Review Article ISSN: 2581-5970 Open Access

UPI JOURNAL OF ENGINEERING AND TECHNOLOGY

Journal home page: https://uniquepubinternational.com/upi-journals/upi-journal-of-engineering-and-technology-upi-jet/

Clustering Algorithms and Bayesian Networks for Distributed Geospatial Data Mining and Knowledge Discovery

Diego Liberati

Italian National Research Council, Italy.

Abstract

A few consolidated methods of data mining approaches developed by ourselves are proposed in the modern framework of geoinformatic remote sensing. These approaches and combinations of them, be it partially or fully, helps in extracting knowledge from huge data sets especially as in geo-informatics. These methods of data mining are quite useful in depicting trends and patterns associated with huge amount of partially correlated data generated at various stations and classifying them based on different variables associated with the process which can also be in a non-linear fashion. These methods are successfully applied individually in various contexts. We suggest that combinations of these approaches when worked upon yield an effective classification of data even in the complicated and distributed field of geo-informatics.

Key words: Bayesian Networks, Unsupervised clustering, Data mining, Algorithms.

1. Introduction

The initiative Global Monitoring of Environment and Security (GMES) jointly set up by European Union (EU) and European Space Agency, under the EU directive Infrastructure for Spatial Information in Europe (INSIPRE) has been one of the stimulating international projects putting together geoscientists with different expertise in order to joint effort in sharing detection tools, locations, methods and co-workers, to make a quality jump in the sector.

In particular, given the opportunity to have many data on several possible locations, one of the typical goals one has in mind is to classify records on the basis of a hopefully reduced meaningful subset of the measured variables.

The complexity of the problem makes it worthwhile to resort to automatic classification procedures.

Then, the questions do arise of reconstructing a

How to cite: Liberati D. Clustering algorithms and bayesian networks for distributedgeospatial data mining and knowledge discovery. UPI Journal of Engineering and Technology 2018; 1(1): 1-7.

Article history:
Received: 10-03-2018, Accepted: 17-03-2018, Published: 17-03-2018

Correspondence to: Diego Liberati, Preventive Italian National Research Council, Italy.
Email: diego.liberati@gmail.com

synthetic mathematical model, capturing the most important relations between variables. Such interrelated aspects will be the focus of the present contribution. Four main general purpose approaches, also useful in the geo-informatics context, will be briefly discussed in the present chapter, underlying cost effectiveness of each one.

In order to reduce the dimensionality of the problem, thus simplifying both the computation and the subsequent understanding of the solution, the critical problems of selecting the most salient variables must be solved.

A very simple approach is to resort to cascading a Divisive Partitioning of data orthogonal to the Principal Directions – PDDP – [1] already proven to be successful in contexts different for applications, like analyzing the logs of an important telecom provider [2] or identifying the salient genes discriminating Leukemia in micro-array samples [3].

A more sophisticated possible approach is to resort to a rule induction method, like the one described in Muselli and Liberati [4]. Such a strategy also offers the advantage to extract underlying rules, implying conjunctions and/or disjunctions between the identified salient variables. Thus, a first idea of their even non-linear relations is provided as a first step to design a representative model, whose variables will be the selected ones. Such an approach has been shown [5] to be not less powerful over several benchmarks than the popular decision tree developed by Quinlan [6].

An alternative in this sense can be represented by Adaptive Bayesian Networks [7] whose advantage is also to be available on a commercial wide spread data base tool like Oracle (www.oracle.com). A possible approach to blindly build a simple linear approximating model is to resort to piece-wise affine (PWA) identification [8].

The joint use of (some of) such four approaches, briefly described in the present contribution, starting from data without known priors about their relationships, will thus allows to reduce dimensionality without significant loss in information, then to infer logical relationships, and, finally, to identify a simple input-output model of the involved process that also could be used for controlling purposes even in a complex field like geo-informatics.

The possibility to resort to such services in a distributed way, when not all of them are available at the geoscientist nor at the same place anyway, thus sharing not only data collected at different places in different modalities, but also complementary expertise in different approaches, do lead to a cooperative enterprise in the growing paradigm of e-Science [9].

1.1. Background

The introduced tasks of selecting salient variables, identifying their relationships from data and classifying possible intruders may be sequentially accomplished with various degrees of success in a variety of ways.

Principal components order the variables from the most salient to the least one, but only under a linear framework.

Partial least squares do allow extending to non-linear models, provided that one has prior information on the structure of the involved non-linearity; in fact, the regression equation needs to be written before identifying its parameters. Clustering may operate even in an unsupervised way without the a priori correct classification of a training set [1].

Neural networks are known to learn the embedded rules with the indirect possibility [10] to make rules explicit or to underline the salient variables.

Decision trees [6] are a popular framework providing a satisfactory answer to the recalled needs.

2. Experimental

2.1. Salient Variable Selection

A learning strategy that looks for a trade-off between a high predictive accuracy of the classifier and a low cardinality of the selected variable subset may be derived according to the central hypothesis that a good variable subset contains variables that are highly correlated with the class to be predicted, yet uncorrelated with each other.

Based on information theory, the Minimum Description Length (MDL) principle [11] states that the best theory to infer from training data is the one that jointly minimizes the length (i.e. the complexity) of the theory itself and the length of the data encoded with respect to it. Thus, MDL can be employed as a criterion to judge the quality of a classification model, by finding a compact encoding of the training data [12]. As described in [13], each feature can be ranked according to its description length that reflects the strength of its correlation with the target. In this context, the MDL measure is again given by weighting the encoding length with the number of bits needed to describe the data [7].

Once all variables have been ordered by rank, no a priori criterion is available to choose the cut-off point beyond which variables can be discarded. One can thus start by building a classifier on the set of the n-top ranked features via one of the following approaches. Then, a new feature is sequentially added to this set, and a new classifier is built, until no improvement in accuracy is achieved.

2.2. Unsupervised Clustering

The approach taken herein may be summarized in the following three steps, the second of which is the core of the method, while the first one constitutes a pre-processing phase useful to ease the following task, and the third one is a post-processing step designed to focus back on the original variables. Then:

1. A Principal Component Analysis (PCA) [14-15] defines a hierarchy in the transformed orthogonal variables according the principal directions of the data set. It is a multivariate analysis designed to select the linear combinations of variables with higher inter-subject co-variances; such combinations are the most useful for classification. More precisely, PCA returns a new set of orthogonal coordinates of the data space, where such coordinates are ordered in decreasing order of inter-subject covariance.

2. The unsupervised clustering is performed by cascading a non-iterative technique-the Principal Direction Divisive Partitioning (PDDP) [1] based upon singular value decomposition [16] and the iterative centroid-based divisive algorithm k-means [17]. Such a cascade, with the clusters obtained via PDDP used to initialize k-means centroids, is shown to achieve best performances in terms of both quality of the partition and computational effort [18]. The whole dataset is thus bisected into two clusters, with the objective of maximizing the distance between the two clusters and, at the same time, minimizing the distance among the data points lying in the same clusters. The classification is achieved without using a priori information (unsupervised learning) thus automatically highlighting data belonging to a (possibly unknown) class.

3. By analyzing the obtained results, the number of variables needed for the clustering may be reduced, by pruning all the original variables that are not needed in order to define the final partitioning hyper-plane, so that the classification eventually is based on a few variables only.

Binary rule inference and variable selection while mining data via logical networks

Recently, an approach has been suggested-Hamming Clustering-related to the classical theory exploited in minimizing the size of electronic circuits, with the additional care to obtain a final function able to generalize from the training dataset to the most likely framework describing the actual properties of the data. In fact, the Hamming metric tends to cluster samples whose code is less distant; this is likely to be natural, if variables are redundantly coded via thermometer (for numeric variables) or only-one (for logical variables) code [4].

The approach followed by Hamming clustering in mining the available data to select the salient variables and to build the desired set of rules consists of the three following steps:

Step 1: A critical issue is the partition of a possibly continuous range in intervals, whose number and limits may affect the final result. The thermometer code may be used to preserve ordering and distance (in case of nominal input variables, for which a natural ordering cannot be defined; the only-one code may instead be adopted). The simple metric used is the Hamming distance, computed as the number of different bits between binary strings. In this way, the training process does not require floating point computation but only basic logic operations. This is one reason for the algorithm speed and for its insensitivity to precision.

Step 2: To generalize and infer the underlying rules during the logical synthesis designed to obtain the simplest AND-OR expression able to satisfy all the available input-output pairs, at every iteration Hamming clustering groups together in a competitive way binary strings having the same output and close to each other. A final pruning phase does simplify the resulting expression, further improving its generalization ability. The minimization of the involved variables intrinsically excludes the redundant ones, thus enhancing the very salient variables for the investigated problem. The low (quadratic) computational cost allows managing quite large datasets.

Step 3: Each logical product directly provides an intelligible rule, synthesizing a relevant aspect of the searched underlying system that is believed to generate the available samples [5].

2.3. Adaptive Bayesian Networks

Naïve Bayes (NB) is a very simple Bayesian network consisting of a special node (the target class) that is parent of all other nodes (the variables) that are assumed to be conditionally independent, given the value of the class. The NB network can be supervisedly quantified against a training dataset of pre-classified instances, by computing the probability associated to a specific value of each variable, given the value of the class label. Then, any new instance can be easily classified making use of the Bayes rule. Despite its strong independence assumption is clearly unrealistic in several application domains; NB has been shown to be competitive with more complex state-of-the-art classifiers [12, 19-20].

In order to relax NB full independence assumption, correlation arcs are added between the variables of a NB classifier, still imposing specific structural constraints [12, 19] in order to maintain computational simplicity on learning. The Adaptive Bayesian Network (ABN) algorithm [7], is a greedy variant, based on MDL, of the approach proposed in [19]: the network is initialized to NB on the top k ranked variables according to their MDL relevance. Next, the algorithm attempts to extend NB by constructing a set of tree over multi-dimensional variables. Interestingly, each multi-dimensional feature can be expressed in terms of a set of if-then rules enabling users to easily understand the basis of model predictions

2.4. Piece-wise Affine Identification

Once the salient variables have been selected, it may be of interest to capture a model of their dynamical interaction. A first hypothesis of linearity may be investigated, usually being only a very rough approximation, when the values of the variables are not close to the functioning point around which the linear approximations computed.

On the other hand, to build a non-linear model is far from easy; the structure of the non-linearity needs to be a priori known, which is not usually the case. A typical approach consists of exploiting a priori knowledge, when available, to define a tentative structure, then refining and modifying it on the training subset of data, and finally retaining the structure that best fits a cross-validation on the testing subset of data. The problem is even more complex when the collected data exhibit hybrid dynamics (i.e., their evolution in time is a sequence of smooth behaviours and abrupt changes).

An alternative approach is to infer the model directly from the data without a priori knowledge via an identification algorithm capable of reconstructing a very general class of piece-wise affine model [8]. This method also can be exploited for the data driven modelling of hybrid dynamical systems, where logic phenomena interact with the evolution of continuous-valued variables. Such approach will be described concisely in the following

Piece-wise affine identification exploits k-means clustering that associates data points in multivariable space in such a way to jointly determine a sequence of linear sub-models and their respective regions of operation without even imposing continuity at each change in the derivative. In order to obtain such a result, the five following steps are executed:

Step 1: The model is locally linear; small sets of data points close to each other likely belong to the same sub-model. For each data point, a local set is built, collecting the selected points together with a given number of its neighbours (whose cardinality is one of the parameters of the algorithm). Each local set will be pure if made of points really belonging to the same single linear subsystem; otherwise, it is mixed.

Step 2: For each local dataset, a linear model is identified through usual least squares procedure. Pure sets belonging to the same sub-model give similar parameter sets, while mixed sets yield isolated vectors of coefficients, looking as outliers in the parameter space. If the signal to noise ratio is good enough, and if there are not too many mixed sets (i.e., the number of data points is enough more than the number of sub-models to be identified, and the sampling is fair in every region), then the vectors will cluster in the parameter space around the values pertaining to each sub-model, apart from a few outliers.

Step 3: A modified version of the classical K-means, whose convergence is guaranteed in a finite number of steps [8], takes into account the confidence on pure and mixed local sets in order to cluster the parameter vectors.

Step 4: Data points are then classified, each being a local dataset one-to-one related to its generating data point, which thus is classified according to the cluster to which its parameter vector belongs.

Step 5: Both the linear sub-models and their regions are estimated from the data in each subset. The coefficients are estimated via weighted least squares, taking into account the confidence measures. The shape of the polyhedral region characterizing the domain of each model may be obtained via linear support vector machines [21], easily solved via linear/quadratic programming.

3. Future Trends

The proposed approaches are now under application in several contexts. The fact that a combination of different approaches, taken from partially complementary disciplines, proves to be effective may indicate a fruitful direction in combining in different ways classical and new approaches to improve classification even in the complex and often distributed field of geo-informatics.

4. Conclusion

The proposed approaches are very powerful tools for quite a wide spectrum of applications in and beyond data mining, providing an up-to-date answer to the quest of formally extracting knowledge from data and sketching a model of the underlying process.

In geo-informatics such tools may be quite useful in order to complement other approaches in processing the huge amount of partially correlated data made available at various stations with complementary sensors, and classifying them on the basis of their identified even non linear profile.

5. Conflicts of Interests

The author(s) report(s) no conflict(s) of interest(s). The author along are responsible for content and writing of the paper.

6. Acknowledgments

NA

7. References

1. Booley DL. Principal direction divisive partitioning. Data Mining and Knowledge Discovery 1998; 2(4): 325-344.

2. Garatti S, Savaresi S, Bittanti S. On the relationships between user profiles and navigation sessions in virtual communities: a data-mining approach. Intelligent Data Analysis 2004; 8(6): 576-600

3. Garatti S, Bittanti S, Liberati D, Maffezzoli P. An unsupervised clustering approach for leukemia classification based on DNA micro-arrays data. Intelligent Data Analysis 2007; 11(2): 175-188.

4. Muselli M, Liberati D. Training digital circuits with Hamming clustering. IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications 2000; 47: 513-527.

5. Muselli M, Liberati D. Binary rule generation via Hamming clustering. IEEE Transactions on Knowledge and Data Engineering 2002; 14: 1258-1268.

6. Quinlan JR. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann, 1994.

7. Yarmus JS. ABN: A Fast, Greedy Bayesian Network Classifier:http://otn.oracle.com/products/bi/pdf/adaptive_bayes_net.pdf, accessed on 09-03-2018.

8. Ferrari-Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. Automatica 2003; 39: 205-217.

9. Bosin A, Dessì N, Fugini MG, Liberati D, and Pes B. Applying Enterprise Models to Design Cooperative Scientific Environments, Lecture Notes in Computer Science, Volume 3812, 2006, pp. 281 – 292.

10. Taha I, Ghosh J. Symbolic interpretation of artificial neural networks. IEEE Transactions on Knowledge and Data Engineering 1999; 11: 448-463.

11. Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modelling. IEEE Transactions on Information Theory 1998; 44: 2743-2760.

12. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning 1997; 29: 131-161.

13. Kononenko I. On biases in estimating multi-valued attributes. International Journal of Computer Applications 1995; 95: 1034-1040.

14. O'Connel MJ. Search program for significant variables. Computer Physics Communications 1974; 8: 49.

15. Hand D, Mannila H, Smyth P. Principles of data-mining. Cambridge, MA: MIT Press, 2001.

16. Golub GH, Van Loan CF. Matrix computations. Johns Hopkins University Press, 1996.

17. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, 1967.

18. Savaresi SM, Boley DL. A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. International Journal on Intelligent Data Analysis 2004; 8(4): 345-363.

19. Keogh E, Pazzani MJ. Learning the structure of augmented Bayesian classifiers, International Journal on Artificial Intelligence Tools 2002; 11(4): 587-601.

20. Cheng G, Greiner R. Comparing Bayesian Network Classifiers, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, Inc., San Francisco, 1999.

21. Vapnik V. Statistical learning theory. New York: Wiley, 1998.

8. Key Terms and Their Definitions

8.1. Hamming Clustering: A fast binary rule generator and variable selector are able to build understandable logical expressions by analyzing the Hamming distance between samples.

Hybrid Systems: Their evolution in time is composed by both smooth dynamics and sudden jumps.

8.2. k-means: Iterative clustering technique subdividing the data in such a way to maximize the distance among centroids of different clusters, while minimizing the distance among data within each cluster. It is sensitive to initialization.

Model Identification: Definition of the structure and computation of its parameters best suited to mathematically describe the process underlying the data.

8.3. PDDP (Principal Direction Divisive Partitioning): One-shot clustering technique based on principal component analysis and singular value decomposition of the data, thus partitioning the dataset according to the direction of maximum variance of the data. It is used here in order to initialize K-means.

8.4. Principal Component Analysis: Rearrangement of the data matrix in new orthogonal transformed variables ordered in decreasing order of variance.

Rule Inference: The extraction from the data of the embedded synthetic logical description of their relationships.

8.5. Salient Variables: The real players among the many apparently involved in the true core of a complex business.

8.6. Singular Value Decomposition: Algorithm able to compute the eigenvalues and eigenvectors of a matrix; also used to make principal components analysis.

8.7. Unsupervised Clustering: Automatic classification of a dataset in two of more subsets on the basis of the intrinsic properties of the data without taking into account further contextual information.