Interpolating Support Information Granules

B. Apolloni\textsuperscript{1}, S. Bassis\textsuperscript{1}, D. Malchiodi\textsuperscript{1}, and W. Pedrycz\textsuperscript{2}

\textsuperscript{1} Dipartimento di Scienze dell’Informazione, Università degli Studi di Milano
Via Comelico 39/41, 20135 Milano, Italy
\{apolloni, bassis, malchiodi\}@dsi.unimi.it

\textsuperscript{2} Department of Electrical and Computer Engineering, University of Alberta
ECERF, 9107 - 116 Street, Edmonton, Alberta, Canada T6G 2V4
pedrycz@ee.ualberta.ca

Abstract. We develop a hybrid strategy combining truth-functionality, kernel, support vectors and regression to construct highly informative regression curves. The idea is to use statistical methods to form a confidence region for the line and then exploit the structure of the sample data falling in this region for identifying the most fitting curve. The fitness function is related to the fuzziness of the sampled points and is regarded as a natural extension of the statistical criterion ruling the identification of the confidence region within the Algorithmic Inference approach. Its optimization on a non-linear curve passes through kernel methods implemented via a smart variant of support vector machine techniques. The performance of the approach is demonstrated for three well-known benchmarks.

1 Introductory Comments

This work concerns the use of techniques of granular computing \cite{1} in the refinement of standard regression models \cite{2}. The underlying concept and the design rationale can be concisely outlined in the following manner (see Fig. 1). Given the experimental data, we commonly confine to the linear regression model as the first possible alternative worth exploring. Once accepted, we then focus on the refinement of the model. From the functional standpoint, there are several essential phases reflecting the rationale. First, the confidence region of the preliminary linear model (formed through the use of the confidence curves for some predefined confidence level) eliminates data points falling outside this region. The remaining data are subject to further usage in model building by endowing them with some properties of information granules. We consider the surroundings of those selected points as true information granules and equip them with bell-shaped membership functions similar to those encountered e.g. in radial basis functions (RBF) \cite{3} (see Fig. 2(a)). In our turn, we connect the shape of the bells around points to the mutual relations between these points as it emerges from a suitable clustering of them. Considering the landscape constituted by a norm on the bells, we may look for a regression curve maximizing the integral of this norm along the curve (see Fig. 2(b)).

© Springer-Verlag Berlin Heidelberg 2006
We consider all this to form a dual objective in respect to support vector machines (SVM) [4]. With the latter we try to draw a line passing along the valleys, with the former along the crests of the fitness landscape. In both cases we manage more complex curves by making use of kernel techniques. The main idea is to exploit well established SVM techniques in order to develop an efficient solution. As it is, however, our dual objective has the drawback of not presenting a saddle point as identifier of the optimal solution. This makes harmless the SVM search for the null value of the duality gap. In this study we overcome this drawback by adopting a proper shift trick. Finally, we use the kernel mathematics to deal with non-linear curves as well.

Clear advantage of this procedure relies on the formation of a unifying processing framework that exploits both types of information, namely granular and statistical. This is particularly beneficial as these two are generally viewed to be mutually exclusive. In the literature, indeed we have a huge vein of works on statistical regression theory (refer to [5] and [6] as some representative examples). Also fuzzy regression has gained some visibility, where the drifts of the model with respect to the observed data come within the fuzziness with which the whole the data generation system (the coefficients of the regression line included) can be defined [7,8]. Both approaches start, however, from the general assumption of the existence of the true model, concealed to the humans apart some air-holes releasing sample observations alternatively framed into either an exact though indeterminate framework or into a context not susceptible of sharp computations. On the contrary our starting point is the sample data that we try to organize into operationally suitable descriptions, distinguishing between local information – in the fuzzy sets realm – and global information – in the realm of statistics – that are jointly owned by them. The benefit of the approach is the substantial easiness with which we may integrate many tools separately assessed in the single frameworks.

The paper is organized as follows: Section 2 describes how the regression model is determined, while Section 3 covers some preliminary numerical experiments.