Survey Paper on Different Methods in Data Mining

Prof.D.G. Vyawahare, Mr. Aniruddha Suresh Wankhade

***Abstract*— In new era the information and data communication technologies are highly used in the Business of Industry. The data warehouse is used in the many business value by improving the effectiveness of managerial decision-making. Basically such a process may open new assumption of dimensions, detect new invasion of patterns, and raises new data with security problems. Rough Sets Theory Based Multimedia Data contain graphics, animations, video, sounds, music, texts etc. This theory represents a promising technique in imperfect data analysis which has found interesting extensions and various applications that handle imperfect knowledge, such as Bayesian inference, fuzzy set, rough set, neural network, decision tree etc.**

***Index Terms*— Neural Networks, Data mining, privacy, databases and security.**

I. INTRODUCTION

This paper contain methods in data mining. Where it is difficult to sort out the different data contain to manipulate it . Different type of data contain distinct of information which is basically very important to sort out with their types for understanding. So data mining types are there for distinguish different types of data like mathematical notation,video,algorithm, fuzzy, neural,inductive logic and many more. Spatial Knowledge Discovery(SDK) is to extract the hidden, implicit, valid, novel and interesting spatial or non spatial patterns, rules, and knowledge from incomplete,large amount, noisy, fuzzy.Data mining is a technique to dig the data from the large database for analysis and executive decision making. In this paper we have present different methods to measure for the data mining.

II. ROUGH SETS

Rough set theory was developed by Zdzislaw Pawlakin the early 1980's. The main goal of the rough set analysis is induction of(learning)approximations of concepts .It introduce mathematical tools to discover patterns hidden in data. It can be used for selection of feature, feature of extraction, data reduction, decision of rule generation, and pattern to extraction (templates, association rules) etc. identifies partial or data total dependencies, redundant data elimination, gives null values approach, missing data, dynamic data and others.

Manuscript received July 22, 2015

Prof. D.G. Vyawahare, completed his ME(CSE) from Government College of Engineering, Aurangabad

Mr. Aniruddha Suresh Wankhade, completed BE(IT) from Anuradha Engineering college, Chikhli. Pursuing ME(CSE) from Anuradha Engineering College, Chikhli.

Rough Sets Theory Based Multimedia Data contain graphics, animations, video, sounds, music, texts etc. Multimedia is defined as combination of more than one media; they may be two types, static and dynamic media explain text, graphics and images are sorted as static media, where objects like animation, music, audio, speech, and video are sorted as dynamic media.

There are many issues and challenges faced by multimedia data providers to fulfill requirements of user. The issues is to arrange and differentiate the huge multimedia data so that the information can be obtained easily at any point of time.

Many studies have been done in multimedia data management. For example, [6] proposed temporal elements such as valid time and transaction time into multimedia data management transactions.

An efficient multimedia data management is highly required because it will improve the process of multimedia information discovery especially for decision application making, marketing of business and intelligent system, etc.

Rough set theory can be used to develop classification model for the particular data sets, where this classification is used to group the data into group which are predefined. Here research explain how rough set theory could be used to predict accuracy of multimedia data whether the data is audio, image or video. Then, web services technology is applied to execute the proposed model under collaborative environment.

III. APPLICATION OF ROUGH SETS

A. Discretization

Discrete symbols are used as new values of the original features. A cut point is a real value c, within the range of a continuous feature, that partitions the [a, b] interval is partitioned into two sub intervals [a, c] and (c, b]. A continuous feature could be partitioned into many sub intervals. A continuous feature with many cut points can make the learning process longer, while a very low number of cut points may affect the predictive accuracy negatively.

Rough sets theory can be applied to compute a measure considering partitioning generated by these cut points and the decisions feature in order to obtain a better of cut points set. We need to set *m* as the number of intervals given by the Scott's formula to determine the bins of a histogram.

Discretization Algorithm:

Input: The original dataset D , and *m* the maximum number of intervals to be considered

For each continuous feature $i\,v$ of Data

For j in 1:m (m is n class. scott ($i\,v$))

Calculate the partition considering j equal width intervals

Evaluate each partition using an association measure based on Rough sets

$\gamma = Pos\ (vi / d\ ) /n$

Stopping criteria: Select the optimal number of partition p

B. *Feature selection*

Feature selection methods determine an appropriate feature subset such that the classification error is optimal. The chosen features permit that pattern vectors belonging to different categories occupy compact and disjoint regions in an *m-dimensional* feature space. Classification and regression algorithms could present problems in their general behavior when redundant features are to be considered. For many investigators the main reason to search for different methods to detect these features.

Feature Selection Algorithm:

Input- Decisional features C, D and Set of conditional
Initialize the best subset of features as the empty set
i. For i in 1:number of conditional features
Have some evaluation measure based on dependency of Rough sets.
End for
ii. Order the features according to dependency measure
iii. Select only the features with high dependency measure.
Output: A subset of features.

C. *Instance selection*

An instance or case is a collection of values taken from an observation considering all the features (conditional and decisional). It is also named a t-uple, sample or data point. Some of the instances in a dataset appear more than once or could be very similar to others, then these could be eliminated since they are redundant. The elimination of similar instances tackle down the redundancy problem. The instance selection problem reduces the training data by searching for the optimal instances and reaching high accuracy of Knowledge Discovery on the unseen data (see Fig. 5.1). Instance selection has the purpose of selecting high quality cases, eliminating noisy data, and inconsistent data. This will produce a reduction of the storage requirement and a speed-up of the computation of posterior Knowledge Discovery and Data Mining tasks [71]. There are various strategies for drawing a representative subset of samples from a dataset. The size of a suitable subset is determined by taking into account the cost of computation, memory requirement, accuracy of the estimator, and other characteristics of the algorithm and dataset. In general, a subset size is determined in such way that the estimates for the entire data set do not differ by more than a stated error margin in more than δ of the samples (Kantardzic, 2003) [40]. It is considered that a KDD task can be executed efficiently using the chosen subset of the original data set.

Instance Selection Algorithm:

Input: The original dataset and the percentage 100p% to be sampled from the positive region. The dataset may contain some continuous conditional feature.
i. Discretize continuous features of the dataset.
ii. Calculate the elementary sets (make partition according to conditional and decisional features).
iii. Calculate the positive region to eliminate the inconsistent cases.
iv. Select 100p% instances from the positive region and save their labels in a list L.
v. Extract specific cases from the original dataset according to list L.
Output: The set of cases to be selected.

III.FUZZY SETS

Fuzzy set consists of elements which have varying degrees of membership.Spatial Knowledge Discovery (SKD) is to extract the hidden, implicit, valid, novel and interesting spatial or non-spatial patterns, rules and knowledge from incomplete, large-amount, ,noisy, fuzzy, random, and practical spatial databases, which include spatial data mining and uncertain reasoning.
In recent years, the term, "spatial data mining and knowledge discovery" (SDMKD) has been connectedly used, in which data mining is a key step or technique in the course of spatial knowledge discovery. With an efficient and rapid improvement of automatic obtaining technologies of spatial data, the amount of data in spatial database have been increased in index movement.
Fuzzy set theory, evidence theory, and neural networks, are powerful computational tools for data analysis and have good potential for data mining as well. But traditional spatial data mining and knowledge discovery did not pay attention to these characteristics.

APPLICATIONS OF FUZZY SETS:

[1]Fuzzy Cluster Analysis
[2]Learning Fuzzy Rule-Based Systems
[3]Fuzzy Decision Tree Induction
[4]Fuzzy Association Analysis
[5]Fuzzy Methods in Case-Based Learning
[6]Possibilistic Networks

IV. NEURAL NETWORK

Neural network more accurately called artificial neural networks, are computational models that consists of a number of simple processing units that communicate by sending signals to one another over a large number of weighted connections. They were originally developed from the inspiration of human brains. In human brains, a biological neuron collect signals from other neurons through a host fine structures called dendrites. The neuron send out spikes of electrical activity through a long ,thin stand called as axon, which slits into thousands of branches .At the end of branch ,a structure called synapse converts the activity from the axon into electrical effects that inhibit or excite activity in which neurons are connected. A neuron receives the excitation input that is sufficiently large compared with its input, it sends a spike of electrical activity. While learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes. Like human brain, neural networks also consist of processing units (artificial neurons) and connections (weights) between them. The processing units transport incoming information on their outgoing connections to other units. The "electrical" information is simulated with specific values stored in those weights that make these networks have the capacity to learn, memorize and create relationship within data.

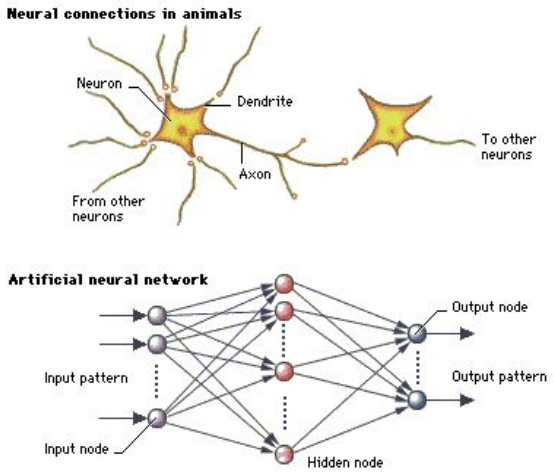

Fig. Neural connection and Artificial neural network.

ANNs have powerful pattern classification and pattern recognition capabilities through learning andgeneralize from experience. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. They are powerful tools to modeling, especially when the underlying data relationship is unknown. After training, ANNs can be used to predict the outcome of new independent input data. Training Of Artificial Neural Networks :A neural network has to be configured such that the application of a set of inputs produces set of outputs required. To set the strengths of the connections exist there are various methods. Set the weights explicitly, using a priori knowledge or try to 'train' the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. There are different learning situations as follows: 1.Supervised learning and 2.• Unsupervised learning 1.Supervised learning : Supervised learning is also called as Associative learning, with input and matching output patterns. 2. Unsupervised learning : Unsupervised learning is also called as Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input.

APPLICATIONS OF NEURAL NETWORKS
• Economic Modeling
• Mortgage Application
Assessments
• Sales lead assessments
• Disease Diagnosis
• Manufacturing Quality Control
• Sports forecasting
• Process Fault detection
• Bond Rating
• Credit Card Fraud
• Detection
• Oil Refinery Production
Forecasting
• Foreign Exchange Analysis
• Market and Customer Behavior
Analysis
• Optimal resource Allocation
• Financial Investment Analysis
• Optical Character Recognition
• Optimzation

V. DECISION TREE

Trees are connected in a acyclic graphs. They are fundamental to computer science (data structures),biology (classification), psychology (decision theory), and many other fields.The decision tree approach is mainly used for generating interesting classification rule.Attribute of decision tree which are required for analysis are taken as starting node.The attribute is first classified in terms of groups and after that important attribute is again taken and classified under certain consideration.

There are three types of Decision tree node:
1.A root node that has no incoming edges and zero or more outgoing edges.
2. Internal node having each of which has exactly one incoming edge and two or more outgoing edges.
3. Leaf or terminal node, each of which has exactly one incoming edge and no outgoing edges.

In decision tree, each leaf node is assigned label. The non-terminal nodes, which include the root and other internal nodes contain attribute test conditions to separate records that have different characteristics.

Application of Data mining (decision trees) :
These techniques are employed to maximize personalized learning and adaptive learning.
Based on decision trees, the PCLS can evaluate learning performance and can personalize the learning path to enhance the learner's creativity.
Moreover, mining data in a way that is based on statistical theories and algorithms can integrate learning information to form potential educational knowledge, which provides critical information to the teacher and the system developer to establish the learning process and the cognitive load of the learners.

VI. INDUCTIVE LOGIC PROGRAMMING

Inductive logic programming is a categorized field of machine learning or data mining which uses logic programming for representation of Example background knowledge and the hypotheses. It will give the known background knowledge, represented a set of example as a logical database of things. ILP system will give hypothesized logic program where it will include all positive and none of negative examples.

Aim- Positive examples + Negative examples + background knowledge => Hypothesis

From observation and background knowledge LIP generate hypothesis and find patterns base on this hypothesis space.
The generated hypothesis will gives the nature of data and can be taken as rough model.

APPLICATION OF INDUCTIVE LOGIC PROGRAMMING
A. LIP have a advantage to fast pattern discovering so it can be use as broad range of application.
B.The expressiveness and illegibility of the first pattern logic hypothesis representation language.
C. The ability to use structured complex and multi-relational data.

D. The compact and natural description of relations.
E. The ability to use different forms of background knowledge

CONCLUSION

From above paper different data mining methods are studied . Rough Sets Theory Based Multimedia Data contain graphi cs, animations, video, sounds, music, texts etc.Fuzzy set con sists of elements which have varying degrees of member-shi p,Neural network more accurately called artificial neural net works, are computational model,Inductive logic programmin g is a categorized field of machine learning,
Decision tree is approach is mainly used for generating inter esting classification rule.

ACKNOWLEDGMENT

It would be my pleasure to express my sincere to my Guide Prof. D.G. Vyawahare in providing helping hand in this paper. His valuable guidence, support and supervision all through this paper title "Survey Paper on Different Methods in Data Mining"

Name- Aniruddha Suresh Wankhade

VII. REFERENCES

[1] Andrew Kusiak, Member, IEEE," Rough Set Theory: A Data Mining Tool for Semiconductor Manufacturing", ieee transactions on electronics packaging manufacturing, vol. 24, no. 1, january 2001
[2] G.Ramani," Rough set with Effective Clustering Method", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2, February 2013
[3] Prachi Patil," Data Mining with Rough Set Using Map-Reduce", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 11, November 2014
[4] Ruying Sun," Data Mining Based on Fuzzy Rough Set Theory and Its Application in the Glass Identification", Modern applied science Volume 3 No.8,August 2009
[5] Fuzzy Rough and Evolutionary Approaches to Instance Selection by Nele Verbiest
[6] M. D. S. Seneviratne and D. N. Ranasinghe," Inductive Logic Programming in an Agent System for Ontological Relation Extraction", International Journal of Machine Learning and Computing, Vol. 1, No. 4, October 2011
[7] Raymond J. Mooney, Prem Melville, Lappoon Rupert Tang," Relational Data Mining with Inductive Logic Programming for Link Discovery", Appears in the Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, Nov. 2002, Baltimore, MD.
[8] Dr. Yashpal singh, alok singh chauhan" neural networks in data mining", Journal of Theoretical and Applied Information Technology© 2005 - 2009 JATIT. All rights reserved.
[9] B.B. Misra and S. Dehuri," Functional Link Artificial Neural Network for Classification Task in Data Mining", Journal of Computer Science 3 (12): 948-955, 2007 ISSN 1549-3636
[10] Chitranjanjit kaur, Pooja kapoor, Meenu Bala," Role of Neural network in data mining", "International Journal for Science and Emerging Technologies with Latest Trends" 2(1): 20-28 (2012)
[11] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari," Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012

Prof. D.G. Vyawahare, completed his ME(CSE) from Government College of Engineering, Aurangabad.Completed BE(CSE) from SSGMC Shegaon, having 10 tear industrial experience and 8 year teaching experience, Currently working as Head of department Anuradha Engineering College, Chikhli
Mr. Aniruddha Suresh Wankhade, completed BE(IT) from Anuradha Engineering college, Chikhli. Pursuing ME(CSE) from Anuradha Engineering College, Chikhli.