IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

A Survey in Data Mining Prospective for handling Uncertainty and Vagueness

Monika Dandotiya[1*], Mahesh Parmar[2]

[1,2]Dept. of Computer Science & Engineering, Madhav Institute of Technology & Science, Gwalior (M.P.), India

DOI: https://doi.org/10.26438/ijcse/v7i4.5661 | Available online at: www.ijcseonline.org

Accepted: 09/Apr/2019, Published: 30/Apr/2019

***Abstract*—** Statistical analysis is used in traditional data mining techniques. But this analysis is less prone to real world scenario. The latest innovations in technology databases contain imprecise & vague data. In the field of data mining, handling such data is always a tedious task. During important decision making task the use of imprecise data causes the inconsistency & vagueness. In this paper to handle uncertain data in data mining various mathematical models like fuzzy set, soft set, rough set & vague set are projected. Various productive approaches have already renewed the Association rule mining. Comparative study of various models defines the idea for using particular set theory. To deal with commercial management & business decision making problem, for generating profitable patterns here we are trying to explore the concept of different set theory. These are also the main benefits of this paper.

Keywords— Data mining, Vagueness, uncertainty, fuzzy set, vague set, Gray set, rough set & association rule mining.

I. INTRODUCTION

Data mining Methods is being exceedingly used for mining the unseen prognostic information from huge databases. Apparent growth of data mining, increase the enormous series of complex applications that is profitable for recovering the yearning information to convert it into information from large databases. The conventional approach in data mining technology offered only statistical analysis in the previous years for mining various problems in large datasets. But in realistic situation due to advancement in the database technology data retrieved in many areas restrained inaccurately precise data. This inaccurate data specifies the existence of uncertainty, incompleteness and vagueness in field of data mining. Such data causes problem during important decision-making task. The problem of decision making, particularly in financial issues is a critical task in business.

Since databases have evolved in many characteristics. Therefore the store data is not always structured for but unstructured too. The notable process to analyze these databases to harmonize some valuable evidence is known as Knowledge Discovery in Databases (KDD). The persistence of KDD process is to mine concealed knowledge which is useful and profitable from large databases. Consider the association rule mining which is important method to locate association among itemset present in database. This method includes basic statistical analysis to find patterns of interest and later generate valid rules that will affect the outcome of its application. As the databases become more advance in storing data of heterogeneous nature, it was evident that statistical techniques will not provide an adequate result. Because of which many reasoning based techniques that could be incorporated with data mining started to surface up.

The large database contains a incredible amount of data about societies, industries & manufacturers, societies etc. which often associated with the uncertainty of various types to cope up with such challenges raised due to uncertain situation knowledge engineering cooperate a vital role. In many areas due to reason such as error in measurement inaccuracy, sampling error and outdated data sources the real data In human life these type of impreciseness could be separated in two categories – uncertainty and vagueness.

In order to mine such complex data in a manner so that it ultimately results in process of important decision making task it is important to recognize the type of uncertainty and then deal with it. Handling uncertain data is biggest challenge in front of computer scientist to deal with such issue some method of soft computing must be integrated which deal to cause with such type of databases. Many mathematical models are proposed as extension of classical set theory like fuzzy set, rough set, soft set, vague set, gray set which are able to deal uncertainty in data mining but same time they have certain superiority over other in dealing particular type of uncertainty. This paper focuses on the various mathematical models that deal with uncertainty and vagueness in data mining.

II. TYPE OF UNCERTAINITY IN DATA MINING

The conventional data mining approach follows numerical and logical implication for finding information in databases. Since the diversified and heterogeneous data are very close to real world they become vulnerable to uncertainty. Through vagueness it can be understand that it cannot be potential to represent the accurate life of data and what will be the conclusion of it when processed. ambiguity occurs when it is impossible to emphasize any value to an object when modeling is done. Uncertainty can be of diverse forms and to recognize specific one is a strenuous work. Various types of uncertaintiy that amalgamated through data are:

a) **Vagueness:** such type of uncertainty arises when modeling of object include the inherent vague values which are not expressed evidently in modeling process.
b) **Ambiguity:** such type of uncertainty occurs when many possible interpretations can made in data due to which object in the model have inflexibility.
c) **Inconsistency:** this is arises when two or more statement in modeling is true at same time which cannot possible.
d) **Imprecision:** such type of uncertainty arises when available information is not explicit to the preferred modeling**.**

To deal with such type of uncertainty in database various soft computing techniques are used to reason with data to some extend and have found there use in research area of clustering, classification, association rule mining in data mining prospective.

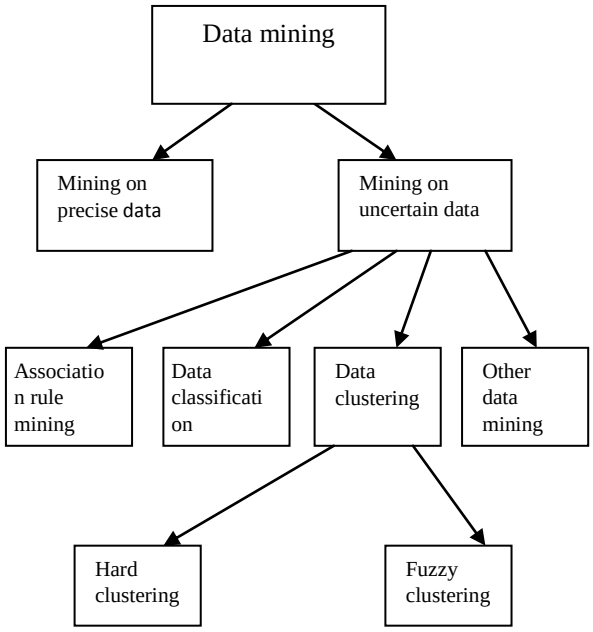

Figure 1: Classification with uncertain data in data mining

2.1 Effect of Uncertainty in mining process

From figure 1 its can understood that effect of uncertain data almost affect every form of mining methods. The effect of uncertainty on association rule mining, clustering and classification is summarized as:

a) **Association Rule mining:** To mine frequent itemset from vague data can be done under probabilistic structure. An itemset is frequent if and only if this itemset's probability is superior than a given probabilistic verge. The expected frequent itemset which is support based frequent itemset measures the uncertainty which is simply extension of the frequent itemset in deterministic data. Traditionally the data model was assumed for association analysis in which the doubtless facts were captured for the transactions for the items that are contained in the transaction but due to wide application in realistic world transactions contain uncertain facts about items in transaction. Such data contains useful information that helps in crucial decision making process and hence cannot ignore form association analysis. Many techniques are implemented to mine uncertain data to find association among the itemset.

b) **Clustering:** In data mining the clustering problem has been well studied that contains five major steps i.e. pattern representations, definition of similarity matric, grouping, data abstraction, output assessment. Clustering over uncertain data are studies in very less amount moreover to handle uncertain data the data clustering techniques need to be modified. The data clustering techniques are classified into two types on uncertain data: hard clustering and fuzzy clustering. The main aim of hard clustering is to improve the accurateness of clustering by allowing predicted data values after data uncertainty is considered. Whereas the main aim of fuzzy clustering is to present the result of data clustering in fuzzy method. Probability density function (pdf) incorporates uncertainty information into existing data mining methods. In order to take uncertainty of data in clustering process, clustering algorithms with objective of decreasing the expected sum of square errors in which data object is specified by an vagueness region with an uncertainty probability density function.

c) **Classification:** In data mining, classification is well studied and generating classification patterns based on uncertain data is still a tedious task. From training dataset rules are being generated to find out the best split attribute and best opening point for this purpose probabilistic information gain is used to discover the best possible split attribute and split point in vague training dataset. For this purpose four techniques are used i.e. uncertain numerical attribute, uncertain categorical data, pruning techniques and prediction techniques. In uncertain numerical attribute an interval is

related with probability density function in which highest value and smallest value are the critical point. There is only one value covered under categorical attribute i.e. called as split point, which is defined by probability distribution over domain and so on.

III. MATHEMATICAL MODELS FOR HANDLING UNCERTAINITY AND VAGUENESS

The real world data is not always exact it contains lot of uncertain and vaguely specified data. Handling such data is biggest challenge in field of data mining. Various mathematical models evolved over time to encounter issues arise due to such data. In the same time these model shows the superiorities over other and their characteristics have practical application in computer science.

3.1. Fuzzy set theory

The most appropriate theory, for dealing with qualms is the theory of fuzzy sets developed by Zadeh. The concept of fuzzy sets provides a well-located means for presenting vague concepts by allowing partial memberships. A fuzzy set (class) A in X is characterized by a membership(characteristic) function $f_A(x)$ which acquaintances with each point in X a real number in the interval [0, 1], with the value of $f_A(x)$ at x representing the "grade of membership" of x in A.

$$f_A(x): U \longrightarrow [0,1] \quad (1)$$

$$X = \{\frac{f_A(x)}{x} \in U, f_A(x) \in [0,1]\} \quad (2)$$

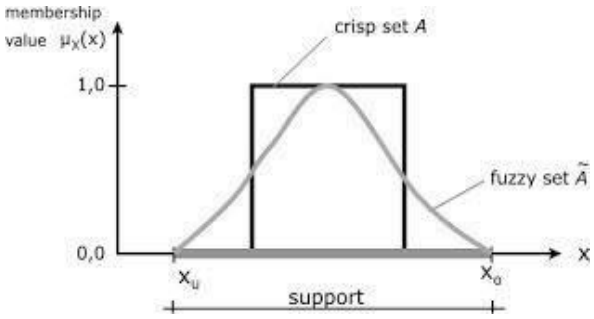

Figure 2: Generalization of fuzzy set over crisp set.

Thus, the quicker the value of $f_A(x)$to unity, the superior the position of membership of x in A. Fuzzy set can be infer by a family of crisp sets, and fuzzy set operators can be definite using usual set operators. The membership values may be interpreted in stipulations of truth values of certain proposition, and fuzzy set operators in terms of logic connectives in many-valued logic. This provides a formulation of fuzzy set theory based on many-valued logic. The fuzzy set theory deals with the ill-definition of the edge of a class through a constant simplification of set characteristic functions.

- **Application domain:** Pattern recognition, communication of information, abstraction.
- **Advantages:** Describes devices in arrangement of numeric and symbolic values, can adopted in to many problems, algorithms can describes with little data so less memory required, robust and not sensitive to change.
- **Disadvantages:** Possibly needs more processing power, takes higher progress time, proof of characteristics is difficult or impossible, insufficiency of parameterization tool, membership function is difficult to found.

3.2. Rough set theory

Rough set theory by Pawlak is suitable mathematical approach for dealing imprecision, incompleteness and uncertainty in data analysis. The rough set attitude is founded on the belief that with every object of the cosmos of conversation we can relate some information (data, knowledge). Objects characterized by the same information are similar in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory. In the rough set approach indiscernibility is definite relative to a given set of practical attribute. Let U be a set called universe and R be an equivalence relation on U, called indiscernibility relation. This pair (U, R) is called an approximation space or Pawlak approximation space. For any ⊆ U, we call the following two subsets the lower and upper approximation with respect to the approximation space (U, R).

$$\underline{R}(X) = \{x \epsilon U | [x]\underline{R} \subseteq U\} \quad (3)$$

$$\overline{R}(X) = \{x \in U | [x]\overline{R} \cap X = \emptyset\} \quad (4)$$

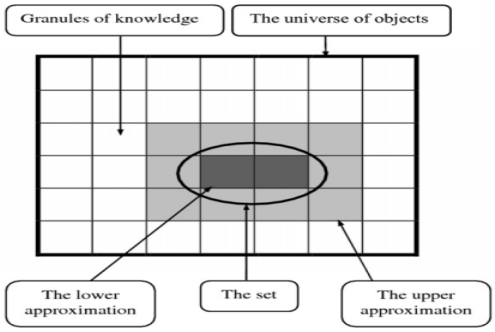

Figure 3: Generalization of rough set

Moreover, if $\underline{R}(X) = \overline{R}(X)$ then X is known a definable set with respect to the approximation space (U, R). Otherwise, X is called rough set in (U, R). Therefore, we suppose that any type of such method is replaced by a pair of specific concepts called the lower approximation consisting of all objects which surely belong to the concept and the upper approximation containing all objects which possibly belong to the concept. The difference between the upper and the

lower estimate constitutes the boundary region of the vague concept. Approximations are two basic operations in rough set theory. Its efficiency has been successfully verified by many related applications such as attribute decline, pattern identification, and so on. The key elements from rough set theory are as follows:

i. There is no supplementary information about the data like thresholds or specialist knowledge is requisite.
ii. For given dataset the discretized attribute values, it is likely to find a subset within the original attributes that are most informative.
iii. Only the facts that concealed in data are analyze.

- **Application:** Classification theory, cluster analysis, dimension theory, field of artificial intelligence, inductive logic, automatic classification, pattern recognition and learning algorithms etc.
- **Advantages:** Easy to understand, no need of additional information about data, suited for concurrent processing, suitable for dataset with inconsistency.
- **Disadvantages:** Algorithms are inefficient for computing the core attribute in large datasets, reliance on absolute information system, does not considers the arithmetic distribution of data in equivalence class.

3.3. Soft set theory

Soft set as a novel mathematical tool for handling with doubts that was free from the insufficiency of the parameterization tools. In the soft set theory the way of defining any object is principally differs from the way in which we use standard mathematics. In it we construct a mathematical model and the perception of the exact solution of this model for an object. This model being too complex in nature, no exact solution is established. Thus, we introduce the estimated solution to that object.

Let U be an initial universe set and a set of parameters. A pair (F, A) is called a soft set over U, where F is a mapping given by

$F: A \rightarrow \mathrm{P(U)}$ such that $\mathrm{F}(\varepsilon) = \emptyset$ if $\varepsilon \notin \mathrm{A}$ (5)

In the soft set theory, the object has an approximate nature since its initialization, and we do not need to introduce the notion of exact solution. We can introduce any parameterization since there is no restriction on the approximate description of the object.

- **Application:** Stability regularization, game theory, soft analysis and operation research.
- **Advantages:** Free from inadequacy of parameterization tool.
- **Disadvantages:** Does not allocate membership standards.

3.4. Vague set theory

A vague set theory V in a universe of discourse U is characterized by a true membership function α_v and a false membership function β_v as follows:

$$\alpha_v : \longrightarrow U[0,1] \quad (6)$$

$$\beta_v : \rightarrow U[0,1] \quad (7)$$

$$\alpha_v(u) + \beta_v(u) \leq 1 \quad (8)$$

Where $\alpha_v(u)$ a lower bound on the grade of membership of u is derived from the evidence for u, and $\beta_v(u)$ is a lower bound on negation of u derived from the evidence against u.

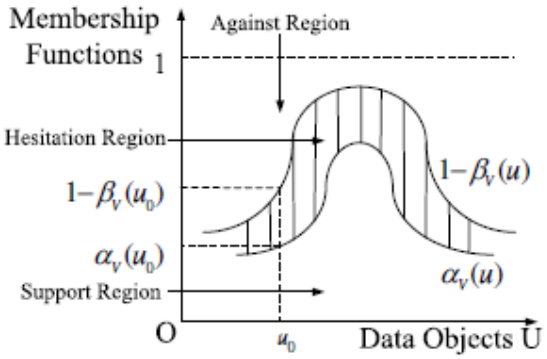

Figure 4: Membership of vague set

The two membership of vague set theory are useful in capturing hesitation information about data which contain valuable information in making strategic planning to incorporate under consideration items. The two membership function of vague set is capable in capturing three type of evidences i.e. support, against and hesitation. The two derived membership i.e. median membership and imprecision membership capture total evidences and imprecision of item in a vague interval.

- **Applications:** Decision making problem in commercial management, E-commerce, education and academia.
- **Advantages:** Two memberships improves drawback of fuzzy set theory that naturally capture hesitation region, easy to work with vague data.
- **Disadvantages:** Cannot parameterize elements appropriately describe a smooth changing of information.

3.5. Grey set

In 1982 Deng Julong proposed Grey set theory that deals with intervals to manage uncertainty. This theory uses small sample and vague conditions for generalizing estimates especially in conclusion making problems. The two membership function is used by this set theory, superior and inferior membership to deal with vagueness. The exact value is not known in the grey number but the series of it is known. The correct crisp value of white number and black number is a number that neither its exact value nor its range is known.

A grey number $\otimes x$ is given from the following

circumstantial situations:

- x : Is element of the universe set U.
- X : The value range that x may hold ($X \subseteq \mathbb{R}$)
- G : Grey set of U.
- $\mu_G{}^*(x): U \rightarrow [0,1]$: The upper membership function.
- $\mu_{G^*}(x): U \rightarrow [0,1]$: The lower membership function.
- x^* : The upper endpoint of x.
- x_* : The lower endpoint of x.

Where $\mu_G{}^*(x) \leq \mu_{G^*}(x)$ and if $\mu_G{}^*(x) = \mu_{G^*}(x)$ then grey set become fuzzy set.

- **Applications:** Can be applied in many real world problems like manufacturing and hydrology.
- **Advantages:** Use two membership and able to find absolute uncertainty as well as relative uncertainty, need small samples, need not to consider the distribution law and trend, easy in modeling and computing.
- **Disadvantages:** Low accuracy, Weak in processing nonlinear information, lack of the ability of self-learning, weighted assignment and its computations are difficult.

IV. COMPARATIVE STUDY OF VARIOUS SET TO HANDLE UNCERTAINTY

Features	Fuzzy Set	Rough Set	Vague Set	Soft Set
Proposed	Fuzzy set theory put forwarded by L.A.Zadeh in 1965	Z. Pawlak proposed rough set theory in 1980's.	Gau and Buehrer proposed this theory.	D. Molodtsov proposed soft set theory in 1999.
Objective	It provides tool for dealing out fuzzy and uncertain information that state some indefinite but erroneous meaning.	Processing uncertain knowledge by mean of analyze and process imprecise.	For decreasing the inconsistency and vagueness, this theory is used.	It is a novel mathematical tool for solving uncertainties that was free from the inadequacy of the parameterization tools.
Application domain	Pattern recognition, communication of information.	Classification theory, cluster analysis..	Decision making problem in commercial management.	Stability regularization, game theory.
Advantages	Describes systems in combination of numeric and symbolic standards, can adopted in to many problems.	Easy to understand, no need of additional information about data.	Two memberships improves drawback of fuzzy set theory that naturally capture hesitation region.	Free from inadequacy of parameterization tool.
Disadvantages	Possibly needs more processing power, takes longer development time.	Algorithms are inefficient for computing the core attribute in large datasets.		

V. BACKGROUND AND RELATED WORK

Fuzzy set theory put forwarded by L.A.Zadeh [3] in 1965 that describes vagueness with accurate mathematical language. It provides influential tool for dealing out fuzzy and vague information that define some indefinite but inaccurate meaning to process information with uncertain boundaries by prepared and formulated mathematical methods.

Polish mathematician Z. Pawlak [4] proposed rough set theory in 1980's. To process uncertain knowledge by mean of examine and practice imprecise, inconsistent and incomplete information effectively is the mathematical theory.

Gau & Buehrer proposed vague set theory in 1993 theory in 1993 [2] for handling uncertain situation which provide two membership functions instead of distinct membership function in theory of fuzzy set. The relation between elements and sets in universe is that of "belonging in definite scope" grade of membership is uttered in interval form also degree for evidences and against evidences are expressed in period form.

Grey systems were projected by Professor Julong Deng [5]. In grey systems, there are three categories of information: The complete information with white, the insufficient information with grey, The totally unknown information with black.

In 1999 D. Molodtsov [15] projected soft set theory that proposed a universal arithmetical tool in dealing with fuzzy, unconsistent and not clear object.

Yanhang Li and Zheng Qin [22] made a comparative analysis of similarity measure between vague set and intuitionistic fuzzy set that provide benefit of selection and application for vague and intuitionistic fuzzy sets. An Lu. And Wilfred Ng gave comparison between both and suggests which one is better in handling vague data.

Eyke Hullermeier [19] in 2008 provide the fuzzy sets's application in data mining & machine learning his research

is anxious with technique for the programmed induction of models and extraction of interesting patterns from experiential data.

Yongxin Tong et al. perform mining of frequent itemset over uncertain databases. They identified that uncertain databases are divided into two groups first one referred to expected support based and second one referred the probabilistic frequent itemset. They perform extensive experiment.

REFERENCES

[1] AgrawalR., Imielinski T., Swami A.N."Mining association rules between sets of items in large databases". In Buneman, P., Jajodia, S., eds.: SIGMOD Conference, ACM Press (1993) 207–216.

[2] Gau W.L., Buehrer, D.J. "Vague sets". IEEE Transactions on Systems, Man, and Cybernetics 23 (1993) Pages-610–614.

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

[4] Pawlak, Zdzislaw. "Rough sets." International Journal of Computer & Information sciences 11.5(1982): 341-356.

[5] J. Deng. "The control problems of grey systems" Systems & Control Letters, 1982.

[6] An Lu and Wilfred Ng "Maintaining consistency of vague databases using data dependencies "Data and Knowledge Engineering, Volume 68,2009,Pages 622-641.

[7] Lu A., Ng W. "Managing merged data by vague functional dependencies". In: Atzeni P., Chu W., Lu H., Zhou S., Ling T.-W. (eds.) ER 2004. LNCS, vol. 3288, pp. 259–272. Springer, Heidelberg.

[8] An Lu and Wilfred Ng "Mining Hesitation Information by Vague Association Rules" Lecture Notes in Computer Science ,Springer Volume 4801/,2008,pg 39-55.

[9] Lu A.,Ng W. "Vague sets or intuitionistic fuzzy sets for handling vague data": Which one is better? In: Delcambre L.M.L., Kop C., Mayr H.C., Mylopoulos J., Pastor, O. (eds.) ER 2005. LNCS, vol. 3716, pp. 401–416. Springer, Heidelberg

[10] Lu. A., Ng. W: Handling Inconsistency of vague relations with functional dependencies. In: ER (2007). LNCS Vol 4321 pp 301-312.

[11] Lu,A., Ke,Y., Cheng ,J., Ng,W.: Mining Vague association rules.In:DASFAA,pp.891-897(2007)

[12] Atanassov, K.T.: Intuitionistic Fuzzy Sets: Theory and Applications (Studies in Fuzziness and Soft Computing). Springer-VerlagTelos (1999).

[13] Pardasani K.R., Anajan Pandey "A Model for Vague association rule Mining in Temporal Database" in Journal of Information and Computing Science, Vol.8, 2013, ISSN 1746-7659, pp. 063-074.

[14] Pardasani K.R., Anajan Pandey " A Model for Mining Course Information Using Vague Association Rule " in International Journal of Computer Applications , Vol 58, ISSN 0975-8887, November 2012.

[15] D.Molodtsov "Soft Set Theory-First Result" An International Journal Computers & Mathematics with Application, Elsevier, Vol. 37,pp. 19-31,1999.

[16] Vivek Badhe, Arvind T.S. " Comparative Analysis of Fuzzy, Rough, Vague and Soft set Theories in Association Rule Mining" in International Journal of Scientific Progress and Research (IJSPR) , Vol-2, ISSN 2349-4689, November 2014.

[17] A. Tiwari, R.K. Gupta and D.P. Agrawal "A survey on Frequent Pattern Mining: Current Status and Challenging issues" Information Technology Journal 9(7) 1278-1293, 2010.

[18] W. Wang, J. Yang and P. Yu "Efficient mining of weighted association rules (WAR)", Proc. Of the ACM SIGKDD Conf. on knowledge Discovery and Data Mining. 270-274,2000.

[19] Eyke Hullermeier "Fuzzy sets in machine learning & data mining" Applied Soft Computing, Science Direct, Elsevier, Vol. 11, pp. 1493-1505, 2008.

[20] Feng Tao, "Mining Binary Relationships from transaction data in weighted Setting" PhD Thesis, School of Computer science, Queen's University Belfast, UK, 2003.

[21] G.D. Ramkumar, Sanjay Ranka, and Shalom Tsur, "Weighted Association Rules: Model and Algorithm" KDD1998, 1998.

[22] N.Pasquier, Y.Bastide, R. Taouil, and L.Lakhal, "Efficient mining of association rules using closed itemset lattices," Information Systems, Vol 24, No. 1, 1999, pp. 25-46.

[23] Bing Liu, Wynne Hsu, and Yiming Ma, "Mining Association Rules with Multiple Supports", Proc. Of the ACM SIGKDD Int'1 Conf. On Knowledge Discovery and Data Mining (KDD-99), SanDiego, CA, USA, 1999.

[24] Jiawei Han and Yongjian Fu. "Discovery of Multiple-Level Association Rules from Large Databases" in the Proceedings of the 1995 Int'1 Conf. on Very Large Data Bases (VLDB'95), Zurich, Switzerland, 2002, pp. 420-431.

[25] F. Tao, F. Murtagh and M.Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework, "Proc, ACM SIGMOD '03, pp. 661-666, 2003.

[26] Yanhong Li, David L.Olson, Zheng Qin "Similarity measures between intuitionistic fuzzy (vague) sets: A comparative analysis" Elsevier, Pattern Recognition Letters 28 (2007) pp. 278-285.

AUTHORS PROFILE

Monika Dandotiya, M.tech (CSE) from MITS Gwalior Area of Interest : Data Mining, Image Processing
E-mail: dandotiyamonika@gmail.com
I am pursuing M,tech in CSE&IT Department in MITS Gwalior. I have received B.E. degree from VITM Gwalior. My area of current research includes Data Mining & Image Processing.

Prof. MAHESH PARMAR, Assistant Professor B.E.(CSE), ME (Computer Engineering) Area Of Interest: Data Mining, Image Processing. E-Mail: maheshparmar@mitsgwalior,in
Mr. Mahesh Parmar as an Assistant Professor in CSE&IT Department in MITS Gwalior and having 10 years of Academic and Professional experience. He received M.E. degree in Computer Engineering from SGSITS Indore. He has guided several students at Master and Under Graduate level. His areas of current research include Data mining and Image Processing. He has published more than 25 research papers in the journals and conferences of international repute. He has also published 02 book chapters. He is having the memberships of various Academic / Scientific societies including IETE, CSI, and IET etc.

