ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Morisita-Based Feature Selection for Regression Problems

Jean Golay, Michael Leuenberger and Mikhail Kanevski

University of Lausanne - Institute of Earth Surface Dynamics (IDYST)
UNIL-Mouline, 1015 Lausanne - Switzerland

Abstract. Data acquisition, storage and management have been improved, while the factors of many phenomena are not well known. Consequently, irrelevant and redundant features artificially increase the size of datasets, which complicates learning tasks, such as regression. To address this problem, feature selection methods have been proposed. This research introduces a new supervised filter based on the Morisita estimator of intrinsic dimension. The algorithm is simple and does not rely on arbitrary parameters. It is applied to both synthetic and real data and a comparison with a wrapper based on extreme learning machine is conducted.

1 Introduction

In data mining, it is often not known a priori what features (or input variables) are truly necessary to capture the main characteristics of a studied phenomenon. This lack of knowledge implies that many of the considered features are irrelevant or redundant. They artificially increase the dimension E of the Euclidean space in which the data are embedded (E equals the number of features). This is a serious matter, since fast improvements in data acquisition, storage and management cause the number of redundant and irrelevant features to increase. As a consequence, unless the sample size N grows exponentially with E, the curse of dimensionality is likely to reduce the overall accuracy of the results yielded by any learning algorithm. Besides, large N and E are also difficult to deal with because of computer performance limitations.

In regression, these issues are often addressed by implementing supervised feature selection methods [1]. They can be broadly subdivided into filters and wrappers. Filters do not use any evaluation criterion involving a learning machine, while wrappers do. Besides, both approaches can be used with search strategies, since an exhaustive exploration of the $2^E - 1$ models (all the combinations of features) is often computationally infeasible. Greedy strategies, such as Sequential Forward Selection (SFS), can be distinguished from randomized ones.

The present paper deals with a new SFS filter algorithm relying on Morisita-based estimates of the intrinsic dimension, M, of data [2, 3]. The Morisita estimator of Intrinsic Dimension (ID) is closely related to the fractal theory and M ($\leq E$) can be interpreted as the dimension of the space where the points of a dataset truly reside (i.e. the data manifold [4]). The proposed algorithm is supervised and designed for regression problems. It does not make use of any threshold, unlike what can be found in related works [5, 6], and it keeps the

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

simplicity of the Fractal Dimension Reduction (FDR) algorithm introduced in [7]. The Morisita estimator of ID is presented in Section 2. Section 3 introduces the Morisita-based filter and Section 4 is devoted to numerical experiments conducted on both synthetic and real data. A comparison with a wrapper combining Extreme Learning Machine (ELM) [8] and an exhaustive search strategy is also carried out.

2 The Morisita Estimator of Intrinsic Dimension

The Morisita estimator of Intrinsic dimension, M_m, is based on the multipoint Morisita index $I_{m,\delta}$ [2, 3] (named after Masaaki Morisita who proposed the first version of the index). $I_{m,\delta}$ is computed by superimposing a grid of Q quadrats of diagonal size δ onto the data points. It measures how many times more likely it is that m ($m \geq 2$) points selected at random will be from the same quadrat than it would be if the N points of the studied dataset were distributed according to a random distribution generated from a Poisson process (i.e. complete spatial randomness). The formula is the following:

$$I_{m,\delta} = Q^{m-1} \frac{\sum_{i=1}^{Q} n_i(n_i-1)(n_i-2)\cdots(n_i-m+1)}{N(N-1)(N-2)\cdots(N-m+1)}$$

where n_i is the number of points in the i^{th} quadrat. For a fixed value of m, $I_{m,\delta}$ is calculated for a chosen range of δ values. If a dataset follows a fractal behavior (i.e. is self-similar), the functional relationship of the plot relating $\log(I_{m,\delta})$ to $\log(1/\delta)$ is linear and the slope is defined as the Morisita slope S_m. Finally, M_m is expressed as:

$$M_m = E - \left(\frac{S_m}{m-1}\right)$$

In practice, each variable is rescaled to $[0, 1]$ and the R chosen δ can be replaced with the edge lengths, ℓ, of the quadrats. In the rest of this paper, only M_2 will be used and it will be computed with an algorithm called Morisita INDex for Intrinsic Dimension estimation (MINDID) [3] whose complexity is $\mathcal{O}(N * E * R)$.

3 The Morisita-based Filter for Regression Problems

The Morisita-Based Filter for Regression (MBFR) relies on three observations following from the works by Traina et al. [7] and De Sousa et al. [5]:

1. Given an output variable Y generated from k relevant and non-redundant input variables X_1, ..., X_k and let $ID(\cdot)$ denote the Intrinsic Dimension (ID) of a dataset, one has that:
$$ID(X_1, \ldots, X_k, Y) - ID(X_1, \ldots, X_k) \approx 0$$
2. Given i irrelevant input variables I_1, ..., I_i completely independent of Y, one has that:
$$ID(I_1, \ldots, I_i, Y) - ID(I_1, \ldots, I_i) \approx ID(Y)$$

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Algorithm 1 MBFR

INPUT: a dataset A with f features $F_{1,\ldots,f}$ and one output variable Y; a vector L of values ℓ; two empty vectors of length f: $SelF$ and $DissF$ for storing, respectively, the name of the selected features and the dissimilarities; an empty matrix Z for storing the selected features. **OUTPUT:** $SelF$ and $DissF$.

1: Rescale each feature and Y to $[0, 1]$.
2: **for** $i = 1$ **to** f **do**
3: **for** $j = 1$ **to** $(f + 1 - i)$ **do**
4: $ID(Z, F_j, Y) - ID(Z, F_j) = Dissimilarity$ (MINDID is used with L)
5: **end for**
6: Store in $SelF[i]$ the name of the F_j yielding the lowest $Dissimilarity$.
7: Store this $Dissimilarity$ in $DissF[i]$.
8: Remove the corresponding F_j from A and add it into Z.
9: **end for**

3. Given a randomly selected subset of X_1, ..., X_k of size r with $0 < r < k$ and $k > 1$, j redundant input variables J_1, ..., J_j related to some or all of X_1, ..., X_r and all the i irrelevant input variables I_1, ..., I_i, one has that:

$$ID(X_1, \ldots, X_r, J_1, \ldots, J_j, I_1, \ldots, I_i, Y)$$
$$-ID(X_1, \ldots, X_r, J_1, \ldots, J_j, I_1, \ldots, I_i) \approx H$$

where $H \in]0, ID(Y)[$ and H decreases to 0 as r increases to k.

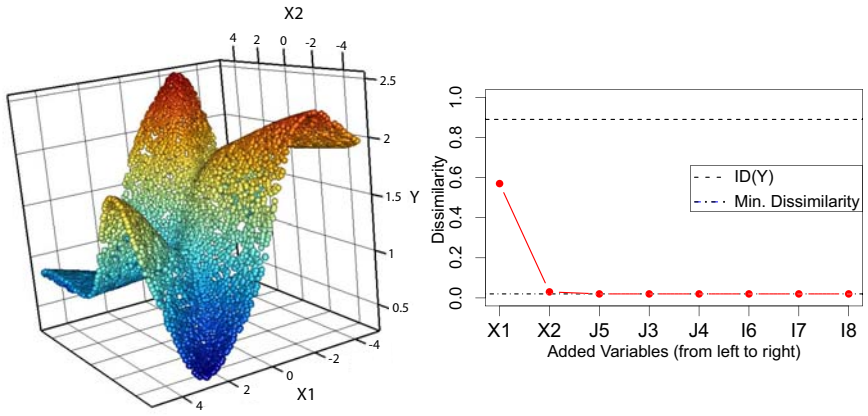

Fig. 1: (left) The functional relationship between the dependent variable Y and the relevant features X_1 and X_2; (right) MBFR applied to 1 simulation of the synthetic dataset.

The difference $ID(\text{features}, Y) - ID(\text{features})$ can thus be suggested as a way of measuring the dissimilarity (i.e the independence) between Y and the selected

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

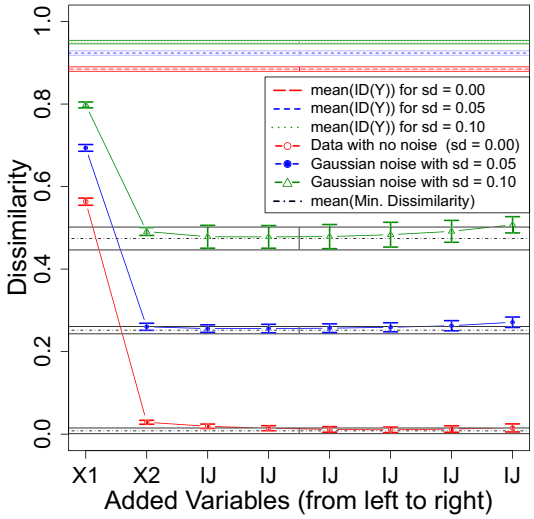

Fig. 2: MBFR applied to simulations of the synthetic dataset with different levels of Gaussian noise (100 simulations per level).

features, among which only the relevant ones (i.e. the non-redundant features on which Y depends) contribute to reducing the dissimilarity. Based on that property, MBFR (See Algorithm 1) aims at retrieving the relevant features available in a dataset by sorting each subset of variables according to its dissimilarity with Y. MBFR implements a SFS search strategy and relies on the MINDID algorithm [3] for the computation of M_2. Its complexity is $\mathcal{O}(N * E^3 * R)$.

4 Experimental Study

4.1 Synthetic Data

The synthetic dataset was constructed as follows. An output variable Y was generated from two uniformly distributed input variables X_1 and X_2 (see Figure 1) by using an Artificial Neural Network (ANN) (one hidden layer of ten neurons, a sigmoid transfer function, randomly generated weights $\in [-2, 2]$, no biases). Three redundant (J) and three irrelevant (I) input variables were also included: $J_3 = log(X_1 + 5)$, $J_4 = X_1^2 - X_2^2$, $J_5 = X_1^4 - X_2^4$, I_6 is uniformly distributed, $I_7 = log(I_6 + 5)$ and $I_8 = I_6 + I_7$. Simulations were generated with $N = 10000$ and with the same ANN weights.

The MBFR algorithm was applied with ℓ^{-1} ranging from 5 to 20. The right panel of Figure 1 shows the result of one simulation run. X_1 and X_2 are easily identified as the relevant features, since they reduce the dissimilarity from $ID(Y)$ to about 0. The same conclusion can be drawn from Figure 2 which presents the effect of three levels of Gaussian noise on three sets of 100 simulations. The minimum dissimilarity increased with the standard deviation

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

of the noise. Eventually, if Y is shuffled (i.e. the dependences with any feature is broken), the dissimilarities stay close to $ID(Y)$ as shown in Figure 3. In Figures 2 and 3, the names of the redundant and irrelevant features were replaced with letters IJ because their rank was not stable over the simulations. It is also worth mentioning that the corresponding dissimilarities do not fluctuate around $ID(Y)$ and this might be related to the increasing dimensionality of the considered space and to the way the dissimilarities are computed (i.e. a subtraction between the ID of two datasets embedded in spaces of different E).

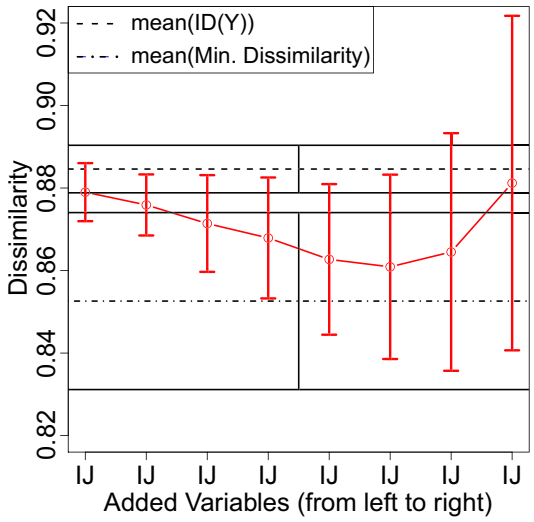

Fig. 3: MBFR applied to 100 simulations of the synthetic data with shuffled Y.

4.2 Real Data

The concrete dataset, downloaded from the UCI machine learning repository, was used. The MBFR algorithm was applied with ℓ^{-1} ranging from 2 to 13 and the selected features turned out to be (see Figure 4): Age, Blast FS, Cement and Superplasticizer. The computation lasted 1.85 seconds using R (Intel Core i7-3770 CPU @ 3.40 GHz with 16.0 GB of RAM under Windows 8).

In order to compare and assess the results obtained with the MBFR algorithm, a wrapper method based on ELM [8] and relying on an exhaustive search was applied: (a) a 5-fold cross-validation was used: 1 fold was iteratively allocated to the set of validation and the remaining 4 folds were assigned to the training set; (b) for each of the $2^E - 1$ subsets of features, ELM models with a number of hidden nodes (the only hyper-parameter of ELM) varying from 1 to 40 were trained and evaluated; (c) the model showing the minimal Mean Squared Error (MSE) was selected. By iterating this process (a to c) 20 times, a total of 100 evaluations of the $2^E - 1$ subsets of features were carried out and the rank of the subset selected by the MBFR algorithm was recorded (for the 100 evalu-

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

ations). The resulting histogram is displayed in Figure 4 and it highlights that the considered subset of features is one of the best from an ELM perspective.

5 Conclusion

MBFR combines the advantages of the existing algorithms of fractal feature selection, while it is designed for regression tasks. It was applied to both synthetic and real datasets. The results are coherent with those yielded by the ELM-based wrapper and with the work by Traina et al. [7] and De Sousa et al. [5]. In future research, MBFR will be applied to challenging datasets and the problem of the sensitivity of the algorithm to the choice of the scale range will be addressed.

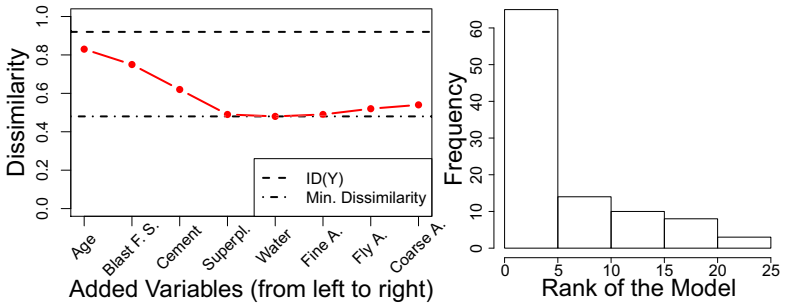

Fig. 4: (left) MBFR applied to the concrete dataset; (right) Histogram of the model ranks based on ELM.

References

[1] I. Guyon, S. Gunn, M. Nikravesh and L. A. Zadeh, editors *Feature Extraction: Foundations and Applications*, Springer, Berlin, 2006.

[2] J. Golay, M. Kanevski, C. D. Vega Orozco and M. Leuenberger, The multipoint Morisita index for the analysis of spatial patterns, *Physica A*, 406:191-202, Elsevier, 2014.

[3] J. Golay and M. Kanevski, A new estimator of intrinsic dimension based on the multipoint Morisita index, *arXiv:1408.0369*, 2014.

[4] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*, Springer, New York, 2007.

[5] E. P. M. De Sousa, C. Traina Jr., A. J. M. Traina, L. Wu and C. Faloutsos, A fast and effective method to find correlations among attributes in databases, *Data Mining and Knowledge Discovery*, 14:367-407, Springer, 2007.

[6] D. Mo and S. H. Huang, Fractal-based intrinsic dimension estimation and its application in dimensionality reduction, *IEEE Transactions on Knowledge and Data Engineering*, 24(1):59-71, IEEE Computer Society, 2012.

[7] C. Traina Jr., A. J. M. Traina, L. Wu and C. Faloutsos, Fast feature selection using fractal dimension. *Proceedings of the* 15^{th} *Brazilian Symposium on Databases*(SBBD 2000), October 15-19 28-30, João Pessoa (Brazil), 2000.

[8] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, Extreme learning machine: Theory and applications, *Neurocomputing*, 70(1-3):489-501, 2006.