A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining

Dr. D. Ashok Kumar[*1], N. Kannathasan[#2]

[*1]*Government Arts College, Tiruchirapalli- 620 022, Tamil Nadu, India*

[*2]*Mahatma Gandhi Government Arts College, Mahe, P.O. New Mahe-673 311, U. T of Puducherry, India*

Abstract— Data mining has emerged as one of the major research domain in the recent decades in order to extract implicit and useful knowledge. This knowledge can be comprehended by humans easily. Initially, this knowledge extraction was computed and evaluated manually using statistical techniques. Subsequently, semi-automated data mining techniques emerged because of the advancement in the technology. Such advancement was also in the form of storage which increases the demands of analysis. In such case, semi-automated techniques have become inefficient. Therefore, automated data mining techniques were introduced to synthesis knowledge efficiently. A survey of the available literature on data mining and pattern recognition for soil data mining is presented in this paper. Data mining in Agricultural soil datasets is a relatively novel research field. Efficient techniques can be developed and tailored for solving complex soil datasets using data mining.

Keywords— ***Data Mining, Pattern Recognition, Soil Data Mining***

I. INTRODUCTION

This Data mining software applications includes various methodologies that have been developed by both commercial and research centers. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining has been used to analyze large datasets and establish useful classification and patterns in the datasets. Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods [16]. This paper outlines research which may establish if new data mining techniques will improve the effectiveness and accuracy of the Classification of large soil datasets. In particularly, this research work aims to compare the performance of the data mining algorithms with soil limitations and soil conditions in respect of the following characteristics: Acidity, Alkalinity and sodicity, Salinity, Low cation exchange capacity, Phosphorus fixation, Cracking and swelling properties, Depth, Soil density and Nutrient content. The use of standard statistical analysis techniques is both time consuming and expensive. If alternative techniques can be found to improve this process, an improvement in the classification of soils may result.

In many developing countries, hunger is forcing people to cultivate land that is unsuitable for agriculture and which can only be converted to agricultural use through enormous efforts and costs, such as those involved in the construction of terraces. Each country is known for its core competence. India's is agriculture. Yet, it only accounts for 17 per cent of the total Gross Domestic Product. With the pressure of urbanization, it is going to be a challenge to produce food for more people with less land and water.

Agriculture or farming forms the backbone of any country economy, since a large population lives in rural areas and is directly or indirectly dependent on agriculture for a living. Income from farming forms the main source for the farming community. The essential requirements for crop harvesting are water resources and capital to buy seeds, fertilizers, pesticides, labor etc. Most farmers raise the required capital by compromising on other necessary expenditures, and when it is still insufficient they resort to credit from sources like banks and private financial institutions. In such a situation, the repayment is dependent on the success of the crop. If the crop fails even once due to several factors, like bad weather pattern; soil type; improper, excessive, and untimely application of both fertilizers and pesticides; adulterated seeds and pesticides etc. then he is pushed into an acute crisis causing severe stress [58]. In addition, the plant growth depends on multiple factors such as soil type, crop type, and weather. Due to lack of plant growth information and expert advice, most of the farmers fail to get a good yield.

Most knowledge of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining the soil types or other properties of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey are acquired by field sampling and supported by remote sensing.

The test dataset using for this research work collected from World Soil Information – ISRIC (International Soil Reference and Information Centre). Version 3.1 of the ISRIC-WISE database (WISE3-World Inventory of Soil Emission Potentials) was complied from a wide range of soil profile data collected by many soil professionals world wide. All profiles have been harmonized with respect to the original Legend (1974) and Revised Legend (1988) of FAO-Unesco. Thereby the primary soil data and any secondary data derived from them can be linked using GIS to the spatial units of the soil map of the world as well as more recent Soil and Terrain (SOTER) databases through the soil legend code.

WISE3 is a relational database, compiled using MS-ACCESS. It can handle data on: (a) soil classification; (b) soil horizon data; (c) source of data; and methods used for determining analytical data. Profile data in WISE3 originate from over 260 different sources, both analogue and digital. Some 40% of the profiles were extracted from auxiliary datasets, including various Soil and Terrain (SOTER) databases and the FAO Soil Database (FAO-SDB), which, in turn, hold data collated from a wide range of sources.

WISE3 holds selected attribute data for 10,253 soil profiles, with some 47,800 horizons, from 149 countries. Individual profiles have been sampled, described, and analyzed according to methods and standards in use in the originating countries. There is no uniform set of properties for which all profiles have analytical data, generally because only selected measurements were planned during the original surveys. Methods used for laboratory determinations of specific soil properties vary between laboratories and over time. Some times, results for the same property cannot be compared directly. WISE3 will inevitably include gaps, being a compilation of legacy soil data derived from traditional soil survey. These can be of a taxonomic, geographic, and soil analytical nature. As a result, the amount of data available for modeling is some times much less than expected. Adroit use of the data, however, will permit a wide range of agricultural and environmental applications at a global and continental scale (1:500000 and broader) [44].

The analysis of these datasets with various data mining techniques may yield outcomes useful to researchers in future.

II. MATERIALS AND METHODS

The rapid growth of interest in data mining is due to the (i) falling cost of large storage devices and increasing ease of collecting data over networks, (ii) development of robust and efficient machine learning algorithms to process this data, and (iii) falling cost of computational power, enabling use of computationally intensive methods for data analysis [37].

Data Mining (DM) represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data [18]. The DM process has developed due to the immense volume of data that must be handled easier in areas such as: business, medical industry, astronomy, genetics or banking field. Also, the success and the extraordinary development of hardware technologies led to the big capacity of storage on hard–disks, fact that challenged the appearance of many problems in manipulating immense volumes of data. Of course the most important aspect here is the fast growth of the Internet.

The core of the DM process lies in applying methods and algorithms in order to discover and extract patterns from stored data but before this step data must be pre–processed. It is well known that simple use of DM algorithms does not produce good results. Thus, the overall process of finding useful knowledge in raw data involves the sequential adhibition of the following steps: developing an understanding of the application domain, creating a target dataset based on an intelligent way of selecting data by focusing on a subset of variables or data samples, data cleaning and pre–processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, the data mining step, interpreting mined patterns with possible return to any of the previous steps and consolidating discovered knowledge.

The DM contains many study areas such as machine–learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization. The most important goal here is to extract patterns from data and to bring useful knowledge into an understandable form to the human observer. It is recommended that obtained information to be facile to interpret for the easiness of use. The entire process aims to obtain high–level data from low level data.

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Typically, a data mining algorithm constitutes some combination of the following three components.

- The model: The function of the model (e.g., classification, clustering) and its representational form (e.g. linear discriminants, neural networks). A model contains parameters that are to be determined from the data.
- The preference criterion: A basis for preference of one model or set of parameters over another, depending on the given data.
- The search algorithm: The specification of an algorithm for finding particular models and parameters, given the data, model(s), and a preference criterion.

A particular data mining algorithm is usually an instantiation of the model/preference/search components. The more common model functions in current data mining practice include:

1. Classification [41], [38], [42], [6], [39]: classifies a data item into one of several predefined categorical classes.
2. Regression [19], [12], [64], [45]: maps a data item to a real valued prediction variable.
3. Clustering [61], [50], [47], [52], [29], [31], [62], and [21]: maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models.
4. Rule generation [60], [35], [40], [43], [23], [55], [53], [67]: extracts classification rules from the data.
5. Discovering association rules [2], [63], [5], and [34]: describes association relationship among different attributes.
6. Summarization [32], [65], [25], [20]: provides a compact description for a subset of data.

7. Dependency modeling [22], [7]: describes significant dependencies among variables.
8. Sequence analysis [10], [33]: models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time.

Though, there are lots of techniques available in the data mining, few methodologies such as Artificial Neural Networks, K nearest neighbor, K means approach, are popular currently depends on the nature of the data.

Artificial Neural Network: Artificial Neural Networks (ANN) is systems inspired by the research on human brain (Hammerstrom, 1993). Artificial Neural Networks (ANN) networks in which each node represents a neuron and each link represents the way two neurons interact. Each neuron performs very simple tasks, while the network representing of the work of all its neurons is able to perform the more complex task. A neural network is an interconnected set of input/output units where each connection has a weight associated with it. The network learns by fine tuning the weights so as able to predict the call label of input samples during testing phase. Artificial neural network is a new techniques used in flood forecast. The advantage of ANN approach in modeling the rain fall and run off relationship over the conventional techniques flood forecast. Neural network has several advantages over conventional method in computing. Any problem having more time for getting solution, ANN is highly suitable states that the neural network method successfully predicts the pest attack incidences for one week in advance.

Pedotransfer functions (PTFs) provide an alternative by estimating soil parameters from more readily available soil data. The two common methods used to develop PTFs are multiple-linear regression method and ANN. Multiple linear regression and neural network model (feed-forward back propagation network) were employed to develop a pedotransfer function for predicting soil parameters using easily measurable characteristics of clay, sand, silt, SP, Bd and organic carbon[51].

Artificial Neural Networks have been successful in the classification of other soil properties, such as dry land salinity (Spencer *et al.* 2004). Due to their ability to solve complex or noisy problems, Artificial Neural Networks are considered to be a suitable tool for a difficult problem such as the estimation of organic carbon in soil.

Support Vector Machines: Support Vector Machines (SVM) is binary classifiers (*Burges, 1998; Cortes and Vapnik, 1995*). SVM is able to classify data samples in two disjoint classes. The basic idea behind is classifying the sample data into linearly separable. Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. In simple words given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

SVM is used to assess the spatiotemporal characteristics of the soil moisture products [4].

Decision trees: The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Decision tree is a new field of machine learning which is involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees and discrimination nets or production rules. Application of data mining techniques on drought related data for drought risk management shows the success on Advanced Geospatial Decision Support System (GDSS). Leisa J Armstrong states that data mining approach is one of the approaches used for crop decision making.

Research has been conducted in Australia to estimate a range of soil properties, including organic carbon (Henderson *et al.* 2001). The nation-wide database had 11,483 soil points available to predict organic carbon in the soil. An enhanced decision trees tool (Cubist), catering for continuous outputs was used for this study. A correlation of up to 0.64 was obtained between the predicted and actual organic carbon levels.

K nearest neighbor: K nearest neighbor techniques is one of the classification techniques in data mining. It does not have any learning phase because it uses the training set every time a classification performed. Nearest Neighbor search (NN) also known as proximity search, similarity search or closest point search is an optimization problem for finding closest points in metric spaces.

K nearest neighbor is applied for simulating daily precipitation and other weather variables (*Rajagopalan and Lall, 1999*).

Bayesian networks: A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data Development of a data mining application for agriculture based on Bayesian networks were studied by *Huang et al. (2008)*. According to him, Bayesian network is a powerful tool for dealing uncertainties and widely used in agriculture datasets. He developed the model for agriculture application based on the Bayesian network learning method. The results indicate that Bayesian Networks are a feasible and efficient.

Bayesian approach improves hydrogeological site characterization even when using low-resolution resistivity surveys [52].

K means approach: K means method is one of the most used clustering techniques in the data mining. The idea behind the K means algorithms is very simple that certain partition of the data in K clusters, the centers of the cluster can be computed as the mean of the all sample belonging to a cluster. The center of the cluster can be considered as the representative of the cluster. The center is quite close to all samples in the cluster.

K Means approach was used to classify the soil and plants (*Camps-Valls et al., 2003*).

Fuzzy logic: Fuzzy logic is a form of multi valued logic derived from Fuzzy set theory to deal with reasoning that is approximate rather than accurate. In contrast with "crisp logic", where binary sets have binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1 and is not constrained to the two truth values of classic propositional logic [46]. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions. Fuzzy logic emerged as a consequence of the 1965 proposal of Fuzzy set theory by Lotfi zadeh [1] [66]. Though fuzzy logic has been applied to many fields, from control theory to artificial intelligence, it still remains controversial among most statisticians, who prefer Bayesian logic, and some control engineers, who prefer traditional two-valued logic.

Fuzzy logic is used to the prediction of soil erosion in a large watershed (B.Mitra et al., ScienceDirect, Nov.1998).

Genetic Algorithm: The Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of Evolutionary Algorithm (EA), which generates solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover.

Soil liquefaction is a type of ground failure related to earthquakes. It takes place when the effective stress within soil reaches zero as a result of an increase in pore water pressure during earthquake vibration (Youd, 1992). Soil liquefaction can cause major damage to buildings, roads, bridges, dams and lifeline systems, like the earthquakes.

Genetic Algorithm approach is used for assessing the liquefaction potential of sandy soils (G. Sen et al. Nat. Hazards Earth Syst. Sci., 2010).

Ant Colony Optimization: The Ant Colony Optimization (ACO) algorithm is probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of ant colony algorithms family, in swarm intelligence methods, and it constitutes some Meta heuristic optimizations. Initially proposed by Marco Dorigo in 1992 in his Ph.D. thesis [13] [17], the first algorithm was aiming to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food. The original idea has since diversified to solve a wider class of numerical problems, and as a result, several problems have emerged, drawing on various aspects of the behavior of ants.

Ant Colony Optimization is applied for estimating unsaturated soil hydraulic parameters (K.C.Abbaspour *et al, ELSEVIER*, 2001).

Particle Swarm Optimization: Particle Swarm Optimization (PSO) is a method for performing numerical optimization without explicit knowledge of the gradient of the problem to be optimized. PSO is originally attributed to Kennedy, Eberhart, and Shri [28] [54] and was first intended for simulating social behavior. The algorithm was simplified and it was observed to be performing optimization. The book by Kennedy and Eberhart [27] describes many philosophical aspects of PSO and swarm intelligence. An extensive survey of PSO applications is made by Poli [48] [49].

Particle Swarm Optimization is used for analysis of Soil erosion characteristics (Li Yunkai et al, Springer, Sep.2009).

Simulated Annealing: Simulated Annealing (SA) is a generic probabilistic Meta heuristic for the global optimization problem of applied mathematics, namely locating a good approximation to the global optimum of a given function in a large search space. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For certain problems, simulated annealing may be more effective than exhaustive enumeration provided that the goal is merely to find an acceptably good solution in a fixed amount of time, rather than the best possible solution. The method was independently described by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983 [30] and by Vlado Cerny in 1985 [9]. The method is an adaptation of the Metropolis Hastings algorithm, a Monte Carlo method to generate sample states of a thermodynamic system, invented by N. Metropolis et al. in 1953 [36].

Simulated Annealing is used for analyzing Soil Properties (R.M. Lark et al., ScienceDirect, March, 2003).

III. RESULTS AND DISCUSSION

The purpose of the study is to examine the most effective techniques to extract new knowledge and information from existing soil profile data contained within ISRIC-WISE soil data set. Several data mining techniques are in agriculture and allied area. Few of techniques are discussed here. K means method is used to forecast the pollution in the atmosphere (*Jorquera et al., 2001*). Different possible changes of weather are analyzed using SVM (*Tripathi et al., 2006*). K means approach is used for classifying soil in combination with GPS readings (*Verheyen et al., 2001*). Wine Fermentation process monitored using data mining techniques. Taste sensors are used to obtain data from the fermentation process to be classified using ANNs (*Riul et al., 2004*).

A brief survey of the related work in the area of soil mining is that the data involved here are high dimensional data and

dimensionality reduction was addressed in classical methods such as Principal Component Analysis (PCA) [24]. There is a growing literature demonstrating the predictive capacity of the soil landscape paradigm using digital data and empirical numerical modeling techniques as specified by Christopher et al., [11]. The Eigen decomposition of empirical covariance matrix is performed and the data points are linearly projected. When the information relevant for classification is present in eigenvectors associated with small eigenvalues are removed, then this could lead to degradation in classification accuracy. Examples of spatial prediction have been provided, across a range of physiographical range of environment and spatial extents, for a number of soil properties by Gessler et al., [21] Tenenbaum et al.,[59] introduced the concept of Isomap, a global dimensionality reduction algorithm. The CCDR (classification constrained dimensionality reduction) algorithm [15] was only demonstrated for two classes and the performance was analyzed for simulated data. Bui et al., [8] demonstrated the potential for the discovery of knowledge embedded in survey of landscape model using rule induction techniques based on decision trees. It has the ability to mimic soil map using samples taken from it, and by implication it also captures the embedded knowledge. Related to agriculture, many countries are still facing a multitude of problems to maximize productivity [26]. Another concept of CCDR plots the classification error probability and its confidence interval using K nearest neighbour classifier [14]. Normally there is a decrease in error probability as dimension increases, and the optimal value is reached when dimension value varies between 12 - 14, which has been proved using entropic graph algorithm. However the food production has improved significantly during last two decades by providing it with good seeds, fertilizers, and pesticides and modern farming equipment [57]. The agriculture sector has seen a tremendous improvement.

IV. CONCLUSIONS

In this research survey, data mining and pattern recognition techniques for soil data mining studied. The survey aims to come out of the techniques being used in the agricultural soil science and its allied area.

The recommendations arising from this research survey are: A comparison of different data mining techniques could produce an efficient algorithm for soil classification for multiple classes. The benefits of a greater understanding of soils could improve productivity in farming, maintain biodiversity, reduce reliance on fertilizers and create a better integrated soil management system for both the private and public sectors.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions.

.

REFERENCES

[1] "Fuzzy Logic". Stanford Encyclopedia of Philosophy. Stanford University. 2006-07-23. Retrieved 2008-09-29.

[2] Agrawal R., Imielinski T., and Swami A., Mining association Rules between sets of items in large databases, in Proceedings of 1993 ACM SIGMOD International Conference on Management of Data, (Washington D.C.), pp. 207-216, May 1993.

[3] Alahakoon D., Halgamuge S.K., and Srinivasan B, Dynamic self organizing maps with controlled growth for knowledge discovery, IEEE Transactions on Neural Networks, vol. 11, pp. 601-614, 2000.

[4] Anish C. Turlapaty, Valentine Anantharaj, Nicolas H. Younan, Spatio-temporal consistency analysis of AMSR-E soil moisture data using wavelet-based feature extraction and one-class SVM, In the Proceedings of the Annual Conference Baltimore, Maryland, March 9-13, 2009.

[5] Au W. H. and Chan K. C. C., An effective algorithm for discovering fuzzy rules in relational databases, in Proceedings of IEEE International Conference on Fuzzy Systems FUZZ IEEE 98, (Alaska), pp. 1314-1319, May 1998.

[6] Banerjee M, Mitra S, and Pal S.K, Rough fuzzy MLP: Knowledge encoding and classification, IEEE Transactions on Neural Networks, vol. 9, pp. 1203-1216, 1998.

[7] Bosc P., Pivert O., and Ughetto L., Database mining for the discovery of extended functional dependencies, in Proceedings of NAFIPS 99, (New York, USA), pp. 580-584, June 1999.

[8] Bui E. N., Loughhead A. and Comer R... Extracting Soil Landscape Rules from Previous Soil Surveys. Australian Journal of Soil Science, 37:495508, 1999.

[9] Cerny V., A thermo dynamical approach to the traveling salesman problem: an efficient simulation algorithm. Journal of Optimization Theory and Applications, 45:41-51, 1985.

[10] Chiang D. A., Chow L. R., and Wang Y. F., Mining time series data by a fuzzy linguistic summary system," Fuzzy Sets and Systems, vol. 112, pp. 419-432, 2000.

[11] Christopher J. Moran and Elsabeth Bui N., Spatial Data Mining for Enhanced Soil Map Modeling. In the Proceedings of the International Journal of Geographical Information Science, 2002.

[12] Ciesielski V and Palstra G, Using a hybrid neural/expert system for database mining in market survey data, in Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96), (Portland, OR), p. 38, AAAI Press, Aug. 2-4, 1996.

[13] Colorni A., Dorigo et M., Maniezzo V., Distributed Optimization by Ant Colonies, actes de la première conference euro penne sur la vie artificielle, Paris, France, Elsevier Publishing, 134-142, 1991.

[14] Costa A. and Hero A. O. Geodesic Entropic Graphs for Dimension and Entropy Estimation in Manifold Learning. In the Proceedings of IEEE Transaction Signal Processing, volume 52, pages 2210-2221, 2004.

[15] Costa J. A. and Hero A. O., III. Classification Constrained Dimensionality Reduction. In IEEE International Conference on Acoustic Speech, and Signal Processing, volume 5, pages 1077-1080, March 2005.

[16] Cunningham S. J and Holmes G. Developing innovative applications in agriculture using data mining, In the Proceedings of the Southeast Asia regional Computer Confederation Conference, 1999.

[17] Dorigo M., Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italie, 1992.

[18] Fayadd, U., Piatesky–Shapiro, G., and Smyth, P, Data Mining to Knowledge Discovery in Databases, AAAI Press / the MIT Press, Massachusetts Institute of Technology. ISBN 0–262–56097–6 Fayap, 1996.

[19] Fayyad U.M, Piatetsky-Shapiro G, Smyth P., and Uthurusamy R., eds., Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996.

[20] George R. and Srikanth R., Data summarization using genetic algorithms and fuzzy logic, in Genetic Algorithms and Soft Computing (F. Herrera and J. L. Verdegay, eds.), pp. 599-611, Heidelberg: Springer-Verlag, 1996.

[21] Gessler P. E., Moore D., McKenzie N. J. and Ryan P... Soil Landscape Modeling and Spatial Prediction of Soil Attributes. In the Proceedings

of the International Journal of Geographical Information Systems, volume 9, pages 421-432, 1995.

[22] Hale J. and Shenoi S., Analyzing FD inference in relational databases, Data and Knowledge Engineering, vol. 18, pp. 167-183, 1996.

[23] Hu X. and Cercone N., Mining knowledge rules from databases: A rough set approach, in Proceedings of the 12th International Conference on Data Engineering, (Washington), pp. 96-105, IEEE Computer Society, Feb. 1996.

[24] Jain A. K. and Dubes R. C., Algorithm for Clustering Data. Prentice Hall, 1998.

[25] Kacprzyk J. and Zadrozny S., Data mining via linguistic summaries of data: an interactive approach, in Proceedings of IIZUKA 98, (Fukuoka, Japan), pp. 668-671, October 1998.

[26] Katyal J. C., Paroda R. S., Reddy M. N., Aupam Varma and N. Hanumanta Rao. Agricultural Scientists Perception on Indian Agriculture: Scene Scenario and Vision. National Academy of Agricultural Science, 2000.

[27] Kennedy J. Eberhart R.C. Swarm Intelligence. Morgan Kaufmann. ISBN 1-55860-595-9., 2001

[28] Kennedy, J., Eberhart, R., "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks. IV. pp. 1942–1948, 1995.

[29] Kiem H. and Phuc D., Using rough genetic and Kohonen's Neural network for conceptual cluster discovery in data mining, in Proceedings of RSFDGrC'99, (Yamaguchi, Japan), pp. 448-452, November 1999.

[30] Kirkpatrick S., Gelatt C.D, Vecchi M.P. Optimization by Simulated Annealing. Science New Series 220 (4598):671–680.Doi:10.1126/science.220.4598.671. ISSN 00368075. , 1983-05-13

[31] Kohonen, Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V., and Saarela A., Self organization of a massive document collection, IEEE Transactions on Neural Networks, vol. 11, pp. 574-585, 2000.

[32] Lee D. H. and Kim M. H, Database summarization using fuzzy ISA hierarchies, IEEE Transactions on Systems Man and Cybernetics. Part B-Cybernetics, vol. 27, pp. 68-78, 1997.

[33] Lee R. S. T. and Liu J. N. K., Tropical cyclone identification and tracking system using integrated neural oscillatory leastic graph matching and hybrid RBF network track mining techniques, IEEE Transactions on Neural Networks, vol. 11, pp. 680-689, 2000.

[34] Lopes C., Pacheco M., Vellasco M., and Passos E., Rule evolver: An evolutionary approach for data mining, in Proceedings of RSFDGrC'99, (Yamaguchi, Japan), pp. 458-462, November 1999.

[35] Lu H.J., Setiono R., and Liu H., Effective data mining using neural networks, IEEE Transactions on Knowledge and Data Engineering, vol. 8, pp. 957-961, 1996.

[36] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E... Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21(6):1087-1092, 1953.

[37] Mitchell T.M, Machine learning and data mining, Communications of the ACM, vol. 42, no. 11, 1999.

[38] Mitra S and Pal S.K, Fuzzy self organization, inferencing and rule generation, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 26, pp. 608-620, 1996.

[39] Mitra S, Mitra P, and Pal S.K, Evolutionary modular Design of rough knowledge-based network using fuzzy attributes Neuro computing, vol. 36, pp. 45-66, 2001.

[40] Mitra S. and Hayashi Y., Neuro-fuzzy rule generation: Survey in soft computing framework, IEEE Transactions on Neural Networks, vol. 11, pp. 748-768, 2000.

[41] Mitra S. and Pal S.K, Fuzzy multi-layer perceptron, Inferencing and rule generation, IEEE Transactions on Neural Networks, vol. 6, pp. 51-63, 1995.

[42] Mitra S., De R.K, and Pal S.K, Knowledge-based fuzzy MLP for classification and rule generation, IEEE Transactions on Neural Networks, vol. 8, pp. 1338-1350, 1997.

[43] Mollestad T. and Skowron A., A rough set framework for data mining of propositional default rules," Lecture Notes in Computer Science, vol. 1079, pp. 448-457, 1996.

[44] Niels H. Batjes, ISRIC-WISE Harmonized Global Soil Profile Dataset (Ver. 3.1) - A Report -2008/2

[45] Noda E, Freitas A.A, and Lopes H.S, Discovering Interesting prediction rules with a genetic algorithm, in Proceedings of IEEE Congress on Evolutionary Computation CEC 99, (Washington DC), pp. 1322-1329, July 1999.

[46] Novak, V., Perfilieva, I. and Mockor, J. Mathematical principles of fuzzy logic Dodrecht: Kluwer Academic. ISBN 0-7923-8595-0, 1999.

[47] Pedrycz W, Conditional fuzzy c-means, Pattern Recognition Letters, vol. 17, pp. 625-632, 1996.

[48] Poli, R. An analysis of publications on Particle swarm optimization applications. Technical Report CSM-469 (Department of Computer Science, University of Essex, UK)., 2007

[49] Poli, R. Analysis of the publications on the Applications of particle swarm optimization". Journal of Artificial Evolution and Applications: 1–10. Doi:10.1155/2008/685175., 2008

[50] Russell S and Lodwick W, Fuzzy clustering in data mining for telco database marketing campaigns, in Proceedings of NAFIPS 99, (New York), pp. 720-726, June 1999.

[51] Sarmadian F., Taghizadeh R., Mehrjardi and. Akbarzadeh A, Optimization of Pedotransfer Functions Using an Artificial Neural Network, Australian Journal of Basic and Applied Sciences, 3(1): 323-329, ISSN 1991-8178., 2009,

[52] Shalvi D and De Claris N, Unsupervised neural network approach to medical data mining techniques, in Proceedings of IEEE International Joint Conference on Neural Networks, (Alaska), pp. 171-176, May 1998.

[53] Shan N. and Ziarko W., Data-based acquisition and incremental modification of classification rules, Computational Intelligence, vol. 11, pp. 357-370, 1995.

[54] Shi, Y. Eberhart, R.C., A modified particle swarm optimizer". Proceedings of IEEE International Conference on Evolutionary Computation. pp. 69–73., 1998

[55] Skowron A., Extracting laws from decision tables - a rough set approach, Computational Intelligence, vol. 11, pp. 371-388, 1995.

[56] Souheil Ezzedine, Yoram Rubin, and Jinsong Chen, Bayesian method for hydro geological site characterization using borehole and geophysical survey data: Theory and application to the Lawrence Livermore National Laboratory Superfund site, Water Resources Research, vol. 35, No. 9, Pages 2671–2683, September, 1999.

[57] Subba Rao. Indian Agriculture past Laurels and Future Challenges, Indian Agriculture: Current Status, Prospects and Challenges. Convention of Indian Agricultural Universities Association, 27:58-77, December 2002.

[58] Sudarshan Reddy S, Vedantha S, Venkateshwar Rao B, Sundar Ram Reddy and Venkat Reddy. Gathering Agrarian Crisis Farmers Suicides in Warangal district. *Citizens Report, 1998.*

[59] Tenenbaum J. B., De Silva and Langford C., A Global Geometric Framework for Dimensionality Reduction. 290(5500):2319-2323, 2000.

[60] Tickle A.B, Andrews R., Golea M., and Diederich J., The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks, IEEE Transactions on Neural Networks, vol. 9, pp. 1057-1068, 1998.

[61] Turksen I.B, Fuzzy data mining and expert system Development, in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, (San Diego, CA), pp. 2057-2061, October 1998.

[62] Vesanto J. and Alhoniemi E., Clustering of the self organizing map, IEEE Transactions on Neural Networks, vol. 11, pp. 586-600, 2000.

[63] Wei Q. and Chen G., Mining generalized association rules with fuzzy taxonomic structures, in Proceedings of NAFIPS 99, (New York), pp. 477-481, June 1999.

[64] Xu K, Wang Z, and Leung K.S, Using a new type of non Linear integral for multi-regression: an application of evolutionary Algorithms in data mining, in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, (San Diego, CA), pp. 2326-2331, October 1998.

[65] Yager R. R., On linguistic summaries of data, in Knowledge Discovery in Databases (W. Frawley and G. Piatetsky-Shapiro, eds.), pp. 347-363, Menlo Park, CA: AAAI/MIT Press, 1991.

[66] Zadeh L.A... "Fuzzy sets", Information and Control 8 (3): 338–353, 1965

[67] Zhang Y.Q., Fraser M.D., Gagliano R.A., and Kandel A., Granular neural networks for numerical-linguistic data fusion and knowledge discovery, IEEE Transactions on Neural Networks, vol.11, pp. 658-667,2000.

AUTHORS BIOGRAPHY

D. Ashok kumar did his Master degree in Mathematics and Computer Applications in 1995 and completed Ph.D., on Intelligent Partitional Clustering Algorithm's in 2008, from Gandhigram Rural Institute–Deemed University, Gandhigram, Tamil Nadu, INDIA. He is currently working as Senior Grade Assistant Professor and Head in the Department of Computer Science, Government Arts College, Tiruchirapalli–620 022, Tamil Nadu, INDIA. His research interest includes Pattern Recognition and Data Mining by various soft computing approaches viz., Neural Networks, Genetic Algorithms, Fuzzy Logic, Rough set, etc.,

N. Kannathasan is a Senior Grade Assistant Professor of Computer Science at the Mahatma Gandhi Government Arts College, Mahe, U.T. of Puducherry, INDIA. Prior to joining MGGAC, Mahe, he served at the Bharathidasan Government College for Women, Puducherry, Velammal College of Management and Computer Studies, Chennai, SRM Arts and Science College, Chennai, and AVC College, Mayiladuthurai. He received his M.Phil. Computer Science from the Bharathidasan University, Tiruchirappalli and M.Sc. Computer Science from Ayya Nadar Janaki Ammal College, Sivakasi.