very small scale, due to their inherent non-linear nature. In this work, a more efficient piece-wise linear regression method is introduced based on a novel integer linear programming formulation. The proposed method partitions one input variable into multiple mutually exclusive segments, and fits one multivariate linear regression function per segment to minimise the total absolute error. Assuming both the single partition feature and the number of regions are known, the mixed integer linear model is proposed to simultaneously determine the locations of multiple break-points and regression coefficients for each segment. Furthermore, an efficient heuristic procedure is presented to identify the key partition feature and final number of break-points. 7 real world problems covering several application domains have been used to demonstrate the efficiency of our proposed method. It is shown that our proposed piece-wise regression method can be solved to global optimality for datasets of thousands samples, which also consistently achieves higher prediction accuracy than a number of state-of-the-art regression methods. Another advantage of the proposed method is that the learned model can be conveniently expressed as a small number of if-then rules that are easily interpretable. Overall, this work proposes an efficient rule-based multivariate regression method based on piece-wise functions and achieves better prediction performance than state-of-the-arts approaches. This novel method can benefit expert systems in various applications by automatically acquiring knowledge from databases to improve the quality of knowledge base.

1. Introduction

In data mining, regression is a type of analysis that predicts continuous output/response variables from several independent input variables. Given a number of samples, each one of which is characterised by certain input and output variables, regression analysis aims to approximate their functional relationship. The estimated functional relationship can then be used to predict the value of output variable for new enquiry samples. Generally, regression analysis can be useful under two circumstances: (1) when the process of interest is a black-box, i.e. there is limited knowledge of the underlying mechanism of the system. In this case, regression analysis can accurately predict the output variables from the relevant input variables without requiring details of the however complicated inner mechanism (Bai, Wang, Li, Xie, & Wang, 2014; Cortez, Cerdeira, Almeida, Matos, & Reis, 2009; Davis & Ierapetritou, 2008; Venkatesh, Ravi, Prinzie, & den Poel, 2014). Quite frequently, the user would also like to gain some valuable insights into the true underlying functional relationship, which means the interpretability of a regression method is also of importance, (2) when the detailed simulation model relating input variables to output variables, usually via some other intermediate variables, is known, yet is too complex and expensive to be evaluated comprehensively in feasible computational time. In this case, regression analysis is capable of approximating the overall system behaviour with much simpler functions while preserving a desired level of accuracy, and can then be more cheaply evaluated (Beck, Friedrich, Brandani, Guillas, & Fraga, 2012; Caballero & Grossmann, 2008; Henao & Maravelias, 2010; 2011; Viana, Simpson, Balabanov, & Toropov, 2014).

Over the past years, regression analysis has been established as a powerful tool in a wide range of applications, including: customer

* Corresponding author. Tel.: +442076792563; fax: +442073882348.
E-mail addresses: lingjian.yang.10@ucl.ac.uk (L. Yang), s.liu@ucl.ac.uk (S. Liu), sophia.tsoka@kcl.ac.uk (S. Tsoka), l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou).

http://dx.doi.org/10.1016/j.eswa.2015.08.034
0957-4174/

demand forecasting (Kone & Karwan, 2011; Levis & Papageorgiou, 2005), investigation of CO_2 capture process (Nuchitprasittichai & Cremaschi, 2013; Zhang & Sahinidis, 2013), optimisation of moving bed chromatography (Li, Feng, P., & Seidel-Morgenstern, 2014b), forecasting of CO_2 emission (Pan, Kung, Bretholt, & Lu, 2014), prediction of acidity constants for aromatic acids (Ghasemi, Saaidpour, & Brown, 2007), prediction of induction of apoptosis by different chemical components (Afantitis et al., 2006) and estimation of thermodynamic property of ionic liquids (Chen, Wu, & He, 2014; Wu, Chen, & He, 2014).

A large number of regression analysis methodologies exist in the literature, including: linear regression, support vector regression (SVR), kriging, radial basis function (RBF) (Sarimveis, Alexandridis, Mazarakis, & Bafas, 2004), multivariate adaptive regression splines (MARS), multilayer perceptron (MLP), random forest, K-nearest neighbour (KNN) and piecewise regressions. We briefly summarise those methodologies before presenting our proposed method.

Linear regression

Linear regression is one of the most classic types of regression analysis, which predicts the output variables as linear combinations of the input variables. The regression coefficients of the input variables are usually estimated using least squared error or least absolute error approaches. The problems can be formulated as either quadratic programming or linear programming problems, which can be solved efficiently. In some cases when the estimated linear relationship fails to adequately describe the data, a variant of linear regression analysis, called polynomial regression, can be adopted to accommodate non-linearity (Khuri & Mukhopadhyay, 2010). In polynomial regression, higher degree polynomials of the original independent input variables are added as new input variables, before estimating the coefficients of the aggregated regression function. Polynomial functions of second-degree have been most frequently used in literature due to its robust performance and computational efficiency (Khayet, Cojocaru, & Zakrzewska-Trznadel, 2008; Minjares-Fuentes et al., 2014).

Another popular variant of linear regression is called least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1994). In LASSO, summation of absolute values of regression coefficients is added as a penalty term into the objective function. The nature of LASSO encourages some coefficients to equal to *0*, thus performing implicit feature selection (Tibshirani, 2011).

Automated learning of algebraic models for optimisation (ALAMO) (Cozad, Sahinidis, & Miller, 2014; Zhang & Sahinidis, 2013) is a mathematical programming-based regression method that proposes low-complexity functions to predict output variables. Given the independent input features, ALAMO starts with defining a large set of potential basis functions, such as polynomial, multinomial, exponential and logarithmic forms of the original input variables. Subsequently an mixed integer linear programming model (MILP) is solved to select the best subset of T basis functions that optimally fit the data. The value of T is initially set equal to *1* and then iteratively increased until the Akaike information criterion, which estimates the generalisation of the constructed model, starts to decrease (Miller et al., 2014). The integer programming model is capable of capturing the synthetic effect of different basis functions, which is considered more efficient than traditional step-wise feature selection.

SVR

Support vector machine is a very established statistical learning algorithm, which fits a hyper plane to the data (Smola & Schlkopf, 2004). SVR minimises two terms in the objective function, one of which is ϵ-insensitive loss function, i.e. only sample training error greater than an user-specific threshold, ϵ, is considered in the loss function. The other term is model complexity, which is expressed as sum of squared regression coefficients. Controlling model complexity usually ensures the model generalisation, i.e. high prediction accuracy in testing samples. Another user-specified trade-off parameter balances the significance of the two terms (Bermolen & Rossi, 2009; Chang & Lin, 2011). One of the most important features that contribute to the competitiveness of SVR is the kernel trick. Kernel trick maps the dataset from the original space to higher-dimensional inner product space, at where a linear regression is equivalent to an non-linear regression function in the original space (Li, Gong, & Liddell, 2000). A number of kernel functions can be employed, e.g. polynomial function, radial basis function and fourier series (Levis & Papageorgiou, 2005). Formulated as a convex quadratic programming problem, SVR can be solved to global optimality.

Despite the simplicity and optimality of SVR, the problem of tuning two parameters, i.e. training error tolerance ϵ and trade-off parameter balancing model complexity and accuracy, and selection of suitable kernels still considerably affect its performance accuracy (Cherkassky & Ma, 2004; Lu, Lee, & Chiu, 2009).

Kriging

Kriging is a spatial interpolation-based regression analysis methodology (Kleijnen & Beers, 2004). Given a query sample, kriging estimates its output as a weighted sum of the outputs of the known nearby samples. The weights of samples are computed solely from the data by considering sample closeness and redundancy, instead of being given by an arbitrary decreasing function of distance (Kleijnen, 2009). The interpolation nature of kriging means that the derived interpolant passes through the given training data points, i.e. the error between predicted output and real output is zero for all training samples. Different variants of kriging have been developed in literature, including the most popular ordinary kriging (Lloyd & Atkinson, 2002; Zhu & Lin, 2010) and universal kriging (Brus & Heuvelink, 2007; Sampson et al., 2013).

MARS

MARS (Friedman, 1991) is another type of regression analysis that accommodates non-linearity and interaction between independent input variables in its functional relationship. Non-linearity is introduced into MARS in the form of the so-called hinge functions, which are expressions with max operators and look like $max(0, X - const)$. If independent variable X is greater than a constant number *const*, the hinge function is equal to *X-const*, otherwise the hinge function equals to *0*. The hinge functions create knots in the prediction surface of MARS. The functional form of MARS can be a weighted sum of constant, hinge functions and products of multiple hinge functions, which makes it suitable to model a wide range of non-linearity (Andrs, Lorca, de Cos Juez, & Snchez-Lasheras, 2011).

The building of MARS usually consists of two steps, a forward addition and a backward deletion step. In the forward addition step, MARS starts from one single intercept term and iteratively adds pairs of hinge functions (i.e. $max(0, X - const)$ and $max(0, const - X)$) that leads to largest reduction in training error. Afterwards, a backward deletion step, which removes one by one those hinge functions contributing insignificantly to the model accuracy, is employed to improve generalisation of the final model (Balshi et al., 2009; Leathwick, Elith, & Hastie, 2006). The presence of hinge functions also make MARS a piece-wise regression method.

MLP

Multilayer perceptron is a feedforward artificial neural network, whose structure is inspired by the organisations of biological neural networks (Hill, Marquez, O'Connor, & Remus, 1994). A MLP typically consists of an input layer of input variables, an output layer of response variables, sandwiching multiple intermediate layers of neurons. The network is fully interconnected in the sense that neurons in each layer are connected to all the neurons in the two neighbour layers (Comrie, 1997; Gevrey, Dimopoulos, & Lek, 2003). Each neuron in the intermediate layers takes a weighted linear combination of outputs from all neurons in the previous layer as input, applies an non-linear transformation function before supplying the output to all neurons of the next layer. The use of non-linear transformation functions, including sigmoid, hyperbolic tangent and logarithmic