

An Inductive Data Mining System Framework

Godfrey Onwubolu

Knowledge Management & Mining, Richmond Hill, Canada

onwubolu_g@dsgm.ca

Abstract. *This paper presents a framework for a unified inductive data mining system based on group method of data handling (GMDH) for modeling, predicting, clustering, and classification for mining fuzzy, noisy and large datasets encountered in real-life applications. In data mining of real-life problems, some major issues that arise include missing data, and very large variables defining the data. For handling missing and noisy datasets, a fast Fourier transform (FFT) signature-based approach integrated with expectation maximization-principal component analysis (EM-PCA) is proposed which automatically reduces large variable dataset to a smaller dimension and consequently results in more a flexible and responsive data mining system for dealing with practical real-life problems. The paper presents a unified inductive data-mining system which is capable of solving data mining functions which differ from existing well known deductive modeling schemes.*

Keywords:

Inductive modeling, GMDH, data mining, FFT-EM-PCA signature-reduction algorithm, complex systems

1 Introduction

Since the formal definitions of knowledge discovery in databases (KDD) as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1],[2], a number of methodologies, languages and software for the standardization of industrial applications of data mining have emerged. Amongst the most widely used by the data mining community are CRISP-DM™ (Cross Industrial Standard Process for Data Mining) and SEMMA (set of functional tools for SAS's Enterprise Miner software) methodologies. Basically, the phases for these KDD methodologies are similar but the six phases for CRISP-DM are as follows:

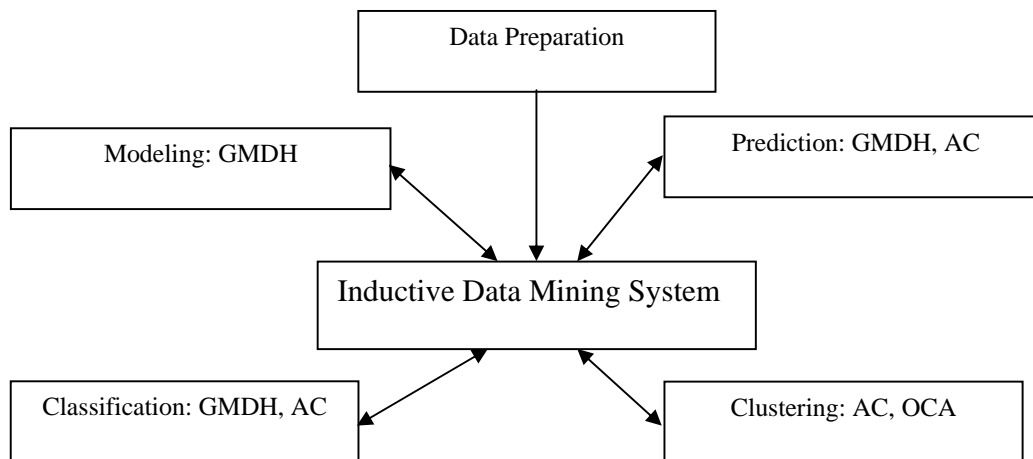
- | | | |
|-----------------------------------|--|-----------------------|
| (1) Business understanding phase. | (4) Data preparation and cleaning phase. | (7) Deployment phase. |
| (2) Build a database. | (5) Data modeling phase. | |
| (3) Data understanding phase. | (6) Evaluation phase. | |

This paper presents a *unified data preparation scheme* (covering phases 2-4) and an *inductive modeling scheme* based on group method of data handling (GMDH) [3] [phase 5] which differs from existing well known deductive modeling schemes (decision trees/rules, regression, cluster analysis, neural networks, association rules, k-nearest neighbor (k-NN), support vector machines (SVM), Bayesian, etc.) used for data mining. Table 1 shows some data mining functions and more appropriate self-organizing (inductive) modeling algorithms and deductive algorithms for addressing these functions. As could be observed, there are mainly three variants of GMDH needed to address most functions of data mining. Consequently, it is easier to design a unified system for variant data mining functions based on GMDH. We refer to this architecture as inductive data mining system as shown in Figure 1; this paper now presents the various aspects of this data mining architecture.

Table 1 Algorithms for self-organizing modeling

Data Mining functions	GMDH Algorithms	Deductive Algorithms
classification	GMDH, AC	Decision trees; Neural networks; k-NN; Naïve Bayes; SVM
clustering	AC ⁺ , OCA ⁺	k-means; spectral clustering; ISODATA
modeling (prediction)	GMDH	CART (classification and regression tree); Regression; Bayesian Belief Networks (BBN); Bayesian Partition model (BPM); Bayesian MARS model; Bayesian multivariate linear splines (MLS); Bayesian Radial Basis (RBF)
time series forecasting	AC, GMDH	CART; Regression; Bayesian versions (as above)
sequential patterns	AC	

⁺Known GMDH-nonparametric model selection methods are: Analog Complexing (AC) and Objective Cluster Analysis (OCA)

**Fig. 1** The inductive data mining architecture

2 Data preparation

2.1 Imputing missing data

It is generally agreed that having a value-added database is essential for data mining applications. Missing data presents problem for the outcome of database application and if imputation of the missing data is implemented, the database will display a better result when utilized for data mining. There are several methods of dealing with data imputation [4],[5] amongst which the following are prominent:

- Mean imputation (MI): This method replaces the missing observation of a certain variable with the mean of the observed values in that variable.
- Regression imputation (RI): missing values are estimated through the application of multiple-regression; variable with missing data is considered as dependent, others variables as predictors.
- Expectation maximization (EM): This method is an iterative two step procedure obtaining the maximum likelihood estimates of a model starting from an initial guess. Each iteration consists of two steps: the expectation (E) step that finds the distribution for the missing data based on the known values for the observed variables and the current estimate of the parameters and the maximization (M) step that replaces the missing data with the expected value.

2.2 EM-PCA approach for simultaneous data imputing and dimensional reduction

An expectation-maximization (EM) algorithm [6] for principal component analysis (PCA) allows a few eigenvectors and eigenvalues to be extracted from large collections of high dimensional data. It is computationally very efficient in space and time; it also naturally accommodates missing information. The key observation is that even though the principal components can be computed explicitly, there is still an EM algorithm for learning them. It can be easily derived as the zero noise limit of the standard algorithms [7, 8] by replacing the usual e-step with the projection above. The algorithm is:

- **e-step:** $X = (C^T C)^{-1} C^T Y$
- **m-step:** $C^{new} = YX^T (XX^T)^{-1}$

where Y is a $p \times n$ matrix of all the observed data and X is a $k \times n$ matrix of the unknown states. The columns of C will span the space of the first k principal components.

2.3 Signature approach for simultaneous data imputing and dimensional reduction

The EM-PCA approach for data imputing may have difficulty dealing with very noisy dataset. Therefore, it is useful to further apply the fast Fourier transform (FFT) which transforms time series data into the frequency domain [8]. In our approach, EM-PCA algorithm is first applied to high dimensional datasets for reduction and dealing with missing data, and the FFT-signature algorithm is further applied for pre-processing very noisy datasets. This EM-PCA-FFT algorithm results in realizing the signatures of the transformed data and further reduces existence of missing information.

3 Framework for Unified Inductive Data Mining

The main components of an inductive modeling and data mining system consist of GMDH-based parametric approaches (MIA, COMBI, Harmonic, OSA) [3] and GMDH-based non-parametric approaches (AC, OCA) [3] shown in Figure 1. In this paper, the connections that exist between GMDH-based non-parametric approaches (AC, OCA) are highlighted and researchers are encouraged to explore the synergy.

3.1 Parametric GMDH

Parametric GMDH variants (MIA, COMBI, Harmonic) [3] is the basis for modeling complex environment. However, due to their deficiencies, hybrid-GMDH approaches based on synergy of GMDH and computational intelligence methods [15] have resulted in significant enhancements which traditional GMDH cannot attain.

3.2 Analog Complexing (AC)

Analog Complexing [3],[10-13] can be considered as a sequential pattern recognition method for stepwise predicting and qualitatively explaining fuzzy objects or multidimensional random processes inherently by complexing (weighted addition) of analogues (similar patterns) taken from historical data. The Analog Complexing algorithm appropriated for forecasting multidimensional stochastic processes is described as a four-step-procedure [12]:

- Step 1: Generation of alternate patterns,
- Step 2: Transformation of analogues,
- Step 3: Selection of most similar patterns,
- Step 4: Combining forecasts.

3.3 Objective Cluster Analysis (OCA)

Objective cluster analysis (OCA) first uses *dipoles* to divide sample data into two subsets, A and B in order to look for the optimal number of clusters. Then it evaluates the consistency of clustering schemes on sets A and B with consistency criterion η_c . The basic steps in OCA are given as follows [3]:

Step 1: Compute the distance samples x_i and x_j

Step 2: Partition the data sample into subsets A and B used as training set and subsets C and D , used as testing set.

Step 3: Clustering

Step 4: Determine the unique optimal clustering scheme

Considering Step 3 of the AC process, it is observed that clustering (OCA) activity is involved. Therefore, for an inductive data mining system, it is recommended that AC and OCA should be integrated to carry out the functions of predictions for short-term data-points as well as the functions of determining the number of clusters and their membership in an a priori manner for new unfamiliar datasets using OCA.

4 Experimentation

Case Study 1: Modeling using GMDH

The case study is based on Breast cancer dataset [14]. This is a 2 class problem with 78 training samples (34 relapse and 44 non-relapse) and 19 testing samples (12 relapse and 7 non-relapse) of relapse and non-relapse. The dimension of breast cancer dataset is 24481. It is not possible to apply traditional GMDH to this type of problem. The hybrid-GMDH approach [15] was employed for classification of the problem, resulting in 100% accuracy.

Case Study 2: Prediction using AC

This case study uses the German economy dataset [12] which is made up of 30 observations and 13 variables. Based on Section 3.2, an Analog complexing (AC) methodology was developed and Figure 2 shows the predictions for x_7 variable (gross domestic product); the last 3 points are predictions.

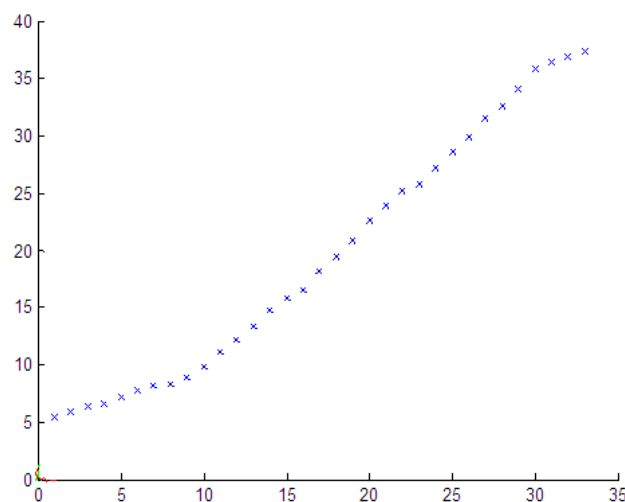


Fig. 2 Analog complexing predictions for x_7 variable (gross domestic product)

Case Study 3: Clustering using OCA

This case study uses the dataset for detecting cyclic disturbances in supply networks [9] which is made up of 72 observations and 43 variables; this dataset is very noisy since it has significant number of voids or missing data. In this paper, we propose a two-stage approach consisting of EM-PCA (section 2.2) and the FFT-signature (section 2.3) components. From the experimentation, it was found that the EM-PCA algorithm could not completely solve the missing data problem. However, by employing the FFT-signature algorithm, the signatures of Figure 3 were realized and the problem of missing data was completely solved. From this figure, it is easy to speculate which variables could be clustered together although manual method of clustering cannot give accurate results. The available data comprise monthly time series data for 72 months and 43 tags or features. Figure 3 shows the power spectra of the pre-processed supply network data. The horizontal axis of the spectra is a normalized frequency axis and represents the sampling frequency. The objective is to identify and find the appropriate number of clusters of the features. In this case any dimensional reduction that needed to be done had to be in the other direction (time/month). Reducing the size of the time/reduction from 72 to 5 using EM-PCA algorithm, applying the FFT-signature algorithm for pre-processing, the objective cluster algorithm (OCA) finds *ab initio*, the clusters needed for practical applications.

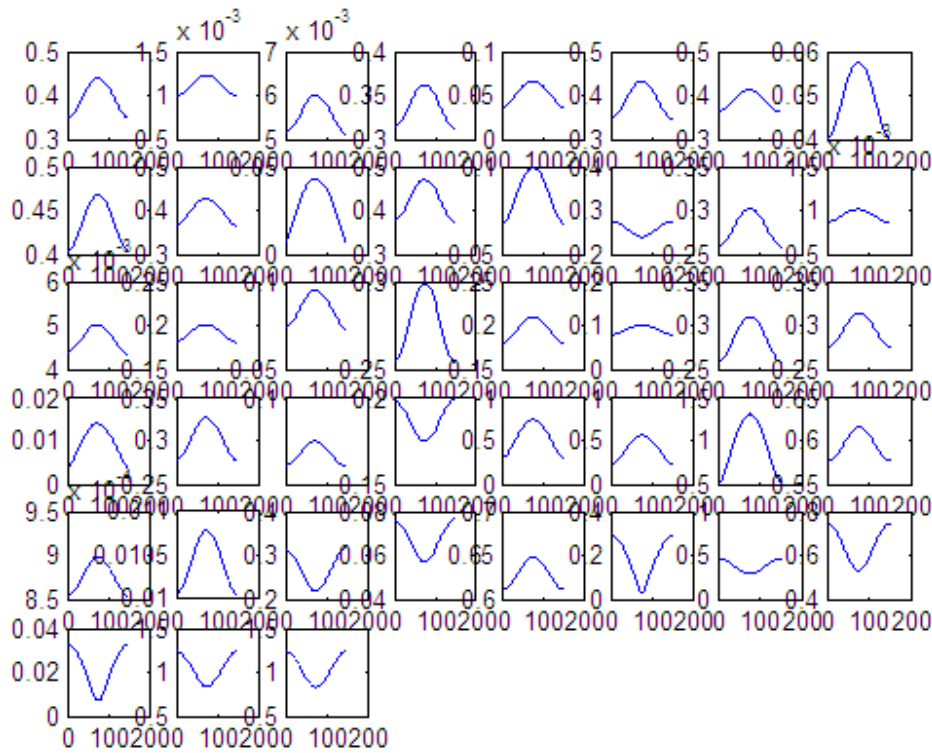


Fig. 3 Signatures for features of dataset

5 Conclusions

This paper presents a framework for realizing a unified inductive data mining system based on group method of data handling (GMDH). Important components should include data preparation, ability to handle long- and short-data samples for modeling and prediction, and clustering. It is further recommended that there should be integration of

AC and OCA sub-modules, which is currently not the case. The paper proposes a signature-based method coupled with EM-PCA for data preparation of very noisy datasets.

Acknowledgement

S.S. Dimov and A.A. Afify of the Manufacturing Engineering Centre, Cardiff University Innovative Manufacturing Research Centre, Cardiff, UK are gratefully appreciated for providing us with the dataset for their earlier investigation on detecting cyclic disturbances in supply networks.

References

- [1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., From data mining to knowledge discovery in databases, American Association for Artificial Intelligence, 1996, 37-53.
- [2] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge, Mass., 1996.
- [3] Madala, H. R., and Ivakhnenko, A. G., Inductive Learning Algorithms for Complex Systems Modeling, *CRC Press Inc.*, 1994, p.384.
- [4] Litle, R. J.A., and Rubin, D.B., *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, 1987
- [5] Yeh, R-L, Liu, C., Shia, B-C., Cheng, Y-T., Huwang, Y-F., Imputing manufacturing material in data mining, *J. Intelligent Manufacturing*, 19, 2008, 109-118.
- [6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B*, 39:1—38, 1977.
- [7] B. S. Everitt. *An Introduction to Latent Variable Models*. Chapman and Hill, London, 1984.
- [8] Zoubin Ghahramani and Geoffrey Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96- 1, Dept. of Computer Science, University of Toronto, Feb. 1997.
- [9] Afify, A. A, Dimov, S. S., Naim, M., Valeva, V., Shukla, V., Data mining: a tool for detecting cyclic disturbances in supply networks, *Proc. IMechE Vol. 221 Part B: J. Engineering Manufacture*, 2007, 1771-1785.
- [10] Lorence, E., N., Atmospheric predictability is revealed by naturally occurring analogues, *J. Atmospheric Science*, 1969, No. 4, pp 636-646
- [11] Lemke F., Mueller J.-A., Self-Organizing Data Mining for a portfolio trading system, *Journal of Computational Intelligence in Finance*, 26(3), 1997, 12-26.
- [12] Mueller J.-A., Lemke F. Self-Organizing Data Mining. Extracting Knowledge From Data. *Trafford Publishing, Canada*, 2003.
- [13] Ivakhnenko, G., Short-term processes forecasting by analogues complexing GMDH algorithm, *Proceedings of 2nd International Conference on Inductive Modeling 2008*, September 15-19, 2008, Kyiv, Ukraine, 241-245.
- [14] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D, Hart, A.M.H, Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature, Nature*, vol. 415, pp. 530-536, 2002.
- [15] Onwubolu, G. C., (ed.), *Hybrid Self-Organizing Modeling Systems*, Springer-Verlag, Heidelberg, Germany, 2009: <http://www.springer.com/engineering/book/978-3-642-01529-8>