WEB BASED DATA MINING THROUGH SOFT COMPUTING TECHNIQUE: A REVIEW

Ms. Kirti Gautam[1], Gurdev Singh[2]
[1] M.Tech student, Department of CSE Jind Institute of Engineering & Technology, Jind (Haryana)
[2]Assistant Professor, Department of CSE, Jind Institute of Engineering & Technology, Jind (Haryana)

Abstract: The present article provides a survey of available literature on data mining using soft computing. Web usage mining is process of extracting useful information from server logs e.g. use Web usage mining is process of finding out what users are looking for on Internet. Web intelligence is area of study & research of application of artificial intelligence & information technology on web within order to create next generation of products, services & frameworks based on internet. A categorization has been provided based on different soft computing tools & their hybridizations used, data mining function implemented, & preference criterion selected by model. The use of different soft computing methodologies is highlighted. Generally fuzzy sets are suitable for handling issues related to understandability of patterns, incomplete/noisy data, mixed media information & human interaction, & could provide approximate solutions faster. Neural networks are nonparametric, robust, & exhibit good learning & generalization capabilities within data-rich environments. Genetic algorithms provide efficient search algorithms to select a model, from mixed media data, based on some preference criterion/objective function. Some challenges to data mining & application of soft computing methodologies are indicated. In following section, evolution of Soft Computing techniques, its development & application as well as its impact within Data Mining have been highlighted.

Index Terms-Fuzzy logic, genetic algorithms, knowledge discovery, neural networks, neuro-fuzzy computing, rough sets, rule extraction.

1. Introduction

The digital revolution has made digitized information easy to capture & fairly inexpensive to store [1], [2]. With development of computer hardware/software and rapid computerization of business, large amount of data have been collected & stored within databases. The rate at which such data is stored is growing at a phenomenal rate. As a result, traditional ad hoc mixtures of statistical techniques & data management tools are no longer sufficient for analyzing this vast collection of data. Several domains where large volumes of data are stored within centralized or distributed databases include following:

- **Financial Investment:** Stock indexes & prices, interest rates, credit card data, fraud detection [3].

- **Health Care:** Several diagnostic information stored by hospital management systems [4].

- **Manufacturing & Production**: Process optimization & troubleshooting [5].

- **Telecommunication network**: Calling patterns & fault management systems.

- **Scientific Domain**: Astronomical observations [6], genomic data, biological data.

Raw data is rarely of direct benefit. Its true value is predicated on ability to extract information useful for decision support or exploration, & understanding phenomenon governing data source. In most domains, data analysis was traditionally a manual process. One or more analysts would become intimately familiar with data and, with help of statistical techniques, provide summaries & generate reports.

2. Data Mining

Data mining is an increasingly important branch of computer science that examines data within order to find & describe patterns. Because we live within a world where we could be overwhelmed with information, it is imperative that we find ways to classify this input, to find information we need, to illuminate structures, & to be able to draw conclusions. Data mining is a very practical discipline with many applications within business, science, & government, such as targeted marketing, web analysis, disease diagnosis & outcome prediction, weather forecasting, credit risk & loan approval, customer relationship modeling, fraud detection, & terrorism threat detection. It is based on methods several fields, but mainly machine learning, statistics, databases, & information visualization.

All these have prompted need for intelligent data analysis methodologies, which could discover useful knowledge from data. The term KDD refers to overall process of **Knowledge Discovery within Databases**. Data mining is a particular step within this process, involving application of specific algorithms for extracting patterns (models) from data. A good overview of KDD could be found within Ref. [8], [9].

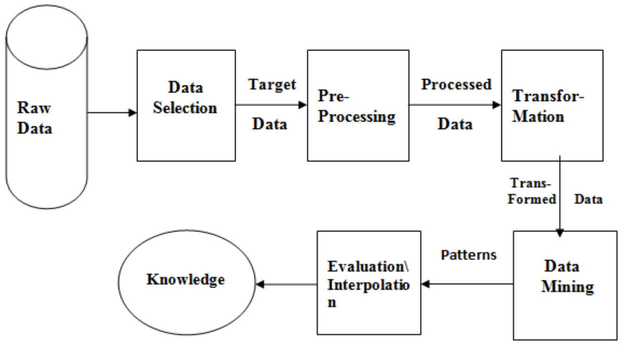

Fig.1. Block diagram for knowledge discovery & data mining

Knowledge Discovery within Databases (KDD) techniques perform data analysis & may uncover important data patterns, contributing greatly to business strategies, knowledge bases, & scientific & medical research. Data mining is an essential step within process of knowledge discovery within databases. Knowledge discovery as a process consists of an iterative sequence of following steps [17]:

i. **Understanding application domain:** includes relevant prior knowledge & goals of application.

ii. **Extracting target data set:** includes selecting a data set or focusing on a subset of variables.

iii. **Data cleaning & preprocessing:** includes basic operations, such as noise removal & handling of missing data. Data from real-world sources are often erroneous, incomplete, & inconsistent, perhaps due to operation error or system implementation flaws. Such low quality data needs to be cleaned prior to data mining.

iv. **Data integration:** includes integrating multiple, heterogeneous data sources.

v. **Data reduction & projection:** includes finding useful features to represent data (depending on goal of task) & using dimensionality reduction or transformation methods.

vi. **Choosing function of data mining:** includes deciding purpose of model derived by data mining algorithm (e.g., summarization, classification, regression, clustering, web mining, image retrieval, discovering association rules & functional dependencies, rule extraction, or a combination of these).

vii. **Choosing data mining algorithm(s):** includes selecting method(s) to be used for searching patterns within data, such as deciding on which model & parameters may be appropriate.

viii. **Data mining:** includes searching for patterns of interest within a particular representational form or a set of such representations.

ix. **Interpretation:** includes interpreting discovered patterns, as well as possible visualization of extracted patterns. One could analyze patterns automatically or semi automatically to identify truly interesting/ useful patterns for user.

x. **Using discovered knowledge:** includes incorporating this knowledge into performance system, taking actions based on knowledge.

3. Soft Computing Technique

Soft computing is an emerging approach to computing which parallels remarkable potential of human mind to reason & learn within an environment of uncertainty & imprecision[12].Soft Computing consists of several computing paradigms like Neural Networks, Fuzzy Logic, & Genetic algorithms. Soft Computing uses hybridization of these techniques. A hybrid technique would inherit all benefits of constituent techniques. Thus elements of Soft Computing are complementary, not competitive, offering their own advantages & techniques to partnerships to allow solutions to otherwise unsolvable problems.

4. Data Mining through Soft Computing Methods

Recently various soft computing methodologies have been applied to handle different challenges posed by data mining. Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, & rough sets) are most widely used within data mining step of overall KDD process. Fuzzy sets provide a natural framework for process within dealing with uncertainty. Neural networks & rough sets are widely used for classification & rule generation. Genetic algorithms (GAs) are involved within various optimization & search processes, like query optimization & template selection. Other approaches like case based reasoning [5] & decision trees [12], [13] are also widely used to solve data mining problems. Each of them contributes a distinct methodology for addressing problems within its domain. This is done within a cooperative, rather than a competitive, manner. The result is a more intelligent & robust system providing a human-interpretable, low cost, approximate solution, as compared to traditional techniques. Let us first describe roles & importance of individual soft computing tools & their hybridizations, followed by various systems developed for handling different functional aspects of data mining. A satisfactory preference criterion is often optimized during mining. It may be stated that there is no universally best data mining method; choosing particular soft computing tool(s) or some combination with traditional methods is entirely dependent on particular application & requires human interaction to decide on suitability of an approach:

a) **Data Mining through Fuzzy Logic:-** As one of principal constituents of soft computing, fuzzy logic is playing a key role within what might be called high MIQ (machine intelligence quotient) systems. Two concepts within fuzzy logic play a central role within its applications. The first is a linguistic variable; that is, a variable whose values are words or sentences within a natural or synthetic language. The other is a fuzzy if-then rule, within which antecedent & consequents are propositions containing linguistic variables [14]. While variables within mathematics usually take numerical values, within fuzzy logic applications, non-numeric linguistic variables are often used to facilitate expression of rules & facts. For example, a simple temperature regulator that uses a fan might look like this:

1. IF temperature IS very cold THEN stop fan
2. IF temperature IS cold THEN turn down fan
3. IF temperature IS normal THEN maintain level
4. IF temperature IS hot THEN speed up fan

There is no "ELSE" – all of rules are evaluated, because temperature might be cold" & "normal" at same time to different degrees.

b) **Data Mining through Neural networks: -** Based on computational simplicity Artificial Neural Network (ANN) based classifier is used. In this proposed system, a feed forward multilayer network is used. Back propagation (BPN) Algorithm issued for training. There must be input layer, at least one hidden layer & output layer. The hidden & output layer nodes adjust weights value depending on error within classification. In BPN signal flow would be within feed forward direction, but error is back propagated & weights are up dated to reduce error. The modification of weights is according to gradient of error curve, which points within direction to local minimum. Thus making it much reliable within prediction as well as classifying tasks.

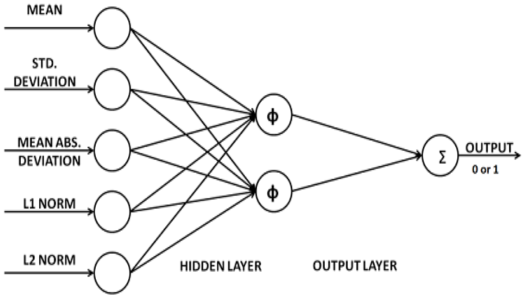

Fig.2. Artificial Neural Network Structure

Unlike fuzzy sets, main contribution of neural nets toward data mining stems from rule extraction & clustering.

I. **Rule Extraction:** In general, primary input to a connectionist rule extraction algorithm is a representation of trained neural network, within terms of its nodes, links & sometimes data set. One or more hidden & output units are used to automatically derive rules, which may later be combined & simplified to arrive at a more comprehensible rule set.

II. **Rule Evaluation:** Here we provide some quantitative measures to evaluate performance of generated rules. This relates to preference criteria/goodness of fit chosen for rules. Let N be a matrix who's element indicates number of patterns actually belonging to class but classified as class:

a. ***Accuracy:*** It is correct classification percentage, provided by rules on a test set defined as where is equal to number of points within class & of these points are correctly classified. User's accuracy: If points are found to be classified into class.

b. ***Kappa:*** The kappa value for class is defined as numerator & denominator of overall kappa are obtained by summing respective numerators & denominators of separately overall classes.

c. ***Fidelity:*** It is measured as percentage of test set for which network & rule base output agree.

d. ***Rule base size****:* This is measured within terms of number of rules. The lower its value, more compact is rule-base.

c) **Data Mining through Neuro-Fuzzy Computing [18]: -** Neuro-fuzzy computation is one of most popular hybridizations techniques. It comprises a judicious unification of merits of neural & fuzzy approaches, enabling one to build more intelligent decision-making systems. This combines generic advantages of artificial neural networks like massive parallelism, robustness, & learning within data-rich environments into system. The modeling of imprecise & qualitative knowledge within natural/linguistic terms as well as transmission of uncertainty is possible through use of fuzzy logic. Besides these generic advantages, neuro-fuzzy approach also provides corresponding application specific merits as highlighted earlier.

The rule generation aspect of neural networks is utilized to extract more natural rules from fuzzy neural networks. The fuzzy MLP [19] & fuzzy Kohonen network [20] have been used for linguistic rule generation & internecine. Here input, besides being within quantitative, linguistic, or set forms, or a combination of these, could also be incomplete. The components of input vector consist of membership values to overlapping partitions of linguistic properties low, medium, & high corresponding to each input feature. Output decision is provided within terms of class membership values. The block diagram of a fuzzy neural network is depicted within Fig. 3.

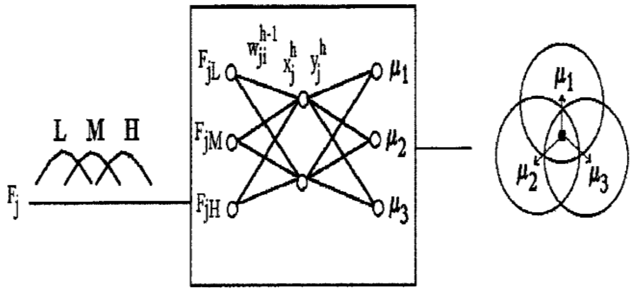

Fig.3. block diagram of a Neuro-Fuzzy network.

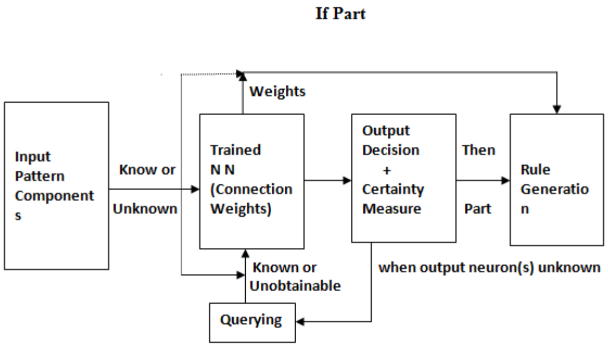

Fig.4. block diagram of inference & rule generation phases.

The models are capable of

a. Inference based on complete and/or partial information;
b. querying user for unknown input variables that are key to reaching a decision;
c. Producing justification for inferences within form of IF THEN rules.

The connection weights & node activation values of trained network are used within process. A certainty factor determines confidence within an output decision. Note that this certainty refers to preference criterion for extracted rules, & is different from notion of certain patterns of (1). Fig. 4 gives an overall view of various stages involved within process of inference & rule generation.

d) Data Mining through Rough Set: - The theory of rough sets has appeared as a major mathematical tool for managing uncertainty that arises from granularity within domain of discourse, i.e., from indiscernibility between objects within a set, & has proved to be useful within a variety of KDD processes. It offers mathematical tools to discover hidden patterns within data & therefore its importance, as far as data mining is concerned, could within no way be overlooked. A fundamental principle of a rough set-based learning system is to discover redundancies & dependencies between given features of a problem to be classified. It approximates a given concept from below & from above, using lower & upper approximations. Some of rough set-based systedms developed for data mining include 1) KDD-R system based on variable precision rough set (VPRS) model; & 2) rule induction system based on learning from examples based on rough set theory (LERS).

Rough set applications to data mining generally proceed along following directions.

a. Decision rule induction from attribute value table [21]–[24]. Most of these methods are based on generation of discernibility matrices & deducts.

b. Data filtration by template generation [25].This mainly involves extracting elementary blocks from data based on equivalence relation. Genetic algorithms are also some-times used within this stage for searching, so that methodologies could be used for large data sets.

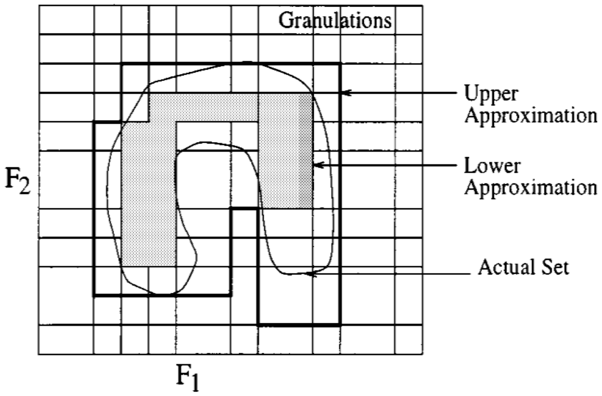

Fig.5. Lower & upper approximations within a rough set.

e) Data Mining through Genetic Algorithms: - search methods, suitable within conditions where search space is large. They optimize a fitness function, corresponding to preference criterion of data mining, to arrive at an optimal solution using certain genetic operators. Knowledge discovery systems have been developed using genetic programming concepts. The MASSON system, where intentional information is extracted for a given set of objects, is popular. The problem addressed is to find common features of a set of objects within an object-oriented database.

Genetic programming is used to automatically generate, evaluate, & select object-oriented queries. GAs are also used for various other purposes like incorporation of multiple data types within multimedia databases, & automated program generation for mining multimedia data.

However, literature within domain of GA-based data mining is not as rich as that of fuzzy sets. We provide below a categorization of few such interesting systems based on functions modeled.

- Regression: Besides discovering human-interpretable patterns data mining also encompasses prediction [8], where some variables or attributes within database are used to determine unknown or future values of other variables of interest. The traditional weighted average or linear multi regression models for prediction require a basic assumption that there is no interaction among attributes. GAs, on other hand, are able to handle attribute interaction within a better manner. Xu et al. [26] have designed a multi-input–single-output system using a nonlinear integral. An adaptive GA is used for learning nonlinear multi regression from a set of training data. Noda et al. [27] use GAs to discover interesting rules within a dependence modeling task, where different rules could predict different goal attributes. Generally attributes with high information gain are good predictors of a class when considered individually. However attributes with low information gain could become more relevant when attribute interactions are taken into account. This phenomenon is associated with rule interestingness. The degree of interestingness of consequent is computed based on relative frequency of value being predicted by it. In other words, rarer value of a goal attribute, more interesting a rule it predicts. The authors attempt to discover a few interesting rules (knowledge nuggets) instead of a large set of accurate (but not necessarily interesting) rules.

- Association Rules: Lopes et al. [28] evolve association rules of IF THEN type, which provide a high degree of accuracy & coverage. While accuracy of a rule measures its degree of confidence, its coverage is interpreted as comprehensive inclusion of all records that satisfy rule. Hence & are defined. Note that quantitative measures for rule evaluation have been discussed within Section III-B2, with reference to neural networks.

f) Data mining through other Hybridizations: - Banerjee et al. have used a rough-neuro-fuzzy integration to design a knowledge-based system, where theory of rough sets is utilized for extracting domain knowledge. In said rough-fuzzy MLP, extracted crude domain knowledge is encoded among connection weights. Rules are generated from a decision table by computing relative reducts. The network topology is automatically determined & dependency factors of these rules are encoded as initial connection weights. The hidden nodes model conjuncts within antecedent part of a rule, while output nodes model disjuncts. A promising direction within mining a huge dataset is to 1) partition it; 2) develop classifiers for each module; & 3) combine results. A modular approach has been pursued to combine knowledge-based rough-fuzzy MLP sub-networks/modules generated for each class, using GAs. An -class classification problem is split into two-class problems.

• 5 Conclusion & Future Scope

Current research within data mining mainly focuses on discovery algorithm & visualization techniques. There is a growing awareness that, within practice, it is easy to discover a huge number of patterns within a database where most of these patterns are actually obvious, redundant, & useless or uninteresting to user. To prevent user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only useful/interesting patterns & present them to user.

Soft computing methodologies, involving fuzzy sets, neural networks, genetic algorithms, rough sets, & their hybridizations, have recently been used to solve data mining problems. They strive to provide approximate solutions at low cost, thereby speeding up process. A categorization has been provided based on different soft computing tools & their hybridizations used, mining function implemented, & preference criterion selected by model.

Web Usage Mining is application of data mining techniques to discover interesting usage patterns from Web data within order to understand & better serve needs of Web-based applications. Usage data captures identity or origin of Web users along with their browsing behavior at a Web site.

Web intelligence is area of study & research of application of artificial intelligence & information technology on web within order to create next generation of products, services & frameworks based on internet.

Recently, several commercial data mining tools have been developed based on soft computing methodologies. These include Data Mining Suite, using fuzzy logic; Braincell, Cognos Thought & IBM Intelligent Miners for Data, using neural networks; & Nuggets, using

GAs. Since databases to be mined are often very large, parallel algorithms are desirable. However, one has to explore a tradeoff between computation, communication, memory usage, synchronization, & use of problem-specific information to select a suitable parallel algorithm for data mining. One could also partition data appropriately & distribute subsets to multiple processors, learning concept descriptions within parallel, & then combining them.

Reference

[1] U. Fayyad & R. Uthurusamy, “Data mining & knowledge discovery within databases,” *Commun. ACM*, vol. 39, pp. 24–27, 1996.

[2] W. H. Inmon, “The data warehouse & data mining,” *Commun. ACM*, vol. 39, pp. 49–50, 1996.

[3] J. A. Major & D. R. Riedinger, “EFD—A hybrid knowledge statistical based system for detection of fraud,” *Int. J. Intell. Syst.*, vol. 7, pp. 687–703, 1992.

[4] R. L. Blum, *Discovery & Representation of Causal Relationships From a Large Time-Oriented Clinical Database: The RX Project*. New York: Spinger-Verlag, 1982, vol. 19. of *Lecture Notes within Medical Informatics*.

[5] R. Heider, Troubleshooting CFM 56-3 Engines for Boeing 737—Using CBR & Data-Mining, Spinger-Verlag, New York, vol. 1168, pp. 512–523, 1996. *Lecture Notes within Computer Science*.

[6] U. Fayyad, D. Haussler, & P. Stolorz, “Mining scientific data,” *Commun. ACM*, vol. 39, pp. 51–57, 1996.

[7] O. Etzioni, “The world-wide web: Quagmire or goldmine?,” *Commun. ACM*, vol. 39, pp. 65–68, 1996.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, Eds., *Advances within Knowledge Discovery & Data Mining*. Menlo Park, CA: AAAI/MIT Press, 1996.

[9] “Special issue on knowledge discovery within data- & knowledge bases,” *Int. J. Intell. Syst.*, vol. 7, 1992.

[10] Aggarwal K , Singh Yogesh, Arvinder Kaur, Malhotra Ruchika (2006) Application of Neural Network for Predicting Maintainability Using Object- Oriented Metrics.Transaction on Engineering, Computing & Technology, Vol. 15.

[11] Aggarwal K, Yogesh Singh (2008) Software Engineering: Program, Documentation & Operating Procedure. New Age International Publishers, 3rd edition.

[12] J. Furnkranz, J. Petrak, & R. Trappl, “Knowledge discovery within international conflict databases,” *Applied Artificial Intelligence*, vol. 11, pp. 91–118, 1997.

.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[14] S.N. Sivanandan, || Principals of Soft Computing||

[15] Calvo R., Partridge M., & Jabri M., —A Comparative Study of Principal Components Analysis Techniques|| .

[16] Terry Quatrani & Grady Booch. Visual Modeling with Rational Rose 2000 & UML. Addison-Wesley, 1999.

[17] U. Fayyad, G. P. Shapiro, & P. Smyth, “The KDD process for extracting useful knowledge from volumes of data,” Commun. ACM, vol. 39, pp. 27–34, 1996.

[18] S. K. Pal & S. Mitra, Neuro-Fuzzy Pattern Recognition: Methods within Soft Computing. New York: Wiley, 1999.

[19] S. Mitra & S. K. Pal, “Fuzzy multi-layer perceptron, inferencing & rule generation,” IEEE Trans. Neural Networks, vol. 6, pp. 51–63, 1995.

[20] , “Fuzzy self organization, inferencing & rule generation,” IEEE Trans. Syst., Man, Cybern. A, vol. 26, pp. 608–620, 1996.

[21] T. Mollestad & A. Skowron, “A rough set framework for data mining of propositional default rules,” within Lecture Notes Comput. Sci., 1996, vol. 1079, pp. 448–457.

[22] X. Hu & N. Cercone, “Mining knowledge rules from databases: A rough set approach,” within Proc. 12th Int. Conf. Data Eng.. Washington, DC, Feb. 1996, pp. 96–105.

[23] A. Skowron, “Extracting laws from decision tables—A rough set approach,” Comput. Intell., vol. 11, pp. 371–388, 1995.

[24] N. Shan & W. Ziarko, “Data-based acquisition & incremental modification of classification rules,” Comput. Intell., vol. 11, pp. 357–370, 1995.

[25] L. Polkowski & A. Skowron, Rough Sets within Knowledge Discovery 1and 2. Heidelberg, Germany: Physica-Verlag, 1998.

[26] K. Xu, Z. Wang, & K. S. Leung, “Using a new type of nonlinear integral for multi regression: An application of evolutionary algorithms within data mining,” Proc. IEEE Int. Conf. Syst., Man, Cybern., pp. 2326–2331, Oct. 1998.

[27] E. Noda, A. A. Freitas, & H. S. Lopes, “Discovering interesting prediction rules with a genetic algorithm,” Proc. IEEE Congr. Evolutionary Comput. CEC ’99, pp. 1322–1329, July 1999.

[28] C. Lopes, M. Pacheco, M. Vellasco, & E. Passos, “Rule evolver: An evolutionary approach for data mining,” within Proc. RSFDGrC’99, Yamaguchi, Japan, Nov. 1999, pp. 458–462.