Mining Frequent Patterns using Self-Organizing Map with MATLAB Neural Network Toolbox

Peddi Kishor, Dr. Sammulal Porika

Abstract - **Frequent pattern extraction is a basic advance in information digging for association rule mining. The vast majority of the proposed methods for extricating frequent patterns depend on the descending closure lemma idea using the support and certainty system. In this paper we examine a new strategy for mining frequent patterns in a transnational database. Self Organizing Map (SOM) is an unsupervised neural system that successfully makes spatially sorted out inside portrayals of the highlights and reflections recognized in the info space. It is a standout amongst the most prominent clustering procedures, and it uncovers existing similarities in the info space by playing out a topology-protecting mapping. These promising properties show that such a clustering method can be utilized to distinguish frequent patterns in a best down way instead of the customary approach that utilizes a bottom up cross section look. All through our tests shown that how a SOM clustering methodology can be utilized for finding frequent patterns.**

Index Terms - **Frequent Patterns, Data mining, Self Organizing Maps, Clustering, Neurons, SOM Algorithm**

I. Introduction

Data mining is the assignment to mining the valuable significant data from data distribution center. It is the wellspring of in-explicit, absolutely legitimate, and possibly valuable and imperative learning from vast volumes of normal data. Data mining is the way toward separating verifiable, beforehand obscure and conceivably valuable data from substantial amounts of data. Through the gradual addition of current data with authentic data, ventures end up possessing bigger data sets in electronic shape than whenever up to this time. Different procedures have been utilized to change over the data into data, including clustering, classification, association rules, and sequential mining, et cetera.

Classification is utilized to make predictions. Association and sequential mining are utilized to describe conduct. Clustering can be utilized for either forecasting or explanation. One of the primary sorts of prescient displaying assignments is arrangement. In characterization based managed learning, data are mapped into classes, which are re-imagined before the data are inspected. The examination on data mining for CRM demonstrated that data mining strategies used to inspire undiscovered helpful learning from a bigger client data. Data mining has the essential objectives of predicting, describing, and building knowledge or information. CRM makes communication of clients with the association by utilizing data innovation. In addition, distinguishing client's need/intrigue better and treating them as needs be can expand their lifetimes. Client division is the gathering of clients into various gatherings in view of their regular traits. It is the fundamental piece of CRM The key component is a course including Self-Organizing Map (SOM) neural system to separate clients into homogeneous gatherings of clients and a choice tree streamlined strategy to recognize important information. Recognizing shoppers by this approach is useful standard for clients and encourages showcasing methodology advancement. When SOM was utilized to recognize the productive gatherings of clients, the factual compressed data and the choice tree inducer were utilized to portray the gatherings of clients. A tree rearranged system was utilized for finding pertinent grouping rules. Client clustering is the most critical data mining procedure utilized as a part of client relationship administration. Client clustering utilizes client buy exchange data to track purchasing conduct and make vital business activities. As an unsupervised data mining strategy, Self organization map clustering is a decent apparatus for exploratory examination, just like the situation when no from the earlier classes have been recognized. The SOM is an extremely visual tool and has solid capacities for managing non-direct connections, missing data and skewed appropriations.

Clustering methods are likewise alluded to as unsupervised learning. Unsupervised learning is a procedure of arrangement with an unclear focus on, that is, the class of each case is unclear. The point is to section the cases into disjoint classes that are homogenous as for the sources of info. Clustering is noticeable amongst the most valuable tasks in data mining process for finding gatherings and distinguishing intriguing disseminations and patterns in the fundamental data. Clustering issue is tied in with parceling a given data set into clusters (bunches) to such an extent that the data focuses in a group are more like each other than focuses in various groups.

II. Literature Study

Review Different journals and articles concerning frequent pattern mining calculations were considered before. Some looked at frequent pattern mining calculations while some adjusted the current calculations to enhance the execution. Peddi Kishor et al [2] proposed a novel computation to find joined positive and negative association rules. Exploratory

Manuscript received December, 2017.
***Peddi Kishor**, Research Scholar, CSE Department, R&D Cell, JNTUH Hyderabad, (e-mail: kish135@gmail.com). Hyderabad, Telangana , India, 9676445121*
***Sammulal Porika**, CSE Department, JNTUH College of Engineering, Kondagattu, Jagitial, Telangana, India, 9441118893, (e-mail:sammulalporika@gmail.com)*

outcomes demonstrated that their technique got proficient and compelling outcomes contrasted with different methodologies. Anis Suhailis Abdul Kadir et al [3] gave the preliminaries of essential ideas of negative association rule and proposed an improvement in Apriori calculation for mining negative association rule from frequent nonappearance and nearness itemset. Relative intriguing quality measures were received to demonstrate that the produced rules are additionally fascinating and solid.

Guimei Liu et al [4] introduced distinctive techniques to manage the false positive blunders in association rule mining. Three various testing remedy approaches-the immediate change approach, the holdout approach and the stage based approach are utilized and broad investigations have been directed to examine their exhibitions. From the outcomes acquired, all the three methodologies control false positives viably yet among the three permutations– based approach has the most noteworthy energy of identifying genuine association rules, yet it is computationally costly. Huaifeng Zhang et al [5] proposed a calculation to find joined association rules. Contrasted and the current association rule, this joined association rule system enables distinctive clients to perform activities straightforwardly. In their investigation, they have concentrated on rule age and intriguing quality measures in consolidated association rule mining. In joined association rule age, the frequent itemsets among itemset bunches are found to enhance effectiveness. Distinctive creators have looked at the exhibitions of various association rule mining calculations by actualizing them on different sorts of datasets.

Jesmin Nahar et al [6] analyzed the different association rule calculations on coronary illness data anticipating sound and wiped out heart status. The three association calculations utilized were Apriori, Prescient apriori and tertius calculation. In light of the test comes about they presumed that Apriori calculation is the most appropriate calculation for this kind of undertaking. A comparative work was finished by Jyoti Arora et al [8] who played out a correlation of different association rule mining calculations on General store data and got the outcomes utilizing Weka data mining instrument. The calculations analyzed incorporate Apriori association rule, Fp-development and Tertius association rule. Subsequent to looking at execution time by these three calculations, creator finds that FP-growth is quicker than other two calculations. Different creators have likewise attempted to join the association rule mining system with either clustering or classification or both. Sunita B. Aher and Lobo L.M.R.J [9] consolidated the clustering (K-means calculation), classification (ADTree classification) and association rule (Apriori) for course recommender framework in E-learning and contrasted the outcomes and utilizing just association rule. The author finds that the joined approach is superior to just Apriori as there is no compelling reason to preprocess the data.

III. Problem Statement

Mining frequent itemsets is a principal and fundamental advance in the finding of association rules. Given T_n a denote to signify a value-based database comprising of n transactions. Each value-based record R from T_n is comprised of a subset of things from the permitted list of items I = {1,2,3 … m}. An itemset or a pattern alludes to any subset of the things contained in vocabulary I. The support of a pattern X is the quantity of value-based records from T_n which contain X. A pattern is viewed as frequent if its support is higher or equivalent to the pre-indicated minimum support (σ). Given a vast value-based database and a minimum support edge (σ), the undertaking of frequent pattern finding is to find patterns that happen in any event σ transactions of the database.

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule:\ X \Rightarrow Y \longrightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

IV. Self Organizing Maps

Self organizing Map (SOM) is used for visualization and investigation of high-dimensional datasets. SOM assist management of high dimensional datasets into lower dimensional datasets, usually 1-D, 2-D & 3-D. This is an unsupervised learning algorithm, does not have need of a target vector because it learns to classify data with no supervision. A SOM is created from a grid of nodes or units to which the input data are submitted. Each node is connected to the input layer and there is no connection among the nodes. SOM is a unsupervised neural network which uses topology preserving procedure and keeps the neighborhood relations in its mapping arrangement.

The frequent pattern discovery using the traditional approach is parameterized by the support and confidence framework. However, if we are to use a clustering approach like SOM for detecting frequent patterns, support threshold would not be an appropriate measure to use. Rather the method using SOM would be parameterized by the chosen map dimension which will play a role in the frequency of the discovered patterns. As already mentioned reduction of map size increases the competition between patterns to be projected onto the output space, and hence patterns with low frequency will disappear from the map. On the other hand, the larger the map the more patterns will be discovered. As a property of SOM is to compress the input space and present the abstraction in a topology preserving manner, a heuristic to determine the map size will be based on the number of input attributes, the size of the data set and the frequency that the user is after.

The Self Organizing Map is a data-investigation technique that pictures closeness relations in an arrangement of data items. For example in economy, it has been connected to the examination of ventures at various levels of reflection, to evaluate their relative budgetary conditions, and to profile their items and clients. Then again, in industry, the observing of procedures, frameworks and apparatus by the SOM strategy has been an essential application, and there the object is to portray the majority of various info states by requested groups of regular states. In science and innovation everywhere, there exist boundless errands where the examination objects must be grouped on the premise of their natural properties, to say the classification of proteins, hereditary arrangements and cosmic systems.

It is a special type of neural network, which works based on unsupervised and competitive learning. Figure 1 shows the schematic diagram of a self-organizing map. The SOM can be viewed as a non-linear generalization of principal component analysis, which can preserve the topology in mapping process. There are two layers in SOM, namely input and competition layers. The multivariate data are fed to the input layer, which are to be mapped to a lower dimensional space. The number of neurons in the competition layer is kept equal to that of data points present in the input layer. Each multivariate data present in the input layer is connected to all the neurons in the competition layer through some synaptic weights. The neurons in the competition layer undergo three basic operations, such as competition, cooperation, and updating, in stages. In the competition stage, the neurons in the competition layer compete among themselves to be the winner. In the next stage, a neighborhood surrounding the winning neuron is identified for cooperation among themselves. The winning neuron along with its neighbors is updated in the third stage. The above three stages are explained below in detail.

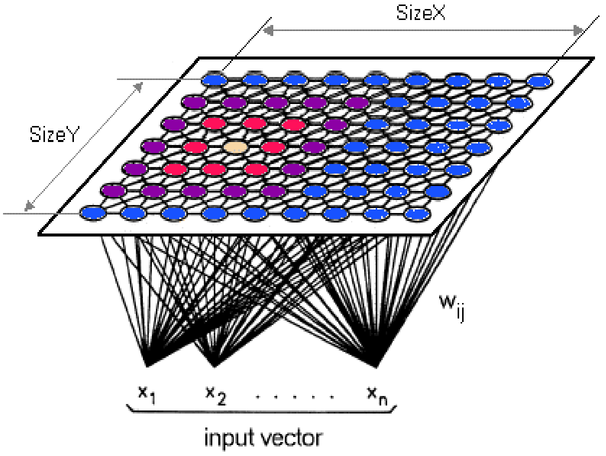
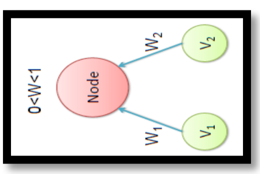


Figure 1: Structure of SOM

In the above figure, each node is connected to the input the same way; no nodes are connected to other nodes. SOMs consist of nodes or neurons. Each neuron has weight vectors attached to them. The dimensionality of the weight vector matches that of the data that the SOM will be trained on. If $X_1, X_2, \ldots\ldots, X_n$, where each "X" is an element of a vector from the training data set, then the corresponding weight vector would be $W_1, W_2, \ldots\ldots, W_n$. Like all neural networks, SOMs need to be trained. The goal is to make them act a certan way, given a particular input. This is achieved by adjusting each nodes weight vectors to resemble those in the training data set.

SOM ALGORITHM:

Step_1: Initialization of each node weights with a random number between 0 and 1.



Step_2: Choose a random input vector from training dataset

Step_3: Calculate the Best Matching Unit (BMU). Each node is examined to discover the vector; its weights are most similar to the input vector. This node is called as the Best Matching Unit (BMU) because its vector is most similar to the input vector. This selection is done by using Euclidean distance principle, which determines similarity between two datasets. The distance between the input vector and the weights of node is calculated in order to find the BMU.

$$Dist = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2}$$

V = the current input vector

W= the node's weight vector

Step_4: Calculate the size of the neighborhood in the region of the BMU. The size of the neighborhood in the region of the BMU is decreasing with an exponential decay function. It shrinks for every iteration until it reaches the BMU.

$$\sigma(t) = \sigma_0 \exp(-\frac{t}{\lambda})$$

σ_0 = *the width of lattice at time zero*

t = *the current time step*

λ = *the time constant*

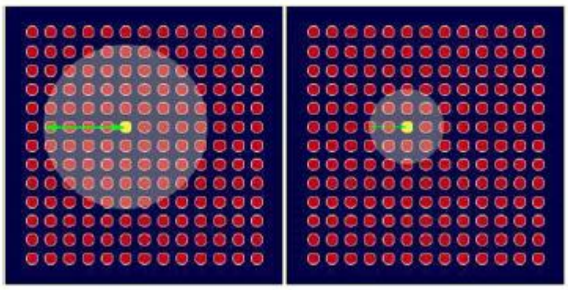


Figure 2: Size of the neighborhood around the BMU

Step_5: Updating of node weights of the BMU and neighboring nodes so that their weight gets more similar to the weight of input vector. The weight of each node inside the neighborhood is adjusted, having greater change for neighbors nearer to the BMU.

$$W(t+1) = W(t) + \Theta(t)L(t)(V(t) - W(t))$$

t = *time-step*

L = *learning rate (decreases with time)*

The learning decay rate is calculated for each iteration.

$$L(t) = L_0 \exp(-\frac{t}{\lambda})$$

While training goes on, the neighborhood regularly shrinks. After completion of training, the neighborhoods have shrunk to zero size.

$$\Theta(t) = \exp\left(-\frac{dist^2}{2\sigma^2(t)}\right)$$

$\Theta(t) =$ *Influence rate*

$\sigma(t) =$ *width of the lattice at time t*

The influence time shows, the amount of influence a node's distance from BMU has on its learning. In the simplest form influence rate is equal to 1 for the entire nodes close up to the BMU and zero for others. After all, from a random allocation of weights and through much iteration, SOM is able to turn up at a map of stable zones. At the end, analysis of data is to be done by individual but SOM is a great technique to retrieve the invisible patterns in the given data. Learning rate is decreased over time. This is to make sure that previous learned inputs aren't discarded. The algorithm stops when a predefined number of iterations have been completed or the average change of weights per iteration drops below a certain predefined value (this value is from now on referred to as delta). As can be seen, no predefined target vectors are set, so the algorithm requires no other input aside from a set of training data.

V. Results & Discussion

In this section, we display some of our preparatory test comes about that approve the accuracy of the proposed system and demonstrate some proper methods for utilizing SOM for frequent pattern extraction. The point of the initial part is to demonstrate how a total arrangement of frequent patterns can be extricated utilizing SOM. As the initial segment demonstrates that the underlying patterns identified by SOM comprise of patterns and perhaps different patterns with high repetition, partially two we utilize a straightforward illustrative case to indicate how confining the map size can create an indistinguishable patterns by utilizing the minimum support structure. At long last to a limited extent three we utilize a more unpredictable syntactic data set to think about the patterns got utilizing SOM. Fluctuating support edges were utilized and the span of the map was balanced in like manner.

Other learning parameters likewise require change with the goal that it turns out to be exceptionally competitive in that restricted space. As a rule we need to change every one of the SOMs parameters in such way with the goal that the self-association and projection that happens in the map is for the most part influenced by the frequent patterns. Just if a specific pattern in the database happens regularly enough at that point the map ought to be influenced. All through our experimentation we have encountered that specific parameters may make the map focalize in less sum time and comparable outcomes could be gotten through the change in various learning parameters. For instance comparable outcomes were acquired via preparing the map for longer time utilizing a littler learning rate. As adequate outcomes were gotten simply after 150 ages, these shows there exist diverse mixes of learning parameters esteems that will create similar outcomes, and consequently for very perplexing data close gauges could be acquired after short preparing. Because of the learning parameters of SOM being subject to each other, the correct rules for copying support can't be given at this beginning period of research into frequent pattern disclosure utilizing SOM. Here we have given some underlying rules to utilizing SOM for frequent pattern extraction and the investigations showed that SOM is in reality ready to separate frequent patterns. There will be more examination required into finding the correct parameters with the goal that the aggregate arrangement of maximal patterns would be ensured.

MATLAB supports unsupervised learning with self-organizing maps and competitive layers. With this tool one can do configuration, prepare, picture, and reenact neural systems. You can utilize Neural System Tool compartment for applications, for example, data fitting, pattern acknowledgment, clustering, time-arrangement expectation, and dynamic framework demonstrating and control. To accelerate preparing and handle vast data sets, you can disperse calculations and data crosswise over multi-center processors, GPUs, and PC bunches utilizing Parallel Figuring Tool. A self-organizing map (newsom) comprises of a competitive layer which can group a dataset of vectors with any number of measurements into the same number of classes as the layer has neurons. The neurons are organized in a 2D topology, which enables the layer to shape a portrayal of the conveyance and a two-dimensional estimation of the topology of the dataset.

iris_dataset

Here is a simple example of the usage of the Toolbox to make and visualize a SOM of a data set. This data set consists of four measurements from 150 Iris flowers: 50 Iris-setosa, 50 Iris-versicolor and 50 Iris-virginica. The measurements are length and width of sepal and petal leaves. The data set is loaded into Matlab and normalized. This information would provide an initial idea of what the data is about, and would indicate how the variables should be preprocessed. The data set consists of 150 samples. Categorize iris flowers based on four attributes. "irisInputs" is an 4x150 matrix, whose rows are:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

"irisTargets" is an 3x150 matrix, where each i^{th} column represents which category the i^{th} iris belongs to with a 1 in one element (and zeros in the other two elements).
1. Iris Setosa 2. Iris Versicolour 3. Iris Virginica

Training automatically stops when the full number (200) of epochs have occurred. Below figures shows experimental results of Iris dataset which are obtained from MATLAB. Here SOM network size is 6X6. The network is trained with the SOM batch algorithm (trainubwb, learnsomb).

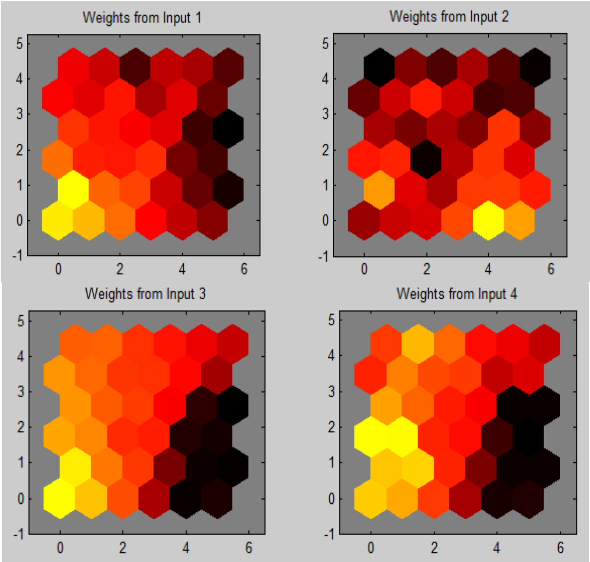


Figure 3: SOM Weight Planes for Iris Dataset

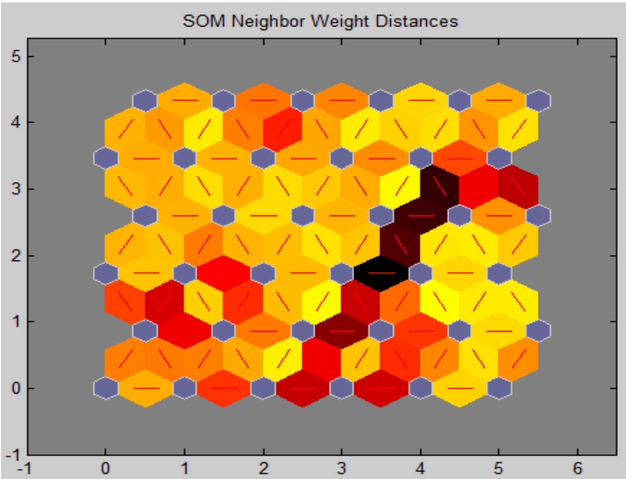


Figure 4: SOM Neighbor Weight Distances for Iris Dataset

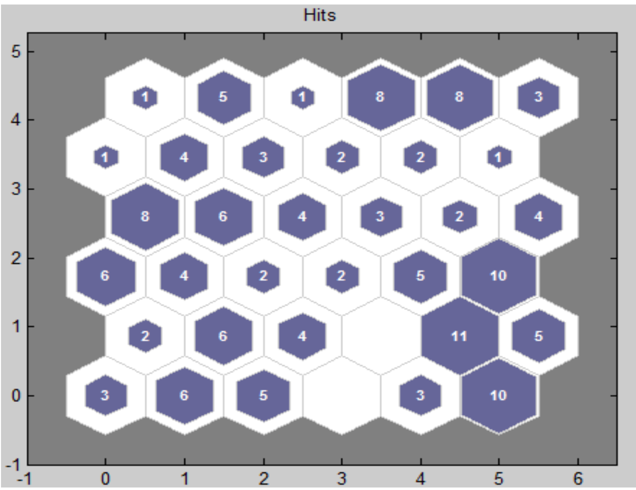


Figure 5: SOM Sample Hits for Iris Dataset

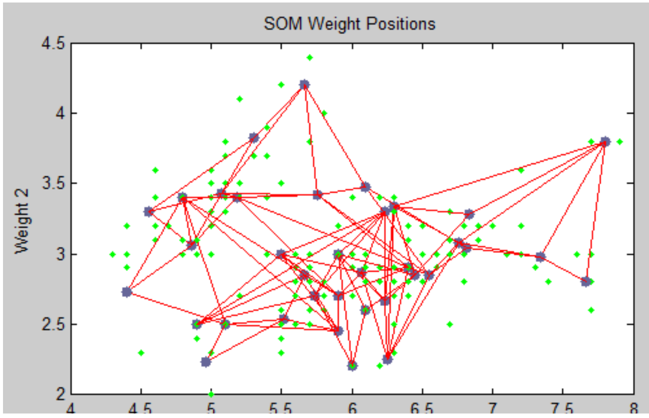


Figure 6: SOM Weight Positions for Iris Dataset

VI. CONCLUSION

This examination has researched an option technique for frequent pattern extraction by utilizing the Self-Organizing Map. All through the analyses we have exhibited how a clustering strategy can be utilized to naturally separate through the infrequent patterns. Frequent patterns are recognized in a best down way, rather than customary approach that utilizes a bottom up cross section search. Besides the SOM approach is parameterized by the measure of the yield measurement instead of the support edge in the conventional approach. Some prompt difficulties that happen while impersonating the support limit by confining the yield estimate were demonstrated. Many components become an integral factor when the support limit is expanded and a more intensive examination on the alteration of SOM"s learning parameters regarding support edge is left as future work. At this stage, we are very happy with the legitimacy and accuracy of the proposed approach when connected to basic data sets. It demonstrates that SOM has some interesting and promising properties for the issue of frequent pattern mining. It is stimulating to investigate more on the productivity of the strategy in contrast with existing Apriori based calculations when connected to demonstrable data with higher multi dimensional quality. Altering the learning parameters of SOM in certain way can give an adequate arrangement in significantly less preparing time. In extremely complex data sets where the customary approach utilizing the support system because of vast measure of hopefuls that should be produced and tried, a clustering way to deal with frequent pattern extraction may should be adjusted where the support limit will be supplanted by the yield space and power of competition boundaries. Other than the way that few patterns could be missed, the strategy would give a decent estimate in situations where the conventional bottom up approaches flop because of natural intricacy. In future work, our exploration will focus on the quantitative examination of SOM mappings, particularly investigation of large datasets, clusters and their properties.

REFERENCES

[1] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD, Washington DC, USA, (pp 207-216).

[2] Peddi Kishor, Dr Sammulal Porika, "An Efficient Approach for Mining Positive and Negative Association Rules from Large Transactional Databases", IEEE International Conference on Inventive Computational Technologies (ICICT), 2016.

[3] Anis Suhailis Abdul Kadir, Azuraliza Abu Bakar and Abdul Razak Hamdan, "Frequent Absence and Presence Itemset for Negative Association Rule Mining ", IEEE,2011.
[4] Guimei Liu, Haojun Zhang and Limsoon Wong, "Controlling False Positives Iin Association Rule Mining" In Proceedings of the VLDB Endowment ACM,2011.
[5] Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang, "Combined Association Rule Mining", PAKDD 2008, LNAI 5012, pp. 1069-1074, 2008 © Springer- Verlag Berlin Heidelberg 2008
[6] Jesmin Nahar,Kevin S.Tickle,Shawkat Ali and YI-Ping Phoebe Chen, "Diagnosis Heart Disease using an Association Rule Discovery Approach" In Proceedings of the IASTED International Conference Computational Intelligence August 2009.
[7] Ila Chandrakar and A. Mari Kirthima, "A Survey On Association Rule Mining Algorithms", In International Journal Of Mathematics and Computer Research, ISSN: 2320-7167, Vol 1, Issue 10, Page No. 270-272, November 2013.
[8] Jyoti Arora, Sanjeev Rao and Shelza, "An Efficient ARM Technique for Information Retrieval In Data Mining " In International Journal of Engineering Research and Technology Vol 2,Issue 10, October 2013.
[9] Lobo L.M.R.J and Sunita B.Aher, "Combination of Clustering, Classification and Association Rule based Approach for Course Recommender System in E-learning ", In International Journal of Computer Applications, Vol 39, February 2012.
[10] Burdick, D., Calimlim, M., & Gehrke, J. (2001). MAFIA: A maximal frequent itemset algorithm for transactional databases. In Intl. Conf. on Data Engineering, April.
[11] Deboeck, G., & Kohonen, T. (1998). Visual explorations in finance using self-organizing maps. London, Springer-Verlag.
[12] Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In ACM SIGMOD Conf. Management of Data, May, Dallas, Texas.
[13] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, September.
[14] Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999) Discovering frequent closed itemsets for association rules. In Proc. 7th Int. Conf. Database Theory (ICDT'99) (pp. 398-416). Jerusalem, Israel, January.
[15] Pei, J., Han, J., & Mao, R. (2000). Closet: An efficient algorithm for mining frequent closed itemsets. In SIGMOD Int'l Workshop on Data Mining and Knowledge Discovery, May.
[16] Tan, H., Dillon, T.S., Feng, L., Chang, E., & Hadzic, F. (2005). X3-Miner: Mining patterns from XML database. Data Mining 2005, Skiathos, Greece.
[17] Wang, J., Han, J., & Pei, J. (2003). CLOSET+: Searching for the best strategies for mining frequent closed itemsets. In Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03), Washington, D.C., August.
[18] Peddi Kishor and Dr. P. Sammulal, SOM: A Novel Approach to Mine Association Rules from Transactional Databases, National Conference on Advanced Computing and Pattern Recognition, JNTUHCEJ, India, 2014.
[19] Peddi Kishor and Dr.P.Sammulal, "Literature Survey on Association Rule Discovery in Data Mining", International Journal of Computer Science and Management Research, Vol 2 Issue 1 January 2013 pp.1408-1414.
[20] Puteri N.E. Nohuddin, Frans Coenen, Rob Christley, Christian Setzkorn, Yogesh Patel, Shane Williams, Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps, In Knowledge-Based Systems, Volume 29, 2012, Pages 104-113, ISSN 0950-7051.
[21] Nohuddin P.N.E., Christley R., Coenen F., Patel Y., Setzkorn C., Williams S. (2011) Social Network Trend Analysis Using Frequent Pattern Mining and Self Organizing Maps. In: Bramer M., Petridis M., Hopgood A. (eds) Research and Development in Intelligent Systems XXVII. Springer, London.
[22] A.J Richardson, C Risien, F.A Shillington, Using self-organizing maps to identify patterns in satellite imagery, In Progress in Oceanography, Volume 59, Issues 2–3, 2003, Pages 223-239, ISSN 0079-6611.

Authors

Mr. PEDDI KISHOR is currently working as HOD & Associate Professor in Computer Science and Engineering department of Sree Chaitanya Institute of Technological Sciences, Karimnagar, Telangana State, India. Currently he is pursuing Ph.D. in CSE branch from JNT University, Hyderabad. He obtained his M.Tech & B.Tech degrees in Computer Science and Engineering from JNT University, Hyderabad. He has over 12 years of experience in teaching various Computer Science subjects at Under Graduate & Post Graduate level. He has published papers in various national & international Conferences/Journals. His research interests include Data Mining, Web Technologies and Mobile Computing.

He is currently working as an Associate Professor in Computer Science and Engineering, JNTUH College of Engineering (JNTUH University), Jagtial, Karimnagar, Telangana State, India. Dr.Sammulal Porika obtained his Ph.D (Computer Science and Engineering) from Osmania University, Hyderabad, India in 2010. He received his B.E. degree from Osmania University in 2000 and M.Tech. from JNT University in Computer Science and Engineering, in 2002. He has been into teaching and research since 2002. In the classroom environment, the interpersonal relationship between teacher and students is an important element contributing to the learning process of students. Dr. Sammulal has always been dedicated to improve the quality of education and uses interactive teaching methods in his every-day teaching process. He has published more than 50 papers in various international conferences/journals. His research interests include Parallel, Distributed, Cluster, Grid and Cloud Computing Algorithms.