Application of Machine Learning in Analysis of Folklorian Data: An Overview

Shasanka Kumar Gayen¹ & Shrinwantu Raha²**

¹²Coochbehar Panchanan Barma University, Coochbehar, West Bengal, India, Pin: 736101
**Corresponding Author

Received: May 29, 2018
Accepted: July 09, 2018

1. Necessity of the Data and Machine Learning Procedures
Though there are a lot of work done on folklore and folkloristic but in most of the cases those is qualitative in nature. Thus whenever we compare between the regional variations of cultural belonging it seems to be ambiguous because sometimes it is hard to distinguish qualitatively. Here we feel the necessity of quantitative data and machine learning to integrate the relation between parameters of folklore study. The inherent culture, belief system, traditional belonging of a particular group of people is the effect of environmental determinism. The physical earth makes the harmony with the traditional way of life of human beings. Relationship between the environmental, physical and cultural parameters can be effectively portrayed through the machine learning algorithms. The climatic phenomena in relation to earth surface features can be mapped with GIS and machine learning helps to validate the whole features of it.

2. Data Structure and Scope of Machine Learning in Folklore:
The word ‘Learning’ completely determines from the outcomes of knowledge, modification of behaviour. Towards the fountain of knowledge data is also required needs to understand. Data may be qualitative, may be quantitative (Iannelli, Martcheva,Milner, 2005). Through Qualitative data, it is hard to recognize the difference between the two successive regions (Kuhn and Johnson, 2013). Quantitative data helps to quantify numerically the ratings of every parameters helps to compare the two specific types of data structure. Machine learning is the instruction to the computer without being explicitly programmed (Samuel 1959). The human behaviour, human belief, cultural perspectives can also be predicted using the machine learning algorithms. The subjective rating of cultural phenomena can be transferred into probability model helps to integrate the application of machine learning algorithms very easily. Being a part of artificial intelligence machine learning also opens the domain of prediction of cultural phenomena very efficiently (Friedman and Kandell, 2000). The neural network model as a part of machine learning artificially helps to determine the input and output values of cultural phenomena within a specific time frame and ultimately help to manifest the inherent changes of those values. Machine Learning also helps in the decision making and planning procedures. Several approaches of machine learning procedures are as the association rule learning, predictive modelling, artificial neural network, deep learning, and support vector machines, cluster analysis, Bayesian network, reinforcement learning, representation learning, similarity learning, sparse decision making learning, genetic algorithm, rule-based machine learning and learning classifier systems.

2.1 Association Rule Learning
Association rule learning is the discoveries of interesting relations between the larger databases within different attributes. It is a interesting process to identify wider relationships between variables (Maitrey and Jha, 2014). The attributes are interrelated within a certain knowledge domain. In MATLAB the useful concept easily portrayed through the FUZZY LOGIC. Fuzzy Logic based on the membership function operated through the Boolean algebra (Kuhn and Johnson, 2013). We can cite an example of Purulia District as a part of this subjective association of rule learning. Figure 1 denotes the association rule feature of every parameter connected in logic within the natural harmony. So, Nature is the universal set can be marked as U. Within this physical environment is a set marked as A; and Socio-cultural environment marked as B so A is defined as:

\[
A = \{ R, S, V, G, C \}
\]

\[
B = \{ P, RE, B, CL, E \}
\]

Where, \( R = \) Relief, \( S = \) Soil, \( V = \) Vegetation, \( G = \) Geology, \( C = \) Climate, \( P = \) Population, \( RE = \) Race and Ethnicity, \( B = \) Belief System, \( CL = \) Customs and Love, \( E = \) Economy

Here we can say that \( B \subseteq A \) or B is a subset of A. Because Population, their culture, love, belief system entirely depends on the physical environment. Figure 2 denotes the structure of the set. Table 1 denotes the subjective scores of every parameters and interrelationships. This interrelationship and parameter wise classification are done to have the better prospective of the model and machine learning.

1800x  IJRAR- International Journal of Research and Analytical Reviews  Research Paper
2.2 Predictive Modelling

Geisser (1965) defined ‘predictive modelling’ as the process by which a model is created or chosen to try to predict the best, the probability of an outcome’. Predictive Modelling by Kuhn and Johnson (2013) assert that it is the process of developing a mathematical tool or model that generates an accurate prediction. For each and every parameter (Figure 1) we can generate primary data or can acquire data from the secondary sources. For the physical parameters we can generate data secondarily through the satellite image and with the help of GIS. After that we can run several programming software to estimate the changing structure of the association matrices between the two different time frames. Different analytical neural network models are in the mood of estimating the associative correlation between the parameters and covariance matrices altogether. Correlation helps to understand the degree of association between the parameters and covariance matrices help to estimate the parameterise standard deviation values. The correlation coefficient ($r$) value helps to estimate the degree of togetherness between the variables and attributes. $-1 \leq r \leq 1$ is the well known specified degree of coefficient between the attributes. The coefficient value $> 1.5$ denotes the correlation as strong association and $< 1.5$ denotes the correlation as weak association.
2.3 Cluster Analysis

Grouping of similar objects into some strict facts called cluster or clusterization. Suppose we are dealing with the 3 tribal groups of a particular region to identify the cultural strengths and tradition of that region. To do so, we will go for the in-depth interviews by selecting sample from the each group. We may get the common properties of every tribes of that particular environment. Same environment will force them with literary same adaptive capacity. Primarily local relief, drainage, geological factors will be same and as subjective grouping if we want to plot it they will be clubbed along the specified axis based on the value choosed in our methods. Again the race, ethnic strategy the rituals, the cultural belonging might be different between the each tribe altogether. So, according the choosen subjective scores those values will be scattered into the diagramme. Such analysis is very much important in the determination of the nature of variables which are helping us to draw the nature of folk tradition of that particular area. Again if we want to progress the data into the Map with Geographical Information Systems this analysis will surely be a great help. The subjective scores can be put as attributes into the map and we can easily differentiate between the many successive portions of the region. Sometimes we can rank the parameters based on their importance and this idea of importance can be fixed based on the in-depth interviews of the target people. Based on the nature of datasets we can choose clusteral procedures. There may be 3 basic cluster analysis approaches:

2.3.1 Connectivity Based Clustering (hierarchical clustering)
It is often said that effective clustering needs the eye of the beholder. The basic idea of this algorithm is that it attracts the nearest neighbourhood value than the farthest value. Nearest neighbourhood values are clubbed into dendrogram form a particular structure or outcome for decision making (Bishop, 2006) Usual choice of distance functions are required to compute such algorithm. Clustering usually represented as dendrograms and folk data with subjective scores can be represented in it and scores will merge after a certain distances. With the minimum object distances we can develop linkage clustering. The Complete Linkage Clustering or unweighted pair group of clustering can be done using the folklorian data. We can give subjective scores and then individual weightage can be given to each parameter. This individual weightage can be relative or can be subjective in nature. The maximum value takes action in computing the Complete Linkage Clustering. If we directly act with subjective scores of the parameter and accordingly we will able to compute unweighted pair group of clustering.

2.3.2 Centroid based Clustering
Centroid based clustering method is based on the centralized vector which is essentially a member of the dataset represented in a particular matrix based data sets. Neighbourhood clustering procedures is inherently matched with this clustering procedure. It partitions datasets into unique structures known as Voronoi diagramme. Centroid based clustering will be helpful in considering multi regional aspects. In such cases Voronoi diagramme will make partition and grouping along the similar kind of variables (Everitt et.al 2011, Friedman and Kendall 1999). To identify this we must do subjective scores or place relative weightage along the selective parameters of folk. From that diagramme we easily able to identify the similar feature exists between the regions because based on the subjective scores similar algorithms will be clubbed and distinguished. Regarding to estimate the similarities and dissimilarities between the regions this method is more useful than the previous one because it partitions between the specific algorithms which is clearer to draw attention.

2.3.3 Density based Clustering
Density based Clustering is done by making density zones based on algorithm. Through contouring we easily demarcate the same values and draw the hypothetical line containing the zonations (Kuhn and Johnson, 2013). So, this method is not abruptly doing Zonation like the previous method but rather this Zonation is smooth, easily understandable sometimes highlighted through the colour bar or code. We need the same algorithm practice with subjective scores in this method.

2.4 Bayesian network Model
The human belief or traditional belief system can be mapped using the Bayesian network model with a set of random variables via a directed acyclic graph (Farahmand B., 2009). Bayesian network in folklore perspective can represent the probabilistic relationship between the population structure and traditional belief and culture of a particular human community. We can select parameters to estimate within a specific population structure and can generate the association rule matrices between the selected parameters to estimate Bayesian network model.
2.5 Reinforcement Learning

Environment induces culture, dress, food habit; settlement-pattern thus helps to build a strong cultural tradition. How a particular group of people or tribal groups are adjusting in particular environment reinforcement learning will be helpful to understand this phenomena. The environmental determinism with the cultural and folk parameters interlinked and those interlinking can be effectively investigated through feed-forward reinforcement procedures.

2.6 Genetic Algorithms

How a population structure is evaluated within particular systems of environment can be estimated using genetic algorithm. This is sometimes a heuristic procedure to generate the natural selection through mutation and crossover to procedure several genotypes. From folklorian point of view the evolution of a particular race within a specific natural environment can be enhanced using this type of algorithm.

References: