Advances in Wireless and Mobile Communications.
ISSN 0973-6972 Volume 10, Number 2 (2017), pp. 325-332
© Research India Publications
http://www.ripublication.com

A GIS BASED APPROACH (GISBA) OF MACHINE LEARNING FOR "INTERACTIVE KNOWLEDGE DISCOVERY PROCESS" [GISBA]

Professor Dr. Sudan Jha

School of Computer Engineering
KIIT University, Bhubaneswar, Odiha – 751024, India

Abstract

It is a well-known fact that the Geographical Information System (GIS) has advanced exponentially and subsequently has yielded out a chunk of huge amount of data especially spatial data. This paper included spatial data analysis with a formal technique. The present data mining techniques existing are not sufficiently capable enough to find out the knowledge from remotely sensed data that too from Spatial Databases. This paper explains the Spatial Analysis using topological, geometric, or geographic properties of remotely sensed data. This paper offers an integrated software systems / tool for data analytics, researchers, academicians, corporate professionals; moreover, some systematic experiments by spatial domain experts. The point to be noted - there is no specific training in machine learning or statistics required. The GISBA system allows data to be analyzed which are located physically in various widely but randomly located in various geographic locations by knowledge discovery instead of fetching and searching them no matter what the network connections may be.

Keywords: GIS, Spatial Data, Knowledge;

1. INTRODUCTION

On or after the adaptation of SAP, ERP packages whether in corporate culture or academic corona or any other domain, the requirement of "valued" data has exponentially grown up. Since "No Data is Dumb", again, Geographic Information System enabled data has been widely utilized. Earlier during 1992s, despite of the availability of various third party utilities, proper integration or migration or handling of data were not implemented in the huge repositories of data. Later, many algorithms

where been worked out and many tools consequently, but again the complexity with relation to time and space remain the same. The throughput of the system was still under the question mark? Moreover, then and now, no specific data mining technique existed, which could be robust and handle the remotely sensed data.

However, one of the most common tools in which few algorithms convertible into a package was introduced in which the most suitable algorithm was selected manually to the given target problem. As an example, one of the tool of "Machine Learning System" started coming into the picture which was capable enough to accelerate the development of algorithms. It not only increased the software reliability but was / is also capable enough to compare the performance of different algorithms. The approach used is based on object-oriented methodology, Object Oriented Analysis, Object Oriented Design and Object Management Technology.

With respect to the spatial databases and with respect to neural networks, the back propagation algorithm, extended with a "pruning" method mainly for classification tasks and a Kohonen [6] network for clustering tasks, spatial databases have advanced into the collection of huge chunk of repositories in various GIS applications ranging from satellite telemetry systems, remote sensing, GPRS to computer cartography and environmental planning. Here it is proposed to define the spatial data mining as a sub module of data mining that deals with the extraction of implicit knowledge and spatial relationship not explicitly stored in spatial databases. But is has also been observed that no GIS system with significant spatial data mining functionality is currently available. Therefore, as a part of solution to the spatial data, different data mining algorithms can be implemented. So that multi-end-users gets benefitted from multiple spatial data mining approaches. This approach is done by integrating all implemented methods in a single environment and thus reduce the user's efforts in planning their management action.

On the other hand, the enormous growth of networks in the hybrid domain has resulted some benefits like possessing a distributed data mining system that extracts knowledge from large local or global databases stored at multiple sites. The JAM system [9], intended for learning from such databases, is a distributed, scalable and Portable agent-based data mining software package that employs a general approach to scaling data mining applications. JAM provides a set of learning programs that compute models from data stored locally at a site, and a set of methods for combining multiple models learned at different sites. However, the JAM software system doesn't provide any tools for spatial data analysis.

Considering our software system, the data mining approach is flexible enough to attempt fetching any spatial data in centralized or distributed environments. In addition to providing an integrated tool for more systematic experimentation to data mining professionals, our software system offers user friendly environment / GUI including targeting non-technical people. The chief objective of this system is to construct a test environment for both standard and spatial data mining algorithms, that could quickly generate performance statistics (e.g. prediction accuracy), compare various algorithms on multiple data sets, implement hybrid algorithms (e.g. boosting, non-linear regression

trees) and graphically display intermediate results, learned structure and final – prediction results. To efficiently achieve this goal, we have introduced a GISBA system which executes programs developed in different environments (C, C++, MATLAB) through a unified modeling controlled approach and a simple Graphical User Interface (GUI).

II. SOFTWARE ORGANIZATION ARCHITECTURE

Software Organization

The Organization of the GISBA software system, shown in Figure 1, represents an integration of data mining algorithms in different programming environments under a unique GUI.

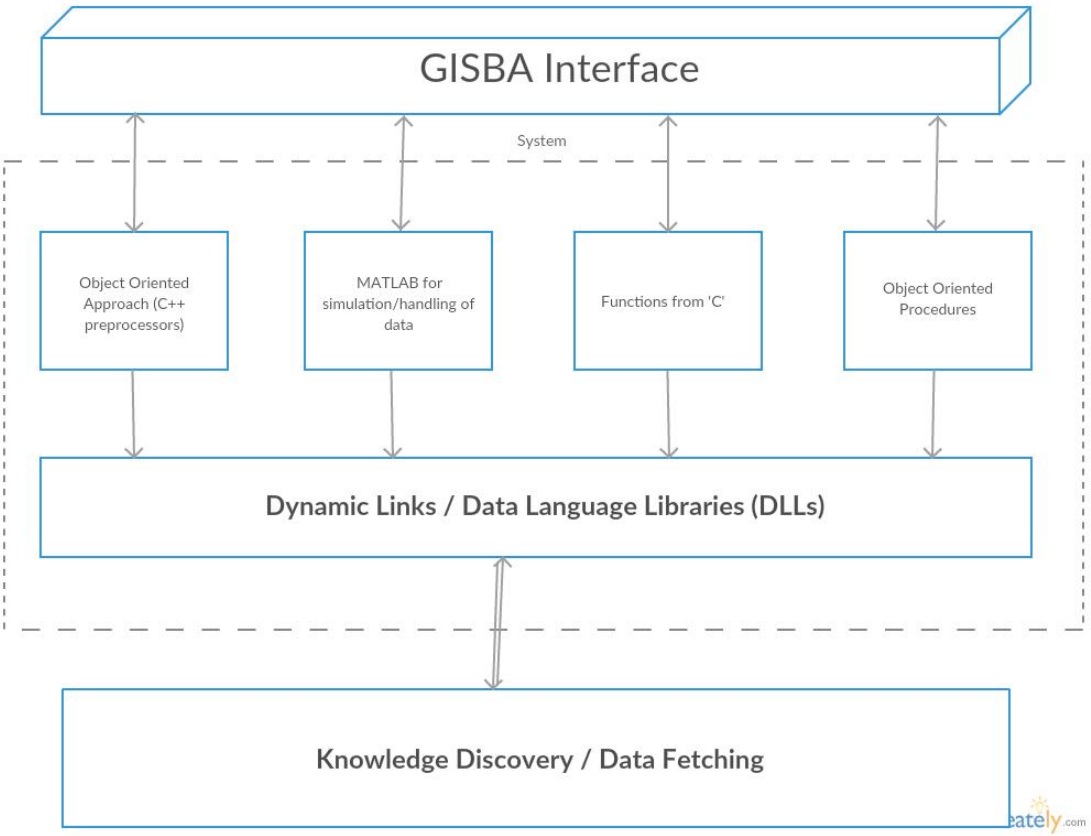

Figure1. An organization of GISBA

The interface allows a user to easily select data mining algorithms and methods. In brief, developed in the Visual Development studio environment, algorithms in MATLAB environment, including C and C++; Visual MATCOM software and

ActiveX controls are been incorporated into the proposed system, thus increasing the compile time and hence the runtime efficiency. Now, coming to GISBA, it is capable to run either from a local or from remotely through connection software, whether Internet connected Machine or LAN without Internet connection (Figure 1). User has accessibility of data from both local and remote machines however, keeping view of the security, a challenge-response password scheme is used in which initially the server sends a random series of bytes as a password and these passwords are encrypted, the server checks them against the 'right' answer after which the decrypted data is used. This though been trapped by malicious users, will be a bit harder to snoop this kind of session than other standard protocols. Learning algorithms help to build prediction models for each remote data set which will be "synthesized" later in order to use the knowledge from all available data sets thus achieving high prediction level. A more advanced distributed GISBA software system includes model and data management over multiple distributed sites under central GUI (Figure 2).

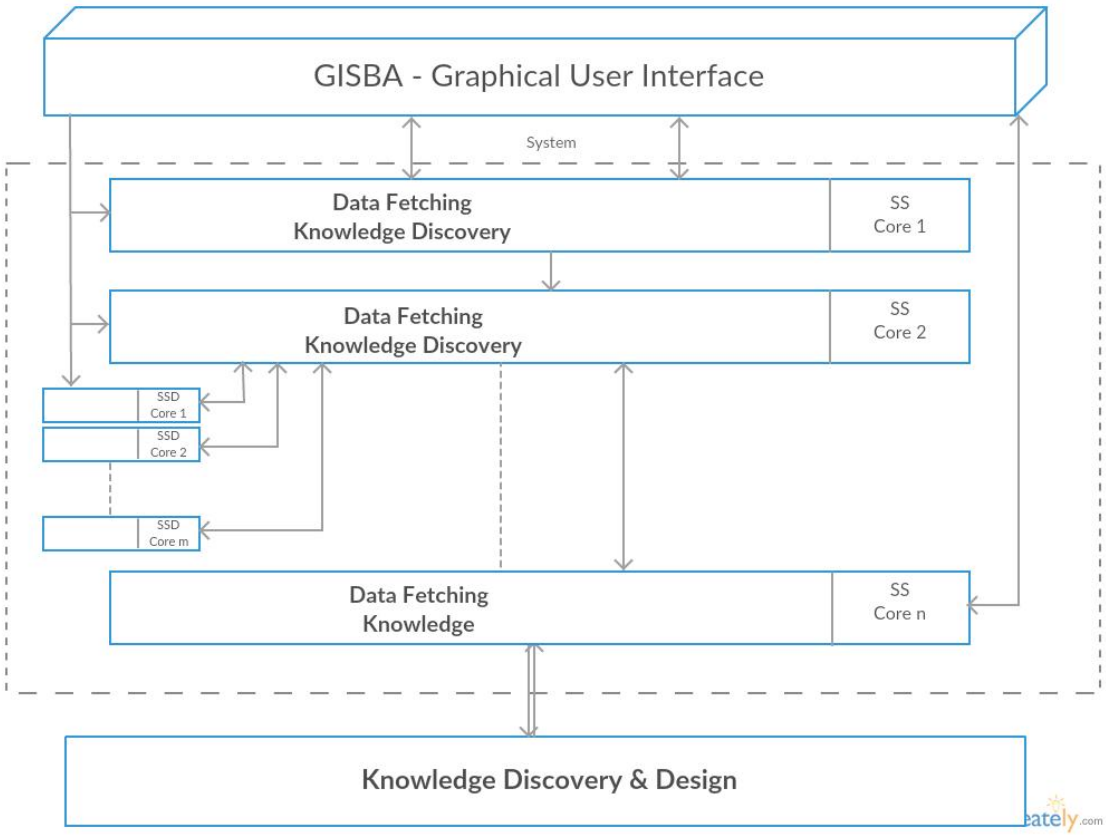

Figure2. The organization of the distributed GISBA Software System

In the figure 2, the package file transfers on a software which is connected at the remote station. A local modeling is being built up by spatial datamining activities at distributed

data bases. All these operations are being done without changing the entity of RAW data. Now the same operations is been applied to the other remote locations via central GUI so as to achieve better global prediction accuracy. Though there are several ways for achieving better accuracy of predictions, here in our first approach, every 'N' users (Figure 2) use some learning algorithm on one or more than one local spatial databases which produces local classifiers. All of these classifiers get combined into a new global classifier by sending them to the central repository, using majority or weighted principle. Finally, each and every user receives the classifier. The second approach that has been incorporated in this paper is combining the classifiers where every user sends their own classifiers using the same methods as before. One complex methods for combining classifiers include boosting over very large distributed spatial data sets by boosting iteration to select a small random sample of data at each distributed site.

GISBA Architecture

In view of complexity nature of spatial data, their modeling and analyzing, the workflow is being divided into six process steps: generation and implemented, data inspection, data preprocessing, data partitioning, modeling and model integration (Figure 3). Since not all spatial data analysis steps are necessary in the spatial data mining process, the data flow arrows in Figure 3 show which preprocessing steps can be skipped. The Figure 3 clarifies the connectivity of the modules, using the original data. Documentation is done through two procedures pre-modeling and post modeling, also called as constructing models. While in pre-modeling information mode, the following files are saved: History file containing the constructed resulting file, the operation performed, name and associated parameters, along with the resulting parameters after the operation. Afterwards, two resulting files are saved for every model: First file with sufficient information for saving of transforming the model to a different site and the second one a file containing all information necessary to describe this model to the user.

III. SOFTWARE FUNCTIONALITIES

The ADAM software system is designed to support the whole knowledge discovery process. Although SDAM includes numerous functions useful for non-spatial data, the system is intended primarily for spatial data mining, and so in this section we focus on spatial aspects of data analysis and modeling.

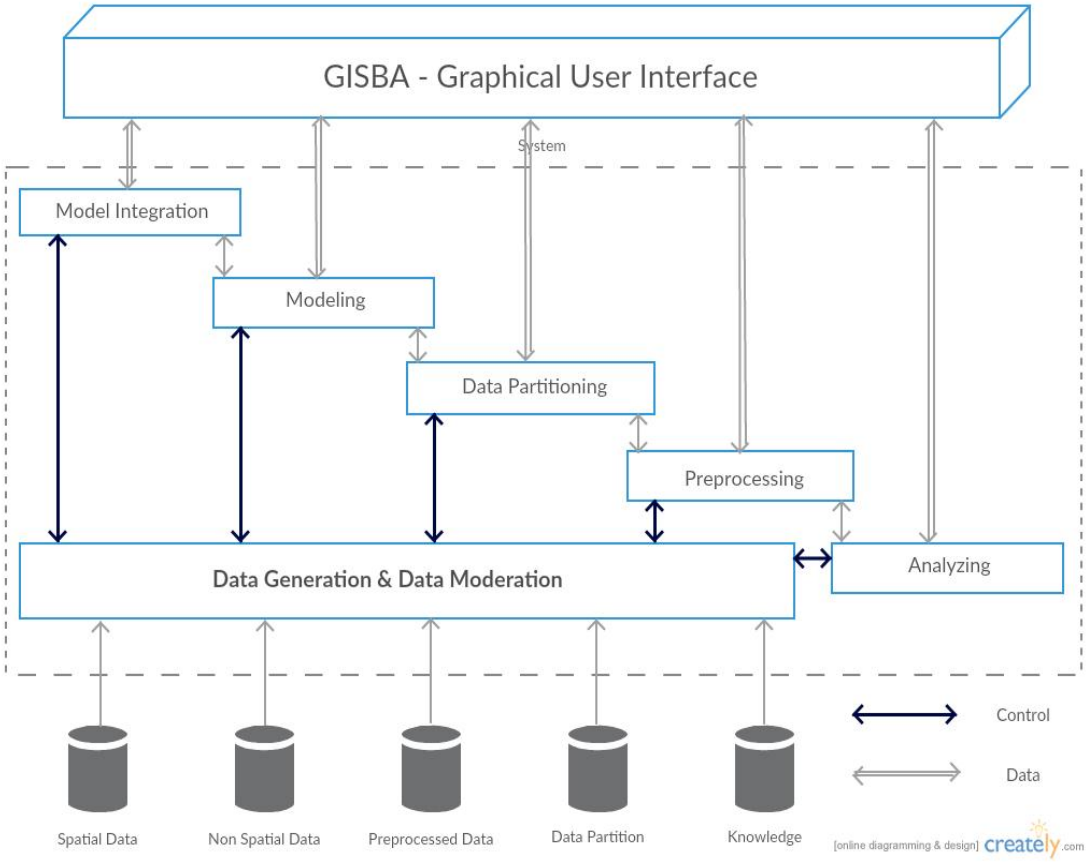

Figure 3. Internal architecture of RSSDAMSS

III. Software Functionalities

GISBA contains large number of "non-spatial data" oriented functions, prime focus is one spatial data mining and thus below is the spatial aspects of data analysis and modeling.

Description of the GISBA Functions

Figure 4 elaborates various functions of GISBA; here DDBMS despite of maintaining the distribution in databases, also serves the requirements from the GUI and including the function of the sub tasks like *Data generation and manipulation*, Data inspection, Data preprocessing, Data Normalization, Feature selection and extraction, Data partitioning especially when user authority management provides multiuser support.

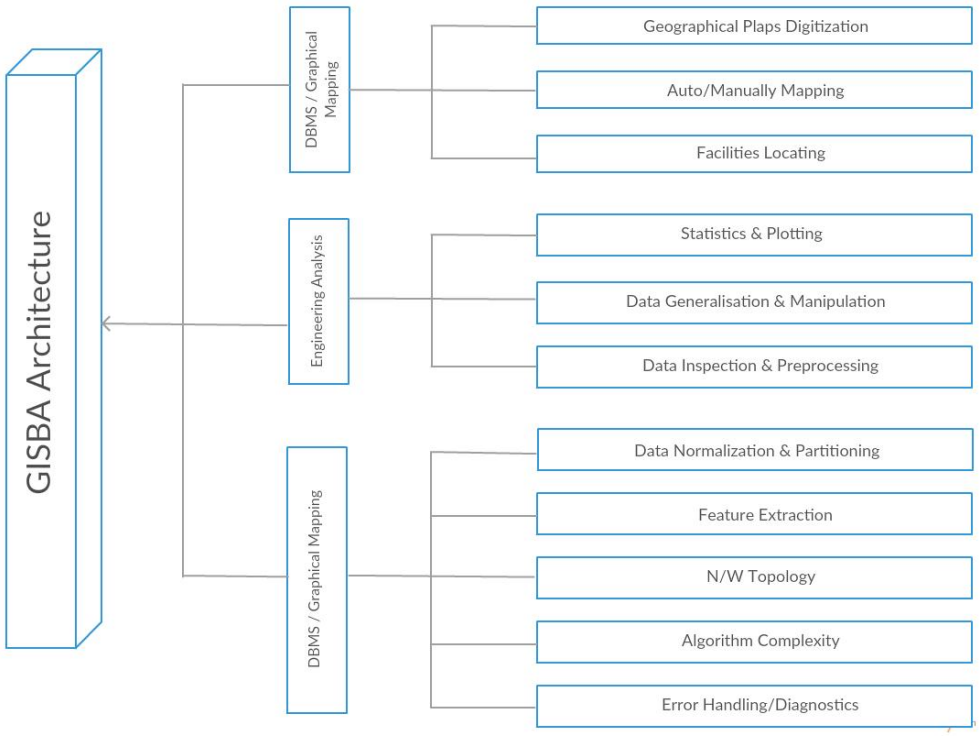

Figure 4. Functions of GISBA Software system

IV. CONCLUSION AND FUTURE WORK

GISBA Software conclusively provides unified and controlled Spatial Data Mining algorithmic techniques along with extensions of non-spatial algorithms for professional data analytics. Secondly it provides the impact of driving attributes and prediction simultaneously for the same actions, for the domain experts. The most important aspect of this work is the support for the remote control of a centralized GISBA software system through LAN and World Wide Web is useful when data are located at a distant location (e.g. a farm in precision agriculture), while a distributed GISBA allows knowledge integration from data located at multiple sites. The system provides extendible environment to include additional data mining algorithms, accordingly functions and furthermore, more advanced distributed aspects in the GISBA software system will be further developed. This paper conclusively focusses on the enhancement of the power and efficiency of the data mining algorithms on very large spatial databases, the discovery of more sophisticated algorithms for remotely sensed spatial data management, and the development of effective and efficient learning algorithms for distributed environments.

REFERENCES

[1] Whei-Min Lin, Ming-Tong Tsay and Su-Wei-Wu, "Application of Geographic Information System to Distribution Information Support", IEEE Transaction on Power system, Vol. 11, No. 1, Feb 1993.

[2] Kohonen, T, "The self-organizing map", *Proceedings of the IEEE*, 78, pp. 1464-1480, 1990.

[3] Kohave, R., Sommerfield D., Dougherty J., "Data Mining using MLC++", *International Journal of Artificial intelligence tools*, Vol. 6, No. 4, pp. 537-566,1997.

[4] Khabaza, T., Shearer, C., "Data mining with Clementine", *IE Colloquium on Knowledge Discovery in Databases*", Digest No. 1995/021(B), pp. 1/1-1/5. London, IEE, 1995.

[5] Quinlan, J. R., "Induction of decision trees", *MachineLearning*, Vol. 1., pp. 81-106, 1986.

[6] Werbos, P., *Beyond Regression: New Tools for Predicting and Analysis in the Behavioral Sciences*, Harvard University, Ph.D. Thesis, 1974. Reprinted by Wiley and Sons, 1995.

[7] Wayne Carr, P.E., Milsoft Integrated solutions, Inc. "Interfacing Engineering Analysis With Automated mapping", Paper No. 94 A4, 1994 Conference IEEE.

[8] Richardson, T., Stafford-Fraser, Q., Wood, K, R, Hopper, A., "Virtual Network Computing", *IEEE Internet Computing*, Vol.3 No.1, pp33-38, Jan/Feb 1998.

[9] Han, J., Koperski, k., Stefanovic, N., " GeoMiner: A System Prototype for Spatial Data Mining", *Proc 1997 ACM-SIGMOD In'l Conf. on management of Data (SIGMOD'97), Tucson*, May 1997.

[10] Han, J., Chiang, J., Chess, S.., Chen, J., Chen, Q., Cheng, S., Gong, q., Kamber, M., Koperski, K., Liu, G., Lu, Y., Stefanovic, N., Winstone, L., Xia, B., Zaiane, O.R., Zhang, S., Zhu, H., "DBMiner: A System for Data Mining in Relational Databases and Data Warehouses", *Proc. CASCON'97: Meeting of Minds*, Toronto, Canada, November 1997.

[11] Stolfo, S.J., Prodromidis, A.L., Tselepis, S., Lee, W., Fan, D., Chan, P.K., "JAM: Java Agents for Metalearning Over Distributed Databases," *Proc. KDD-97 and AAAI97 Work. On AI Methods in Fraud and Risk Management,* 1997.

[12] MATLAB, *The Language of Technical Computing*, The MathWorks Inc., January 1998.

[13] MIDEVA, *MATCOM & Visual Matcom, User's Guide*, MathTools Ltd., October 1998.

[14] Fan W., Stolfo S. and Zhang J. "The Application of AdaBoost for Distributed, Scalable and On-line Learning", *Proceeding of The Fifth ACM SIFKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, August 1999.