Techniques for Machine Learning based Spatial Data Analysis: Research Directions

M. Gangappa
Associate Professor
Dept. of CSE
VNR VJIET, Hyderabad-India

C. Kiran Mai, PhD
Professor
Dept. of CSE
VNR VJIET, Hyderabad-India

P. Sammulal, PhD
Assistant Professor
Dept of CSE
JNTUH College of Engineering, Karimnagar

ABSTRACT

Today, machine learning techniques play a significant role in data analysis, predictive modeling and visualization. The main aim of the machine learning algorithms is that they first learn from the empirical data and can be utilized in cases for which the modeled phenomenon is hidden or not yet described. Several significant techniques like - Artificial Neural Network, Support Vector Regression, k-Nearest Neighbour, Bayesian Classifiers and Decision Trees are developed in past years to acheive this task. The first and most complicated problem concerns is the amount of available data.

This paper reviews state-of-the-art in the domain of spatial data analysis by employing machine learning approaches. First various methods have been summarized, which exist in the literature. Further, the current research scenarios which are going on in this area are described. Based on the research done in past years, identification of the problem in the existing system is also presented in this paper and have given future research directions.

Keywords

Spatial data Analysis, ANN, SVM, Feature Extraction, Classification, Rough sets

1. INTRODUCTION

The process of spatial data extraction along with knowledge exploration is an efficient derivation of implicit, relevant, unknown, potentially useful, spatial or non-spatial knowledge, which may involve different rules, precision, patterns and constraints from incomplete, noisy, random and unstructured data.

Predictive modeling encompasses a variety of statistical techniques from machine learning that analyze the present and historical factual data to make the predictions about the future events. Machine learning, being evolutionary domain and which comes under the field of artificial intelligence, equips the computing machines, the capacity to being train. Machine learning concentrates on the improvement of PC programs that can change when presented with any kind of new information. Here, computer machines scan through information to search for patterns. Machine learning utilizes that information to recognize designs in information and change program activities appropriately. Machine learning mechanisms are categorized into three categories i.e. supervised, unsupervised and reinforcement. Supervised algorithms can apply what has been realized in the past to new information. Unsupervised algorithms can draw assumptions from datasets. In reinforcement, A PC program communicates with a dynamic environment in which it must perform a definite goal.

Some of the significant and popular machine learning algorithms are as -.

(1) Regression algorithm
(2) Instance-based algorithms
(3) Regularization algorithms
(4) Decision Tree algorithms
(5) Bayesian algorithms
(6) Clustering algorithms
(7) Association Rule Learning algorithms
(8) Artificial Neural Network based algorithms
(9) Deep Learning algorithms
(10) Dimensionality Reduction algorithms
(11) Ensemble algorithms

A satellite image using ANN [1] used it's capability to characterize the satellite pictures by utilizing distinctive algorithm which are back-propogation algorithm and K-mean mechanism with different methodologies. ANN's classifier is correlated with Maximum Likelihood (ML) along with unsupervised (ISODATA) conventional classifier. ANN's classification is proper classification and based on training data set. ML and ISODATA is based on remote sensing applications. In order to correlate and compare the execution performance of image classification, classification accuracy is calculated. Various data applications and software tools [2] have been developed for geospatial information: local characterization of environment information, mapping of persistent environment and pollution information, including the utilization of automatic algorithms, improvement of monitoring networks.

A machine learning mechanism for automated exploration and building of expert system for image components analysis expert systems incorporating GIS data [3] employs an inductive learning algorithmic procedure for the generation of production rules from training data. This method is easier to build a knowledge base scenario for a rule-based expert system compare to the conventional knowledge acquisition approach. Image catagorization using Support Vector Regression and Artificial Neural Network [4] composed two areas of machine learning i.e. ANN and SVM. In this paper, image is being divided into muliple sub-images based on its features. An ANN classified each sub-images into responsive class.

Finally, all ANN classified results are also being complied by SVM classifier.

1.1 Motivation to the Problem

Spatial data classification, being an interesting as well as challenging task, has fascinated many researchers in earlier years. In existing literature, several classification mechanisms for classification of images eg. Artificial Neural Networks, Genetic Algorithms, Support Vector Machines, decision trees etc. been proposed. In earlier years, ANN procedure had been proved as a notable classification methodology. The execution efficiency of any specific neural network is characterized by its topological structural nature along with the learning method used. However, there is no exact rule from which one can determine the structural network topological behaviour of ANN which is used in the process of classification. Thus, ANN classifier can be proved to be a vulnerable classifier if the structure adopted here, is having an improper as well as inappropriate amount of neurons in hidden layers. The process of spatial data analysis utilizing SVM is also much complex procedure for understanding the structure, also it requires pre-information about knowledge data sets e.g. stochastic and probabilistic information.
Rough set theory, which evolved in early 1980s, is a tool which has its importance to knowledge retrieval and classification through machine learning, decision support systems, inductive reasoning etc. Prime benefit of RST is that it doesn't need any kind of preliminary information about the sample data set e.g. probability distribution, modular probability assignment etc.

1.2 Organization order of the paper

In rest of the paper, Section 2 gives the overview of the required preliminaries. Section 3 discusses about related work. The Various existing techniques for spatial data analysis are discussed in section 4. The future research directions are discussed in section 5. And finally section 6 concludes the paper.

2. PRELIMINARIES

Some of the preliminaries utilized in spatial data analysis process are summarized as below:-

2.1 Artificial Neural Network

Artificial Neural network is a data refining and processing prototype, stimulated via the manner organic apprehensive structures, inclusive of the mind, process information. The important thing of this paradigm is the unconventional structure of the information processing systems. They can give reasonable answers for issues, which are for the most part described by non-linear ties, high dimensional, noisy, complex, loose, and imperfect or mistake inclined sensor information. ANN consists of 3 layers as- Input layer, middle processing layer and output layer. In the mid layer, network topology is selected according to the input given. The structure is shown as Fig.1 below -

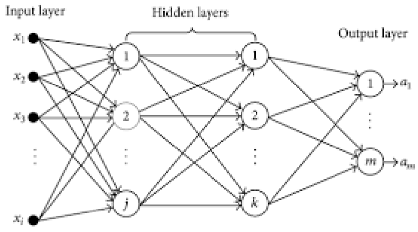

Fig. 1. ANN Pictorial representation

2.2 Support Vector Machine

SVM employs linear hyperplanes along with non-linear hyperplanes classification task for the given attributes characters in the space. SVM is capable of providing good performance and approximate accuracy in classification. SVM can be employed for classification as well as regression challenges but it is mostly used for catagorization. In this algorithmic procedure, for every data object plotting is done as a point in n-dimensional space, Where n represents: no. of features appearence present. Each feature value being the value of an specific coordinate. Afterwards, classification will be employed by discovering the correct hyper-plane that seperate the two classes well. The structure is shown as Fig.2 below.

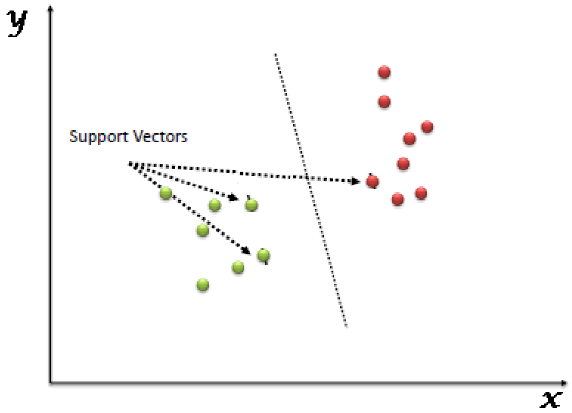

Fig. 2. SVM Pictorial representation

2.3 Rough Set Theory

This section contains summarization of basic details of rough set theory which was originally proposed by Z. Pawlak. [14] In some way, initial theory of rough sets is reffered to as ”Pawlak or classical Rough Sets”.
Suppose I is an information system(IS) which is equal to set (U, A) where A: finite set of attributes and U: non-empty finite set of objects such that: $a : U \rightarrow V_a$, for every $a \in A$ where V_a represent set of value that attribute a may take. Each attribute a and object x in U will get a value $a(x)$ from V_a using information table. An associated equivalence relation $IND(P)$ with any $P \subseteq A$ is following:

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\}$$

where, IND(P): P-indiscernibility relation

3. RELATED WORK

This section presented the work which has been done over past years. In 1990, W. G. Aref et. al [5] proposed a mechanism to information management in geographical applications. This paper describes about geographical information management. In August, 1991 G. Aref et. al. [6] presented a way to extend a DBMS with contiguous operations. In this paper, an information structure is presented that fits the necessities for an efficient processing of structural queries inside the extended database context. In september 1991, Walid G. Aref and Hanan Samet [7] proposed the optimization strategies for spatial querry processing. This paper describes about the major application of standard uncertain query processing and optimization mechanisms in the context of an integrated spatial database environment.

In 1996, Usma Fayad et. al. [8] suggested a primary step toward a unifying framework for knowledge Discovery in Database. This paper describes relationship between data minning, knowledge discovery and other related techniques. Also define the basic data minning algorithm, KDD process and application issues. In 1997, Huang et. al. [3] proposed a machine-learning algorithmic procedure for geo-spatial image reasoning with GIS data. This paper includes an inductive method to generate the production rules from the training dataset. This method is easier to build a knowledge base for a rule-based system as compare to the conventional knowledge acquisition procedure. In 2002, Hung-Chih Wu et. al. [9] proposed a data mining methodology for spatial modeling in Small Area Load Projection. This paper contains investigational data analysis, seeking to explore beneficial patterns within spatial data which aren't apparent to the information user and are beneficial inside the spatial load forecast. In 2008, M. Kanevski et. al. [2] proposed a machine learning algorithms for geo-spatial data applications and software tools. This paper presents a survey of a few contemporary utilizations of ML for geospatial information: local characterization of environment information, mapping of persistent environment and pollution information, including the utilization of automatic algorithms, improvement of monitoring networks. In March 2008, Juan Carlos Niebles et. al. [10] proposed an unsupervised learning mechanism of human action categories utilyzing spatial-temporal phrases. Space-time interest point extraction will be used to represent the video sequence as a collection of spatial-temporal phrases. The probability distribution of the intermediate topics corresponding to human action categories and spatial-temporal words will be learnt automatically by the algorithm. This paper gave method to handle noisy feature points. In 2009, Jeremy Mennis et. al. [11]performed development of the methodology and practice to fetch and extract the beneficial records and information from huge and complex spatial databases. In 2012, Le Hoang Thai et. al. [4] proposed an image classification mechanism using Support Vector Machine and Artificial Neural Network. In this paper, image is being divided into muliple sub-images based on it's feature. An ANN classified each sub-images into responsive class. Finally, all ANN classified results will be compiled by SVM.

In 2014, Nur Anis Mahmon et. al. [1] presented a review on catagorization of satellite image using Artificial Neural Network (ANN). ANNs classification is proper classification and based on training data set. In 2016, Abdullah Saeed Ghareb et. al. [15] presented a methodology for hybrid attributes selection, which is based on the enhanced GA for text categorization. This technique utilizes a hybrid search technique that combines the advantage of filter based feature selection techniques with an enhanced GA (EGA) in a wrapper approach. In 2001, Nicolas Gilardi [12] proposed a local machine learning representation for spatial data analysis.

4. VARIOUS EXISTING TECHNIQUES FOR SPATIAL DATA ANALYSIS

4.1 Spatial image classification using ANN

A satellite image using ANN [1] used the capability of ANN's characterizing the satellite pictures utilizing distinctive algorithm which is back-propagation algorithm and K-mean algorithm with different methodologies. ANN's classifier is compared with maximum likelihood(ML) and unsupervised(ISODATA) conventional classifier. ANN's classification is based on training data set. ML and ISODATA are based on remote sensing applications. To find out the comparison result of performance of the image classification, classification accuracy and Kappa Coefficient were calculated. Image processing is a technique of collecting the information to manipulate the digital images. Normally there are four steps to classify an image.

—Preprocessing - to reduce haze and to detect the band ratio etc.

—Training sample - to select a special region for describing the pattern.

—Decision - to pick out the best technique to compare the pattern according the object.

—Evaluate the correctness of the classification.

Image categorization is one of the important tasks in view of environmental application. This study accentuate the use of different neural network algorithm like back propagation etc.

Supervised learning feed-forward neural network which is back-propagation algorithm is used for image classification in this paper. The training sample of normalized process has been performed before training and classify LU/LC of image satellite. This process is used to avoid the saturation in the network broadcasting process. The normalized formula is shown below:

$$x^{'} = (x - x_{min})/(x_{max} - x_{min})$$

where:
$x^{'}$: mean value of normalized input
x:input vector
x_{min}:original entire training samples set for minimum values
x_{max}:original entire training samples set for maximum values
The most important stage to get the best achievement of classification task, is selection of the training set.

4.2 Machine Learning and AI based Approaches for Geo-Spatial Data Applications and Software Tools

Machine Learning methodological approaches for geo spatial data applications and the software tools [2] presents a survey of a few contemporary utilizations of ML for geospatial information: local characterization of environment information, mapping of persistent environment and pollution information, including the utilisation of automatic algorithms, improvement of monitoring networks.

One of the potential solutions that can be found in machine learning algorithms, specially, distinct architectured in artificial neural networks and theory of statistical learning. i.e. kernel-based method, SVR and support vector machine. It is obvious that these type of approach are data driven like black/grey boxes majorly count on the quantity and quality of data. Hence, it is practicible and become important to use distinct statistical or geostatistical tools for controling the data analysis quality and use ML for modeling. There exist so many resources for machine learning procedures that includes theory structured tutorials and sophisticated software tools. Generally there are three basic tasks of statistical reasoning from data. i.e. probablity density modelling, classification and regression. Two other basic difficulties are: integration/assimilation of data and networks designing/redesigning and models based on science for example meteorological models etc. Few geospatial problems in data analysis and related methods that can solve them, are specified as below:-

—Spatial predictions: geostatistics, deterministic interpolators, machine learning.

—Modelling knowledge predictions having uncertainities: geostatistics, machine learning.

—Multivariate joint predictions of several variables: machine learning (multi-task learning).

—Modeling of probability density function locally according to risk mapping: machine learning (Mixture Density Networks), geostatistics (indicator kriging, simulations).

—Modelling of conditional simulations (spatial Monte Carlo simulations), spatial variability and uncertainty: geostatistical conditional stochastic simulations.

4.3 Image Catagorization/Classification employing Support Vector Regression Machine and Artificial Neural Network

Image Classification using SVM and Artificial Neural Network [4] composed two area of machine learning i.e. ANN and SVM. In this paper, image is being divided into multiple sub-images based on it's feature. An ANN classified each sub-image into responsive class. In this paper, image is presented in large representation space after image processing and its feature extraction. Then, it is divided into sub-space in order to reduce dimensions of image's feature and to analyze it easily. The feature vector would be extracted from ecah sub-space image. The feature vector will be input for ANN. ANN will have three layer i.e. input, hidden and output layer. The feature vector dimension is equal to number of nodes in input layer. The number of output node is equal to number of class.
Suppose, They have k sub-spaces. So, k classification results for sub-spaces will be there i.e. $CL_{SS_1}, \cdots, CL_{SS_k}$. Integration of all results is the main problem. The simple method to integrate is- to find mean value or weighted mean value:

$$\text{Mean Value: } CL = 1/k \sum_{i=1}^{k} (CL_{SS_i})$$

$$\text{Weighted Mean Value: } CL = 1/k \sum_{i=1}^{k} w_i(CL_{SS_i})$$

Where w_i: resultant weight of catagorization result of structured subspace SS_i , and satisfies: $\sum_{l=1}^{R} w_l = 1$
ANN_SVM, both the algorithmic procedures are easy to employ for the specific classification problem and design.

5. RESEARCH DIRECTIONS

The problem identification and further future research directions are presented as follows:-

5.1 Problem Identification

The process of spatial data extraction along with knowledge discovery meant for an efficient extraction of implicit, relevant knowledge from entirely incomplete, noisy and unstructured data in huge spatial databases. Predictive modeling hold within a diverse nature of statistical procedures from machine learning that analyze the present and pre-knowledge to make the predictions about the future events. The first and most complicated problem concerns the amount of available data.

5.2 Future Research Directions

The future research directions will be as follows:-

—Today, the size of unstructured spatial data is huge. So, to process such inconsistent, incomplete and vague data by computers is a challenging task. In todays real world problems, feature selection is an essential aspect due to presence of irrelevant features in the data. In recent past years, Rough set theory evolved as an efficient machine learning technique, which has become an important tool to perform FS. Reduct generation or attribute selection intends to discover a minimal possible attributes subset which can convey the same knowledge as it was exemplified by the original features.
We will propose new attribute/feature selection algorithms using RST as our research work. One extra advantage of RST is that it requires no prior information like - statistical and probabilistic information of data. Further, only the selected relevant features of the data will take part in the classification process. This process makes the predictive modeling task more accurate and practically efficient.

—Some machine learning algorithms used today, perform attributes selection but does not give the guarantee of dimensionality reduction of huge spatial data.
To perform operations in the data, present in higher dimensions may be more computationally complex procedure as well as the computational overhead is huge in further training and testing phases of classification. Modern machine learning technique eg. RST can be utilized to perform above tasks which also gives the guarantee of dimensionality reduction for the given input data.

—Some traditional machine learning techniques perform the feature extraction and classification tasks more accurately only for the complete information system.
So, by using Rough Set based machine learning approach, we are able to deal with the incomplete and inconsistent information systems, more accurately and efficiently.

6. CONCLUSION

Analysis of an spatially structured data is a challenging domain of research, having applications in diverse areas e.g. - medical imaging, GIS, robot sensory motion predictive planning and design etc. A specific machine learning mechanism for automated exploration and building of expert system for image components analysis expert systems incorporating GIS data employs an inductive learning algorithmic procedure for the generation of production rules from training data. Predictive modeling encompasses a hetrogeneous kind of statistical techniques from machine learning that analyze the present and pre-knowledge facts to make the predictions about the future occuring events. The first and most complicated problem concerns the amount of available data. This paper presents the problem and challenges of spatial data analysis and state-of-the-art in this area. The problem identification in this domain as well as future research directions are also discussed at the end of this paper.

7. REFERENCES

[1] Mahmon, Nur Anis, and Norsuzila Yaacob. A review on classification of satellite image using Artificial Neural Network (ANN), Control and System Graduate Research Colloquium (ICSGRC), IEEE, 2014.

[2] M. Kanevski, A. Pozdnoukhov, V. Timonin. Machine Learning Algorithms for Geospatial Data Applications and Software Tools. EMSs 2008: International Congress on Environmental Modeling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making 4^{th} Biennial Meeting of iEMSs.

[3] Xueqiao Huang and John R. Jensen. A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data. Photogrammetric Engineering and Remote Sensing, Vol. 63, No. 10, October 1997, pp. 1185-1194.

[4] Hai, Tran Son, and Nguyen Thanh Thuy. Image Classification using Support Vector Machine and Artificial Neural Network. International Journal of Information Technology and Computer Science (IJITCS) 4.5 (2012).

[5] W. G. Aref and H. Samet. An approach to information management in geographical appli- cations. In Proceedings of the 4th International Symposium on Spatial Data Handling, pages 589 - 598, Zurich, Switzerland, July 1990.

[6] W. G. Aref and H. Samet. Extending a DBMS with Spatial Operations. Advances in Spatial Databases - 2nd Symposium, SSD'91, vol. 525 of Springer-Verlag Lecture Notes in Computer Science, pages 299 - 318, Zurich, Switzerland, August 1991.

[7] W. G. Aref and H. Samet. Optimization strategies for spatial query processing. In Proceedings of the 17th International Conference on Very Large Databases (VLDB), pages 81 - 90, Barcelona, Spain, September 1991.

[8] Usma Fayad, Gregory Piatetsky-Shapiro and Padhraic Smyth. knowledge Discovery in Database. From: KDD-96 Proceedings. Copyright 1996, AAAI (www.aaai.org).

[9] Hung-Chih Wu and Chan-Nan Lu, Senior Member. A Data Mining Approach for Spatial Modeling in Small Area Load Forecast, 0885-8950/02 17.00 (2002) IEEE.

[10] Juan Carlos Niebles, Hongcheng Wang, Li Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. IJCV (2008) 79: 299 - 318 DOI 10.1007 /s11263-007-0122-4.

[11] Jeremy Mennis, Diansheng Guo. Spatial data mining and geographic knowledge discovery - An introduction, Computers, Environment and Urban Systems, Volume 34, Issue 2, March 2010.

[12] Nicolas Gilardi. Local Machine Learning Models for Spatial Data Analysis. Journal of Geographic Information and Decision Analysis. (2001), vol. 4, no. 1, pp. 11 - 28.

[13] Petrosino A., Salvi G.(2006). Rough fuzzy set based scale space transforms and their use in image analysis. International Journal of Approximate Reasoning, 41, 212 - 228.

[14] Pawlak Z.(1982). Roughsets. International Journal of Computer and information Sciences, 11, 341-356.

[15] Abdullah Saeed Ghareb, Azuraliza Abu Bakar, Abdul Razak Hamdan. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems With Applications 49 (2016) 31 - 47.