ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Classified Data Organizing Maps: A Survey

Deepika Kourav[1], Asha Khilrani[2]

Department of CSE, TIT & Science Bhopal[1,2]

Abstract: Self-Organizing Map (SOM) is in the greater concern in these days due to the hierarchy creating in document organization. It will be better for those problems where we needed clustering and visualization. Our direction of this paper is to find better clustering and classification techniques which will be profound in document organization. So in this paper we discuss on the deficiencies find in the traditional techniques and about the possible solutions which can be better in document organization.

Keywords: SOM, Clustering, Classification, Visualization

I. INTRODUCTION

One of the publicly hands-on models is the self-organizing map (SOM) sculpture [1]. The SOM learns immigrant swaggering dimensional statistics and maps them on a degraded, usually 2, dimensional map in a topology-preserving manner [2]. Focus is, data put in order converge in high-dimensional gap strength on top of everything else be close in the mapped low-dimensional space. Such capabilities vindicate the SOM gross widely applied in data visualization and clustering tasks [3].

In real-world exigency, purposefulness maker's bear perpetually suitably add to ambivalent objectives and an expansive fulfill gap with contrastive runner alternatives [4][5]. The multi-criterion making also provide a better combat in the near future. Its involves span rigorous spaces like the design space, incorporating the defining variables of the candidate solutions, and the intention space, constituting the mapping of each candidate solution to the multiple objective functions values[6]. The latter is the space where optimality is get under way, tradeoffs are explored, and decisions are normally reached. So there is the need of classification based on multiple decision criteria which can be heuristic, it will be possible by user defined constraints and multiple selective constraints.

It can be better to find a proper clustered way to organize the documents, then apply some classification criteria which will be satisfied some threshold value to provide the constrained way of these issue. It can be achieved through association rule mining [7], we can use partitioning technique also because it can reduce the searching time and enhance the searching capability [8][9].

For classification we can use association rule mining with some clustering techniques like K-means and fuzzy c-means, it will be a better option [10]. Then we can optimize it using several optimization techniques like Ant Colony optimization (ACO), Particle swarm Optimization, Mimetic algorithm etc.[11][12][13].

Subset superset partitioning can be used for partitioning and better classification [14]. The remaining of this paper is organized as follows.

We discuss clustering and classification techniques in Section 2. In Section 3 we discuss about literature review. In section 4 we discuss about problem domain. In section 5 we discuss about the proposed framework. Conclusions are given in Section 6. Finally references are given.

II. CLUSTERING AND CLASSIFICATION TECHNIQUES

For classifying pair accustomed approximate worn k-means algorithm are a handful of the broadly authorized clustering tools that are expedient in a trade mark of painstaking and domain applications. K-means groups the materials in conform with regard to their circumstance sentiment into K intrepid clusters. Data categorized into the alike cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance.

The steps involved in a K-means algorithm are given subsequently [15]:

1) K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2) The data are assigned to the group that is adjacent to the centroid.
3) The positions of all the K centroids are recalculated as soon as all the data are assigned.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

Association Rule mining is one of the important and most popular matter mining techniques. Federation head up mining gluteus Maximus be efficiently used in any decision making processor decision based leadership generation. In data mining appointment in consequently so we courage find the frequent patterns to know the effective patterns from the huge data. Change we find positive and negative rules [16]. If we agree to the beyond everything phenomena change we come to the point that the rule generation is also huge. In this compounding we metaphysical join aspects of optimization techniques by which we can optimize the association rules. So hybridization is needed [17]. Turn to course mining is a efficacious movement capable of identifying in a used of objects (called items) those which demonstrate similar behavior. For event, in a Stock Exchange, clientele object encode are kept as storekeeper business each includes a set of items purchased together. Analyzing the used of merchant may discuss to fact mosey are frequently purchased together.

The Ant Colony Optimization algorithm is mainly inspired by the experiments run by Goss et al. [18] which using a grouping of real ants in the real environment. They study and observe the behavior of those real ants and suggest that the real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two. This ant behavior was first formulated and arranged as Ant System (AS) by Dorigo et al. [19][20]. Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed [19]. In ACO algorithm, the optimization problem can be expressed as a formulated graph G = (C; L), where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C [21]. Optimization technique can provide better classification strategy and it is used in different area of engineering[22][23][24][25].

III. LITERATURE REVIEW

In 2011, Avrilia Floratou et al. [26] proposed a new algorithm called FLexible and Accurate Motif DEtector (FLAME). FLAME is a flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics.

In 2011, Shawana Jamil et al. [27] focus on investigation of mining frequent sub-graph patterns in DBLP uncertain graph data using an approximation based method. The frequent sub-graph pattern mining problem is formalized by using the expected support measure. Here n approximate mining algorithm based Weighted MUSE, is proposed to discover possible frequent sub-graph patterns from uncertain graph data.

In 2011, Ashwin C S et al. [28] proposed an apriori- based method to include the concept of multiple minimum supports (MMS in short) on association rule mining. It allows user to specify MMS to reflect the different natures of items. Since the mining of sequential pattern may face the same problem, we extend the traditional definition of sequential patterns to include the concept of MMS in this study. For efficiently discovering sequential patterns with MMS, we develop a data structure, named PLMS-tree, to store all necessary information from database.

In 2011, K. Zuhtuogullari et al. [29] observe that an extendable and improved item set generation approach has been constructed and developed for mining the relationships of the symptoms and disorders in the medical databases. The algorithm of the developed software finds the frequent illnesses and generates association rules using Apriori algorithm. The developed software can be usable for large medical and health databases for constructing association rules for disorders frequently seen in the patient and determining the correlation of the health disorders and symptoms observed simultaneously.

In 2010, Hsin-Chang Yang et al. [30] suggest that the SOM has main disadvantage of the need to know the number and structure of neurons prior to training, which are difficult to be determined. Several schemes have been proposed to tackle such deficiency. Examples are growing/expandable SOM, hierarchical SOM, and growing hierarchical SOM. These schemes could dynamically expand the map, even generate hierarchical maps, during training. Encouraging results were reported. Basically, these schemes adapt the size and structure of the map according to the distribution of training data. That is, they are data-driven or data oriented SOM schemes. In this work, a topic-oriented SOM scheme which is suitable for document clustering and organization will be developed. Their proposed SOM will automatically adapt the number as well as the structure of the map according to identified topics. Unlike other data-oriented SOMs, our approach expands the map and generates the hierarchies both according to the topics and their characteristics of the neurons. The preliminary experiments give promising result and demonstrate the plausibility of the method.

In 2013, Hsin-Chang Yang et al. [31] two major deficiencies of classical SOM are the need of predefined map structure and the lack of hierarchy generation. Several approaches have been devised to tackle these deficiencies. They suggest that both structural and topical constraints which specified by the user could be used to guide the

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

learning process. Preliminary experiments demonstrate improvements over previous algorithm on text categorization task.

IV. PROBLEM DOMAIN

After studying several research papers we observe the the following problem findings:

1) Data partitioning can be easily implemented for reducing the size and searching.
2) Different levels of Constraints are also applied.
3) Leraning can be applied in different steps to refine the search.
4) Clustering and classification model can be applied together.
5) Some optimization techniue can be applied for defining a threshold limit.
6) Classification can be heuristic for applying the meta search.
7) Need to dynamize the map structure and organize it in proper hierarchy.
8) Tree slave structure is also useful for extracting the data in breadth first seach way.

V. ANALYSIS

We provide the analysis of the paper in table 1.

Table 1: Analysis

Authors	Technique	Achieve
Andreas Rauber[32]	Hierarchical SOM	The motivation was to provide a model that adapts its architecture during its unsupervised training process according to the particular requirements of the input data.
Sebastián Moreno[33]	Robust Growing Hierarchical Self Organizing Map	The outliers introduce an influence to the GHSOM model during the training process by locating prototypes far from the majority of data and generating maps for few samples data.
Carolina Saavedra.[34]	KDSOM	a hybrid model called K-Dynamical Self Organizing Maps (KDSOM) consisting of K Self Organizing Maps with the capability of growing and interacting with each other are proposed.
Héctor Allende[35]	Robust Self-organizing Maps	They propose a variant to the learning algorithm that is robust under the presence of outliers in the data by being resistant to these deviations.
Khabia et al. [36]	Classification of Web Results	They suggest feature selection and learning clustering of text documents can be used, which requires small amount of training data. Clustering process is itself feature learning step.
Rodrigo Salas[37]	FASOM	A hybrid algorithm called Flexible Architecture of Self Organizing Maps (FASOM) that overcomes the Catastrophic Interference and preserves the topology of Clustered data in changing environments.

VI. CONCLUSION

In this paper we survey several aspects of SOM and the flaws presented in the previous technique. We also find some useful trends in the previous technique which can be incorporated with clustering and association to form a hybrid technique for proper classification and maintaining the document hierarchy. The scopes are in the direction of hybrid framework with the formation of advance structural classifier.

REFERENCES

[1] T. Kohonen, Self-Organizing Maps. Berlin: Springer-Verlag, 1997.
[2] A. Rauber, M. Dittenbach, and D. Merkl, “Towards automatic contentbased organization of multilingual digital libraries: An English, French and German view of the Russian information agency Nowosti news,” in Proceedings of the Third All-Russian Scientific Conference on Digital Libraries: Advanced Methods And Technologies, Digital Collections, September 11-13 2001, pp. 11–13.
[3] A. Rauber, D. Merkl, and M. Dittenbach, “The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data,” IEEE Transactions on Neural Networks, vol. 13, no. 6, pp. 1331–1341, 2002.
[4] M. Bagajewicz and E. Cabrera. Pareto optimal solutions visualization techniques for multiobjective design and upgrade of instrumentation networks. Industrial and Engineering Chemistry Research, 42(21):5195–5203, 2003.
[5] W. Berger, H. Piringer, P. Filzmoser, and E. Gr ¨oller. Uncertainty aware exploration of continuous parameter spaces using multivariate prediction. Computer Graphics Forum, 30(3):911 – 920, 2011.
[6] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multi objective Selection Based on Dominated Hypervolume. European Journal of Operational Research, 2007.
[7] Dubey, Ashutosh K., and Shishir K. Shandilya. "A novel J2ME service for mining incremental patterns in mobile computing." Information and Communication Technologies. Springer Berlin Heidelberg, 2010.
[8] Pragati Shrivastava, Hitesh Gupta,” A Review of Density-Based clustering in Spatial Data”, International Journal of Advanced Computer Research (IJACR), Volume-2 Number-3 Issue-5 September-2012.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

[9] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.
[10] Shyi-Ching Liang, Yen-Chun Lee and Pei-Chiang Lee , "The Application of Ant Colony Optimization to the Classification Rule Problem", 2011 IEEE International Conference on Granular Computing.
[11] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3 Number-1 Issue-9 March-2013.
[12] Arezoo Modiri and Kamran Kiasaleh," Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011.
[13] Yao Liu and Yuk Ying Chung, "Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning", IEEE 2011.
[14] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support", Conseg-2012.
[15] Gupta Chetan, Amit Sinhal, and Rachana Kamble. "Intrusion Detection based on K-Means Clustering and Ant Colony Optimization: A Survey." International Journal of Computer Applications 79 (2013).
[16] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3 Number-1 Issue-9 March-2013.
[17] Anshuman Singh Sadh, Nitin Shukla, "Apriori and Ant Colony Optimization of Association Rules", International Journal of Advanced Computer Research (IJACR), Volume-3 Number-2 Issue-10 June-2013.
[18] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels. Self-organized Shorcuts in the Argentine Ant. Naturwissenschaften, 76:579–581, 1989.
[19] M. Dorigo, Gianni Di Caro, and Luca M. Gambardella. Ant Algorithms for Discrete Optimization. Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA, Universite Libre de Bruxelles, Brussels, Belgium, 1998.
[20] M. Dorigo and M. Maniezzo and A. Colorni. The Ant Systems: An Autocatalytic Optimizing Process. Revised 91-016, Dept. of Electronica, Milan Polytechnic, 1991.
[21] M. Dorigo and G. Di Caro. New Ideas in Optimisation. McGraw Hill, London, UK, 1999.
[22] Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. "A Survey on Breast Cancer Scenario and Prediction Strategy." In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, pp. 367-375. Springer International Publishing, 2015.
[23] Vitthal Manekar, Kalyani Waghmare, " Intrusion Detection System using Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) " , International Journal of Advanced Computer Research (IJACR), Volume-4, Issue-16, September-2014 ,pp.808-812.
[24] Suman Mishra, Prateek Gupta, " An efficient Optimization method for Data Classification " , International Journal of Advanced Computer Research (IJACR), Volume-4, Issue-14, March-2014 ,pp.383-388.
[25] Animesh Dubey, Rajendra Patel and Khyati Choure, " An Efficient Data Mining and Ant Colony Optimization technique (DMACO) for Heart Disease Prediction " , International Journal of Advanced Technology and Engineering Exploration (IJATEE), Volume-1, Issue-1, December-2014 ,pp.1-6.
[26] Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel," Efficient and Accurate Discovery of Patterns in Sequence Data Sets", IEEE Transactions On Knowledge and Data Engineering, VOL. 23, NO. 8, August 2011.
[27] Shawana Jamil, Azam Khan, Zahid Halim and A. Rauf Baig," Weighted MUSE for Frequent Sub-graph Pattern Finding in Uncertain DBLP Data", IEEE 2011.
[28] Ashwin C S, Rishigesh.M and Shyam Shankar T M," SPAAT-A Modern Tree Based Approach for sequential pattern mining with Minimum support",IEEE 2011.
[29] K. Zuhtuogullari and N. Allahverdi ,"An Improved Itemset Generation Approach for Mining Medical Databases", IEEE 2011.
[30] Yang, Hsin-Chang, Chung-Hong Lee, and Kuo-Lung Ke. "TOSOM: A Topic-Oriented Self-Organizing Map for Text Organization." World Academy of Science, Engineering and Technology 65 (2010): 1100-1104.
[31] Yang, Hsin-Chang, Chung-Hong Lee, and Chun-Yen Wu. "Incorporating user constraints into topic-oriented self-organizing maps." Foundations of Computational Intelligence (FOCI), 2013 IEEE Symposium on. IEEE, 2013.
[32] Rauber, A.; Merkl, D.; Dittenbach, M., "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data," Neural Networks, IEEE Transactions on , vol.13, no.6, pp.1331,1341, Nov 2002.
[33] Sebastián Moreno, Héctor Allende, Cristian Rogel, Rodrigo Salas," Robust Growing Hierarchical Self Organizing Map", Lecture Notes in Computer Science Volume 3512, 2005, pp 341-348.
[34] Carolina Saavedra, Héctor Allende, Sebastián Moreno, Rodrigo Salas," K-Dynamical Self Organizing Maps", Lecture Notes in Computer Science Volume 3789, 2005, pp 702-711.
[35] Allende, Héctor, et al. "Robust self-organizing maps." Progress in Pattern Recognition, Image Analysis and Applications. Springer Berlin Heidelberg, 2004. 179-186.
[36] Apeksha Khabia and M. B. Chandak, " A Cluster Based Approach for Classification of Web Results " , International Journal of Advanced Computer Research (IJACR), Volume-4, Issue-17, December-2014 ,pp.934-938.
[37] Salas, Rodrigo, et al. "Flexible architecture of self-organizing maps for changing environments." Progress in Pattern Recognition, Image Analysis and Applications. Springer Berlin Heidelberg, 2005. 642-653.