Class Based Clustering with Cuckoo Search Rank Optimization for Text Data Categorization

Vinita Jain
Research Scholar, Department of CSE
PCST, Bhopal, India
E-mail-vinitamtech@gmail.com

Neha Shrivastava
Asst. Professor
PCST, Bhopal, India
E-mail- nehashrivas@gmail.com

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 have proposed associative cluster to categorize the text data and then apply cuckoo search algorithm for finding the optimize rank in the related class hierarchy. For experimentation we have applied our approach on 8 different data and achieve better results in comparison to the previous methods.

Keywords: - Cuckoo search Optimization, K-Means, Parser, data categorization.

1 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 rest of paper is organized as follows. In Section II Related work. The Section III state the problem IV discusses proposed methodology. In section V discuss performance evaluation and result analysis followed by a conclusion in Section VI.

II RELATED WORK

In 2011, Avrilia Floratou et al. [22] 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. [23] focus on 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. [24] 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. [25] 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. The subset superset rules with association is also suggested in [26] and for the cancer the same rules are suggested with optimization in [27].

In 2010, Hsin-Chang Yang et al. [28] 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. [29] 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 learning process. Preliminary experiments demonstrate improvements over previous algorithm on text categorization task.

III PROBLEM DOMAIN

We have studied various research and journal papers related to intrusion data classification. According to our research we have analyzed that many of the papers focuses on the problem of better classification of intrusion data and to use an optimized technique for it. Few review of summary described here and implicated with their respective author.

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

1) Data partitioning can be easily implanted for reducing the size and searching. [30]
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. [31]
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. [32]
8) Tree slave structure is also useful for extracting the data in breadth first seach way

IV PROPOSED METHODOLOGY

In this paper we have presented a hybri algorithm for text data categorization. In our approach we have used three different approaches in a single framework. The first approach is the pruner which prune the data so that text ranking can be easily identified. We are using four different pruner. First pruner is Number Pruner (N Pruner) for the identification and removal of numeric data. Second pruner is Delimiter Pruner (D Pruner) for the identification and removal of Delimiter. Third pruner is Universal Pruner (U Pruner) for the identification and removal of noun and verb. Third pruner is Database manual Pruner (DT Pruner) for the removal of unwanted data manually. Then the associated items with their frequency and parse removed data will be sent to the next step. This is the refined data which will be goinf for the class ased clustering and optimization by Cuckoo Search Optimization (CSO). This process is better understand by the below figure and the algorithms. We have applied K-Means and make it the seperation of 10 slots each and user can put the minimum frequency for the categorization and selection as per algorithm given in phase II. Then we have applied CSO algorithm with two objective function named maximization and minimization. After finding the max nad min value we will select the to 10 results from the overall categorization. The results shows the success of our algorithm with the help of different category of data.

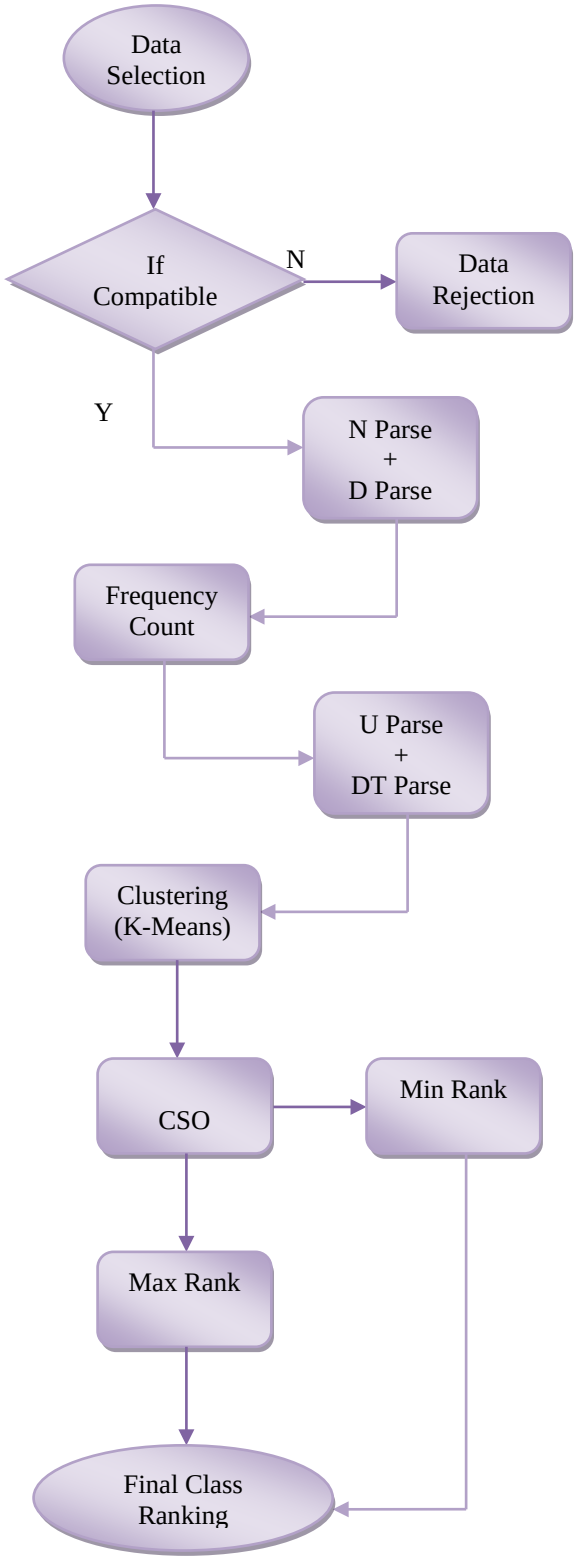

Figure 1: Flowchart for proposed work.

Algorithm : Associated K-Means with Cuckoo Search

Phase-I

Step1: A textual data set will be considered first.
Step 2: It is first parse to remove the textual data and all the delimiters by the help of N parse and D parse.
Step 3: Frequency of the individual text is calculated from the whole document.
Step 4: Then U parse and DT Parse can be applied for more filteration.
Step 5: Received output will be send to Phase II.

Phase-II [33,34]

First data points are initialized. In our case the data points are the text values.
Step 1: X = {x1,x2,x3,……..,xn}
Then center points are initialized. In our case it is distributes based on 10 categories of 10 different slots.
Step 2: V = {v1,v2,…….,vc} be the set of centers.
Step 3: User selects 'c' cluster centers.
Step 4: then the distance between the data points and the center points are calculated.
Step 5: Dole out the information point to the group focus whose separation from the bunch focus is least of all the bunch focuses.
Step 6: Recalculate the new cluster center using:
$\mu i=1|ci|\sum j\in cixj,\forall i$
where, 'ci' represents the number of data points in ith cluster.
Step 7: Recalculate the separation between every information point and new got bunch focuses.
Step 8: On the off chance that no information point was reassigned then stop, overall rehash from step 3.

Phase-III [35,36]

There are three basic rules of CSO. The rules are following:

1. Each one cuckoo lays one egg at once, and dumps its egg in a haphazardly picked home.
2. The best homes with high caliber of eggs will persist to the cutting edge;
3. The quantity of accessible hosts homes is altered, and the egg laid by a cuckoo is found by the host feathered creature with a likelihood in (0,1).

Step 1: Objective function: f(min) & f(max) < > [0,1]
Step 2: Create a starting populace of n host homes;
Step 3: While (t<MaxGeneration)
Get a cuckoo haphazardly (say, i) and supplant its answer by performing Lévy flights
Step 4: Evaluate its quality/fitness F_i
For maximization
Choose a nest among n (say, j) randomly;
Step 5: if (F_i>F_j),
Supplant j by the new arrangement;
end if

A division (p_a) of the more terrible homes are relinquished and new ones are manufactured;
Keep the best arrangements/homes;
Rank the arrangements/homes and find the current best;
Pass the current best answers for the cutting edge;
end while.

V EXPERIMENTAL DETAILS AND RESULT ANALYSIS

In this paper we perform experimental process of proposed method. For result analysis we have selected several data file and out of these we have put 8 data for the result consideration. After analyzing the results from below Figure, we will achieved 97 % accuracy as shown in table 1.

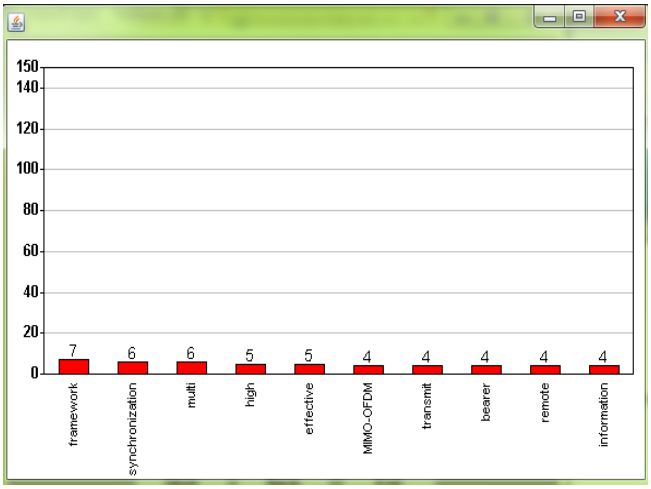

Figure 2: OR of Data 1.

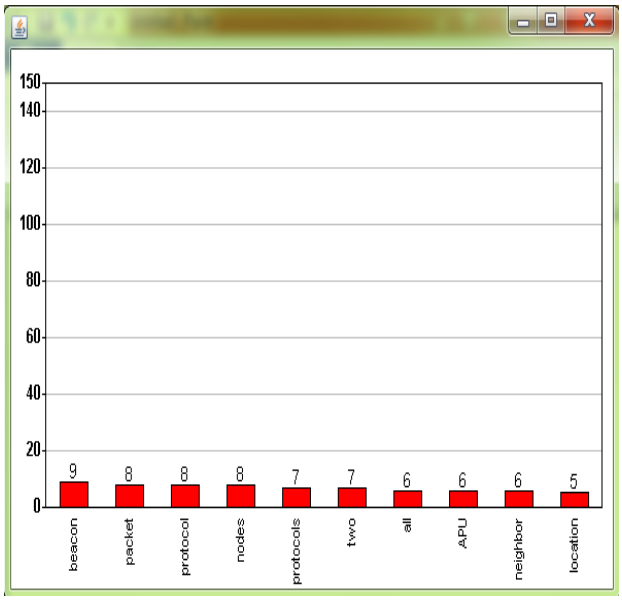

Figure 3: OR of Data 2.

Method	Accuracy (%)
small/shallow C-TOSOM	87.64
small/deep C-TOSOM	92.65
large/shallow C-TOSOM	92.89
TOSOM	92.47
Clustering with Cuckoo Search rank Optimization	97

Table 1: Comparison table for performance evaluation.

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 is presented in this paper. By our algorithm we have achieved better results in comparison to the traditional technique.

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.

[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] 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.

[23] Shawana Jamil, Azam Khan, Zahid Halim and A. Rauf Baig," Weighted MUSE for Frequent Sub-graph Pattern Finding in Uncertain DBLP Data", IEEE 2011.

[24] 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.

[25] K. Zuhtuogullari and N. Allahverdi ,"An Improved Itemset Generation Approach for Mining Medical Databases", IEEE 2011.

[26] Dubey, A.K.; Dubey, A.K.; Agarwal, V.; Khandagre, Y., "Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support," Software Engineering (CONSEG), 2012 CSI Sixth International Conference on , pp.1,6, 5-7 Sept. 2012.

[27] 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.

[28] 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.

[29] 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.

[30] Sharma, Prashant. "Association Rule Mining with enhancing List Level Storage for Web Logs: A Survey." International Journal of Advanced Technology and Engineering Exploration vol 1, issue 1 (2014).

[31] Kai Li, Lijuan Cui, "A Kernel Fuzzy Clustering Algorithm with Generalized Entropy Based on Weighted Sample" , International Journal of Advanced Computer Research (IJACR), Volume-4, Issue-15, June-2014 ,pp.596-600.

[32] D.Bujji Babu, R.Siva Rama Prasad, Y.Umamaheswararao , " Efficient Frequent Pattern Tree Construction " , International Journal of Advanced Computer Research (IJACR), Volume-4, Issue-14, March-2014 ,pp.331-336.

[33] Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, no. 7 (2002): 881-892.

[34] Bennett, K., P. Bradley, and Ayhan Demiriz. "Constrained k-means clustering." Techinical Repport, Microsoft Corp (2000).

[35] Yang, Xin-She, and Suash Deb. "Cuckoo search via Lévy flights." In Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, pp. 210-214. IEEE, 2009.

[36] Yang, Xin-She, and Suash Deb. "Engineering optimisation by cuckoo search." International Journal of Mathematical Modelling and Numerical Optimisation 1, no. 4 (2010): 330-343.