NEURO FUZZY APPROACH TO DATA CLUSTERING: A FRAMEWORK FOR ANALYSIS

Farhat Roohi, M.Phil, MSc.

Dept of Electronics And IT, University of Kashmir Srinagar J&K, India

Abstract

Clustering and pattern recognition have been used for various purposes since times immemorial. However, with the advancements in computing technology and information and communication technology clustering has achieved great significance in data analysis, data mining, knowledge management, artificial intelligence, control processes, image recognition, etc. The ocean of valuable data which is being added every day to the already mountains of data poses a big challenge to the world of computing, as it is practically impossible to handle such a large quantity of data individually. As this data contains valuable hidden information in it, therefore, needs to be understood not only to solve human problems but also to avoid many disasters for betterment of human welfare. Therefore, there is pressing need for better and faster methods of data management which starts with better methods of data storage. Data storage itself is a problem unless it is known where to store which data. Again cluster analysis comes to the rescue and allows data managers to store infinite data in finite or fewer clusters. While, on the one hand the clustering solves the problem of data storage, it also helps in extracting valuable information from the data through its analysis. The information in the data, however, can be deciphered through understanding sequences, associations, and patterns in the data, which needs better and faster learning methods for which fuzzy and neurofuzzy methods have emerged. Against this backdrop the current paper attempts to educate about this field and provides a framework for neurofuzzy cluster analysis.

Keywords: Clustering, Fuzzy, Neural Network, Neurofuzzy

Introduction

Human and natural activities and phenomena have been generating valuable data since their inception. Most of this data have not been used by humans due to lack of data analysis techniques. The valuable hidden information in the data needs to be understood not

only to solve many human problems but to avoid many disasters for betterment of human welfare. Through the analysis of the data better and informed decisions and courses of actions can be taken in almost all the fields of scientific and human activities. The information in the data, however, can be deciphered through understanding sequences, associations, and patterns in the data, which has been the endeavor of knowledge workers, scientists and scholars for last many decades. However, with the emergence of the knowledge economy and revolutionizing breakthrough technologies in computing technologies, data analysis has taken a central role in all the fields of research and development. Today data analysis forms an important pillar in the development of knowledge economies where every day billions of data are processed, manipulated and arranged in different patterns with the help of different computing technologies. However, it is not only the best methods of data analysis but the exponential rate at which data is being generated every day through formidable challenges to the scholars and the world of computing. While data management is a complex function, grouping data into different groups as soon as they are generated brings an order in the data. This will help in reducing the infinite number of data points into finite data groups or clusters, with intra-cluster homogeneity and inter-cluster heterogeneity on the differentiating variables. It will make it practically possible to manage data for storage and information extraction besides helping in minimizing the computation complexity required in further data processing. However, the effectiveness of such dispensation depends on the underlying grouping or clustering variable. Thus, cluster analysis forms the first and most important step in data processing. Against this backdrop the present paper attempts to educate academia and scientists about the nuances of data clustering and also propose a framework for it. The rest of the paper discusses data clustering, fuzzy systems, artificial neural works and Neurofuzzy framework.

Data Clustering

While data analysis today underlies many computing applications, it can be classified either as exploratory or confirmatory. However, the key element in both types of procedures is the grouping, or classification/clustering of data. Data clustering, having an immense number of applications in every field of life, is the organization of a collection of patterns into clusters based on similarity. It refers to the process of partitioning a set of data into a set of meaningful subclasses called clusters. Clustering, a useful and fundamental tool for analysis of data, is in essence the problem of finding a partition of a data set so that, under some definition of "similarity," similar items are clubbed together in one partition and different items are clubbed in different parts. Clustering finds useful applications in many

exploratory patterns-analysis, grouping, machine-learning, and decision-making situations, including data mining, image segmentation, document retrieval, and pattern classification. It has been addressed in many contexts and many disciplines by researchers and scholars, which is a reflection of its wide appeal and application as one of the steps in exploratory data analysis. Even though there is an increasing interest in the use of clustering methods in pattern recognition (Anderberg 1973), image processing (Jain and Flynn 1996) and information retrieval (Rasmussen 1992; Salton 1991), clustering has a rich history in other disciplines (Jain and Dubes 1988) such as biology, geography, geology, archaeology, psychology, psychiatry, marketing and finance. The vast and diversified literature on clustering is also an evidence of its importance and interdisciplinary nature.

Clustering, in general, is accomplished by analysis of input data for automatic characterization, detection and classification. Jain and Dubes (1988) argue the typical pattern clustering activity involves pattern representation (optionally including feature extraction and/or selection), the definition of a pattern proximity measure appropriate to the data domain, clustering or grouping, data abstraction (if needed), and assessment of output (if needed). The approach to clustering in vogue broadly fall under statistical, fuzzy and machine learning techniques. The statistical techniques analyses the data's linear characteristics and classifies it accordingly. The fuzzy set theory technique introduces uncertainty similar to human thinking in the classification process and thus making it robust. The machine learning technique such as artificial neural network (ANN) captures the nonlinear characteristics of data, resulting in better classification.

Fuzzy System

Cluster analysis in traditional terminology means crisp or hard partitioning. In these methods every given object is stringently classified into a specific group in such a way that each object goes to one and only one cluster. Thus the cluster boundaries, defined for the objects or data elements, are very sharp. Contrary to this, in reality the features or attributes of the objects are not sharp as they may be having some tendency to be the part of some other class to some extent. In order to model this reality, the fuzzy sets theory provides a powerful tool for soft partitioning of the data sets. Thus, clustering by using fuzzy concepts, called fuzzy cluster analysis, became a useful tool for data analysis. Fuzzy clustering can more objectively reflect the real world as it obtains the degree of uncertainty of samples belonging to each class and expresses the intermediate property of their memberships. Therefore, it has become the main subject of study on cluster analysis. Research and practice have shown that fuzzy clustering is advantageous to crisp clustering in that the total commitment of a vector to

a given class is not required in each of the iterations. Fuzzy methods have great ability to detect not only hyper volume clusters, but also clusters that are thin shells - curves and surfaces. When compared to crisp approach, fuzzy approach is more successful in avoiding local minima of the cost function and can model situations where clusters overlap.

Fuzzy clustering has been effective in solving the problem of assigning each data point to one and only one of the clusters by assigning a membership grade indicating each data point's degree of belongings to each cluster. Bezdek (1974) propounded fuzzy c-means (FCM) algorithm, which is one of the best known fuzzy clustering approaches. Its use for various applications is well described and analyzed. FCM method works on the optimization of a specific cost function, and it operates well when the clusters are compact or isotropic. Various variants of FCM have emerged as some researchers have worked to decrease the time of consumption others have worked to improve the accuracy. This method has been applied to various data types particularly in the area of image segmentation with slight variations.

One more approach to fuzzy clustering is to determine the partitions from binary fuzzy relations between pairs of samples in a dataset by using the transitive closure technique. While a cluster in this method can be defined as a fuzzy set whose elements are similar to each other, the fuzzy relations between pairs of these elements are not less than a certain level. The advantages of this method include that the number of clusters does not have to be specified in advance and every fuzzy equivalence relation induces a partition in each of its cut. However, the drawbacks includes that these fuzzy equivalence relations cannot be easily found, since the constraints are so restrictive that very few relation functions exist and the computation of transitive closure is complicated.

Artificial Neural Network

For last more than three decades artificial neural networks (ANNs) have been used extensively for clustering. ANN seeks to emulate the architecture and information representation patterns of the human brain. Its architecture depends on the goal to be attained. Patterns are presented at the input which are associated with the output nodes with differential weights. An iteratively process is followed to adjust the weights between the input nodes and the output nodes until a termination criterion is satisfied. This process of weight adjustment, called learning, lends continuous learning or artificial learning capability to the system, which can be either supervised or unsupervised learning in ANN. The supervised learning demands an output class declaration for each of the inputs. The unsupervised learning network itself recognizes the features of the input and self organizes

the inputs. It follows two approaches, the parametric approach and the non-parametric approach. Former approach involves combining classification and parameterization and the lateral approach involves partitioning the unclassified data in subsets using adaptive resonance theory (ART) which encompasses a wide variety of neural networks (NN) which is based on neurophysiology including prior knowledge and adaptive to learning. This phenomenon, known as stability-plasticity dilemma, forms the basis of competitive learning.

Competitive learning exists in biological neural networks. Competitive or winner–take–all, neural networks (Jain and Mao 1996) are used often to cluster input data. Similar patterns are grouped together by the network which is represented by a single unit called neuron. The grouping of patterns is done automatically on the basis of data correlations. The weight update procedures or learning, however, are rather similar to those in several classical clustering approaches. Familiar artificial neural networks used for clustering include Kohonen's learning vector quantization (LVQ) and self-organizing map (SOM) (Kohonen 1984), and adaptive resonance theory models (Carpenter and Grossberg 1990). The SOM gives an intuitively appealing two-dimensional map of the multidimensional data set, and it has been successfully used for vector quantization and speech recognition (Kohonen 1984). SOM, however, generates a suboptimal partition if the initial weights are not selected properly. Additionally, various parameters such as learning rate and a neighborhood of the winning node in which learning takes place controls its convergence. At different iterations it is possible that a particular input pattern can fire different output units, which brings up the stability issue of learning systems. A system is stable if after a finite number of learning iterations no pattern in the training data changes its category. This problem is closely related to plasticity, which is the ability of the algorithm to adapt to new data. As iterations progress the learning rate should be decreased to zero which gives stability, and also affects the plasticity. Carpenter and Grossberg (1990) argue that the ART models are supposed to be stable and plastic. However, ART nets are order-dependent; which means that different partitions are obtained for different orders, in which the data is presented to the net. Further, the number and the size of clusters created by ART net depend on the value selected for the vigilance threshold. Vigilance threshold is used to decide whether a pattern is to be assigned to one of the existing clusters or a new cluster has to be started. Further, both SOM and ART are suitable for detecting only hyper spherical clusters (Hertz et al. 1991).

Neurofuzzy System- Framework

While a fuzzy system can describe the knowledge it encodes but it can't learn or adapt its knowledge from training examples. On the other hand a neural network can learn from

training examples but cannot explain what it has learned i.e it is impossible to interpret the result in terms of natural language. Neural networks and fuzzy systems, therefore, have complementary strengths and weaknesses. The merger of neural networks and fuzzy logic in neurofuzzy models offers learning as well as readability. Because of this, many researchers have made attempts to integrate these two methods to create hybrid models that can combine the advantages of both. In the conventional approach to fuzzy clustering the model designer based on a priori knowledge fixes the membership functions and the consequent models. However, in cases where this set is unavailable and instead a set of input-output data is observed from the process, the components of fuzzy system i.e. membership and consequent models can be represented in a parametric form and the parameters are tuned with the help of neural networks. In such situations the fuzzy methods turn into neurofuzzy methods.

Neurofuzzy methods combine the uncertainty handling capability of fuzzy systems and the learning ability of neural networks. Thus, neurofuzzy (NF) computing has become a popular framework for solving complex problems in general and clustering problems in particular. In case the knowledge about clustering or any general problem can be expressed in linguistic rules, then a fuzzy inference system (FIS) can be built, and if it is in data, or can be learned from a simulation or training then artificial neural networks (ANNs) can be applied. In order to build a FIS, it is required to specify the fuzzy sets, fuzzy operators and the knowledge base. In the same vein, for constructing an ANN it is required to specify the architecture and learning algorithm. As the shortcomings of these approaches are complementary, therefore integration of these two systems is a natural corollary. In this new system the learning capability is an advantage from FIS viewpoint, and the formation of linguistic rule base is the advantage from ANN viewpoint. The requirements of such a system are that it should identify the structure of the data, identify the parameters and extract the rules. For this, different researchers have used different approaches to extract initial fuzzy rules from the given input-output data. However, for developing fuzzy systems clustering techniques have emerged as a powerful alternative approach. Clustering of numerical data now forms the basis of several classification and system-modeling algorithms. Clustering works to identify natural grouping of data from a large data set to produce a concise representation of a system's behavior which can be managed in an efficient manner. Thus neurofuzzy offers a great opportunity in the field of clustering.

The Neurofuzzy framework discussed in this paper has two major phases, structure identification and parameter identification (fig. 1). The structure identification relates to determining the sufficient number of rules required to properly model the available data and

the number of membership functions for input and output variables. In the parameter learning phase coefficients of each rule are tuned e.g. the shape and positions of membership functions. The learning scheme is generally comprised of two steps. The number of rules nodes that is the structure of the network and initial rule parameters or weights are determined in the first step. It is done by using structure identification and in the second step all parameters are adjusted using parameter identification. In this system fast computation speed is achieved by having parameters requiring much less tuning. There is however a need for effective methods to tune the membership functions so as to minimize the output error or maximize the performance index.

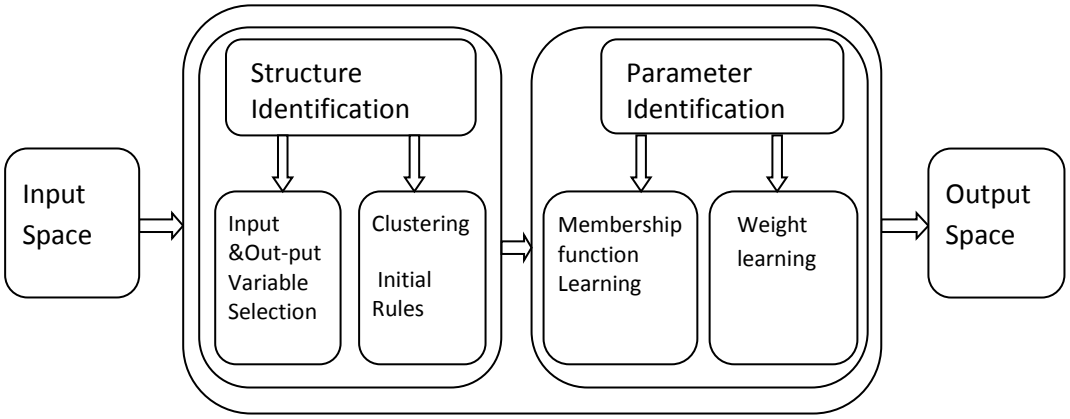

Fig. 1 Neurofuzzy Framework

The basic objective of structure identification is to obtain the integrated neurofuzzy network structure which includes: sufficient input and output variables; appropriate membership functions or antecedents and consequents description; and the proper size of hidden layers of network or the numbers and the expression of the rule. While the selection of input variables is responsible for explaining the whole operating region, the selection of output variables describes the behavior of system in certain operating region. The antecedents and consequents formalized description indicates the kind of structure of the system model. Rule generation generally done through fuzzy clustering. This is done through different fuzzy partitions that can be sorted in three different types: input-space clustering (E.Kim, et al. 1997), output-space clustering (Emami, et al, 2000) and input-output-space clustering (Hellendoom and Driankov 1997). In the input-space input data is classified into various sub-groups through clustering, which may be based on the data correlation. Then the output data are fitted by some particular functions like linear, Gaussian or Sigmoid functions in each of the sub-groups. In the output-space clustering, the output space is clustered first. Then the input space fuzzy partitions are derived by separately projecting the output space partition

onto each input space. In input-output-space clustering, also known as product-space clustering, the input data and the output data are combined based on the data causality, after which, cluster analysis is carried out in the whole space. Clustering analysis is carried out through fuzzy clustering, which generally uses an unsupervised type of learning, having a direct method based on the estimation of the probability function or indirect method based on the similarity metric of the data. However, the output of fuzzy clustering analysis is determined by the training data, similarity metrics, distance metrics, and clustering criteria. As such many experts and researchers focus on the appropriate selection clustering criterion to improve the clustering performance These include (Kim et al, 1998) , who takes into consideration the correlation among components of sample data, (Yeung and Wang, 2000), who introduces the similarity matrix composed with feature weights into the objective function and (Qiu et al, 2002), who constructs a target function using entropy etc.

As compared to the structure identification, the parameter identification has lesser ambiguity and can give a better estimated solution. Generally, parameter identification is achieved in two ways: the family of gradient algorithm and the family of least squares estimation. So far many gradient identification algorithms have been proposed, like E.Kim, et al. (1997), uses the gradient descent algorithm to precisely adjust parameters of the fuzzy model instead of nonlinear optimization methods, and (Wong and Chen, 1999) adopts gradient descent approach and clustering method to automatically construct a multi-input fuzzy model and so on. However, because of its successful applicability to many practical problems many of the researchers have successfully used least square estimation as a basic identification technique and therefore, this methodology of generally recommended for the present framework as well.

Conclusion

This paper has introduced a neurofuzzy system model for data clustering, which includes two steps: the structure identification and parameter identification. It is applied to generate fuzzy rules automatically, and then fix on the size of the neurofuzzy network. This system which combines the neural networks and the fuzzy set theory has emerged as a great breakthrough in the field of clustering, which is a process of grouping data items based on a measure of similarity. Clustering per se is a subjective process, which means that the same set of data items repeatedly need to be partitioned differently for various applications. It makes clustering difficult as a single algorithm or approach will inadequate to solve all the clustering problems. This problem is taken care of by the neurofuzzy system as it is a self

learning system and generates patterns and rules automatically. Thus this paper argues the use of this system in all the learning and artificial intelligence systems

References:

Anderberg, M. R. Cluster Analysis for Applications. Academic Press, Inc., New York, NY. 1973

C. Wong, C. Chen. A hybrid clustering and gradient descent approach for fuzzy modeling [J]. IEEE Trans.on Systems,Man and Cybernetics, Part B, 29(6) :686 - 693, 1999.

Carpenter, G. and Grossberg, S.. ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. Neural Networks 3, 129–152, 1990.

D.S. Yeung, X. Z. Wang. Using a neuro-fuzzy technique to improve the clustering based on similarity [C]// 1EEE Int. Conf. on Systems, Man, and Cybernetics, 5 : 3693 – 3698, 2000.

E. Kim, M.Park, S.Kirn, et al.'A transformed inpnt-domain approach to fuzzy modeling [Jl. IEEE Trans. on Fuzzy Systems,6(4): 596 - 604, 1998.

E.Kim, M.Park, J.Seunghwan, et al.A new approach to fuzzy modeling [J]. 1EEE Trans. on Fuzzy Systems,5(3):328- 337, 1997.

H. Hellendoom, D. Driankov: Fuzz)" Model Identification Selected Approaches [M]. Berlin: Springer, 1997.

Hertz, J., Krogh, A., and Palmer, R. G. Introduction to the Theory of Neural Computation. Santa Fe Institute Studies in the Sciences of Complexity lecture notes. Addison- Wesley Longman Publ. Co., Inc., Reading, MA. 1991.

J. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Ithaca, NY: Cornell University, 1974.

Jain, A. K. and Dubes, R. C. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. 1988.

Jain, A. K. AND Flynn, P. J. Image segmentation using clustering. In Advances in Image Understanding: A Festschrift for Azriel Rosenfeld, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ, 65–83. 1996.

Jain, A. K. and Mao, J. Artificial neural networks: A tutorial. IEEE Computer 29 (Mar.), 31–44. 1996.

Kohonen, T. Self-Organization and Associative Memory. 3rd ed. Springer information sciences series. Springer-Verlag, New York, NY. 1984.

M. Emami, A. A. Goldenberg, I. Turksen. Systematic design and analysis of fuzzy-logic control and application to. robotics, Part I [J]. Modeling, Robotics and Autonomous S) stems, 33 (2,3) : 65 – 88, 2000.

Rasmussen, E. Clustering algorithms. In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419–442. 1992.

Salton, G. Developments in automatic text retrieval. Science 253, 974–980. 1991.

X.Qiu,Y.Tang, D.Meng, et al.A new fuzzy clustering method based on distance and density [C]// IEEE Int. Conf. on Systems, Man and Cybernetics,7 : 5 – 9, 2002.