A Survey of Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data

Bhagyashree Bhoyar
ME Computer (Engineering), Dr. D. Y. PATIL INSTITUTE OF ENGINEERING AND TECHOLOGY, Savitribai Phule Pune University, Pune, India .

Pramod Patil
Department of Computer Engineering DYPIET, Savitribai Phule Pune University, Pune,Indai.

Priyanka Abhang
ME Computer (Engineering), Dr. D. Y. PATIL INSTITUTE OF ENGINEERING AND TECHOLOGY, Savitribai Phule Pune University, Pune, India

Abstract-Big Data describes a technology used to store and process the exponentially increasing dataset which contains structured, semi structured and unstructured data that has to be mined for valuable information. It deals with 3 V's: Volume, Variety and Velocity of processing data. It is associated with cloud computing for the analysis of large data sets in real time. Volume refers to the huge amount of data it collects, Velocity refers to the speed at which it process the data and Variety defines that data does not mean just numbers, dates or strings but also geospatial data,3D data, audio, video, social files, etc. In this system develop a new approach to feature subset ranking for feature selection in classification problems with the goal of using a small number of features to achieve better classification performance.

Keyword:- Data mining, Data models, Big data, Classification algorithms, Decision trees, Load modeling, Algorithm design and analysis, Particle Swarm Optimization, Feature Selection, Meta heuristics, Swarm Intelligence, Classification, Big Data.

I. INTRODUCTION

In recent years, many companies are collecting huge amount of data, often generated continuously as a sequence of events and coming from various locations. Credit card transactional flows, telephone records, sensor network data, network event logs are just some examples of data streams. The design and development of fast, efficient, and accurate techniques, able to extract knowledge from these huge data sets, too large to fit into the main memory of computers, cause significant challenges. First of all, data can be analyze only once, as it arrives. Second, using an entire data stream of a long period could mislead the analysis results due to the weight of outdated data, considered in the same way as more recent data. Since data streams are continuous sequences of information, the underlying clusters could change with time, thus giving different results with respect to the time horizon over which they are computed.

The big data mining is now kept on blooming in different online services and provides a best service to end users or customers. These tools are very useful to end users in providing quality service and an efficient tool to be used in system detected by cyber-attacks. These big data helps the users to retrieve the data as per their wish. Hence feature selection and classification plays an important role in this big data to retrieve or search a data from a variety of big data sets. Also more efficient algorithms must be implemented when dealing with big data. Big data depends on 3 V's such as Volume, Velocity and Variety. These 3 V's are main characteristic function in selecting clustering techniques. First V, Volume of data is very important in clustering process as they require storage space. Second V, Velocity deals with the processing speed according to data flow. Third V, Variety depends on the data types. It may be image, text or a video generated from various sources such as mobile phones, camera or sensors. The feature selection and clustering are the two steps very important in systems which use different domains such as pattern recognition, machine learning, bio-informatics, data mining, semantic ontology and in image retrieval. In both families of data mining algorithms, stream-based and batch-based, classification has been widely adopted for

supporting inferring decisions from big data. In supervised learning, a classification model or classifier is trained by inducing the relationships between the attributes of the historical records and the class labels which are usually the predictor features of all the data and their predicted classes respectively. Subsequently, the classifier is used to predict appropriate classes given unseen samples. In classifier applications, feature selection attempts to select a subset of the most influential features by excluding irrelevant and redundant features in order to enhance accuracy and speedup model training time for the classifier. In the past many computer science researchers studied about using heuristics to tackle the feature selection problems.[1]

The purpose of the study is to explore the effectiveness of PSO and APSO in datamining structure and enhancing classification accuracy. To achieve this objective, the following tasks must be conducted:-

1. To develop CFS/PSO and APSO Structure for feature extraction
2. To analyze and Minimize the Number of Features
3. To evaluate Classification Performance
4. To evaluate Features Accuracy
5. To develop Classification using SVM and KMENAS

II. COMPARISON OF VARIOUS PSO BASED DATA MINING METHODS

Table. Comparison of various PSO based data mining methods

Paper referred	**Clustering Algorithm**	**Dataset**	**Evaluation parameters**	**Future Work**
DW van der Merwe AP Engelhrecht [26]	Gbest PSO, Hybrid PSO and K-means algorithm	Iris , Breast Cancer, Wine, Automotives	Quantization error, Inter cluster distance and intra cluster distance	Extend the fitness function to optimize the inter and intra cluster distances, Experiment on higher dimensional problems-and large number of patterns
Neveen I. Ghali, Nahed El-Dessouki, Mervat A. N., and	PSO, Exponential Particle Swarm Optimization (EPSO)	Breast cancer, Iris, Yeast, Lences, Glass	Quantization error	-------
Surat Srinoy and Werasak Kurutach [32]	Hybrid artificial ant cluster algorithm and kmean particle swarm	KDD'99 data set	Recognition of known network attacks	-------
Esmaeil Mehdizadeh[34]	Fuzzy PSO alogrithm	Artificial data set, iris, wine and image	Objective function value and CPU time	-------
Hesam Izakian, Ajith Abraham, Václav Snášel[33]	Hybrid fuzzy c-means fuzzy particle swarm algorithm for clustering	Iris , Cancer, Wine, glass, CMC, vowel	Objective function values	-------
T. Niknam, M. Nayeripour and B.Bahmani Firouzi [35]	Particle swarm optimization - ant colony optimization (PSO-ACO)	Iris, Wine, Vowel and CMC	Function value, Standard deviation and number of function evaluation	-------
K. Premalatha and A.M.	PSO with local search	Iris , Wine, glass	Fitness value , Inter and Intra Cluster	-------

Xiang Xiao, Ernst R. Dow, Russell Eberhart, Zina Ben Miled and Robert J. Oppelt [37]	Hybrid SOM –PSO algorithm	Yeast data set and rat data set	Average merit, execution time	-------
N. M. Abdul Latiff, C. C. Tsimenidis, B. S. Sharif and C.	Binary PSO with multi-objective clustering approach (DCBMPSO)	100 nodes	Number of cluster, network lifetime and delivery of data messages	To investigate the DC-BMPSO algorithm properties such as the effect of varying algorithm parameter, *init p* on the
Sandeep Rana, Sanjay Jasola, Rajesh Kumar	PSO in sequence with K-Means	Artificial problem, Iris and wine	Quantization error, Inter and Intra Cluster distance	Variations in PSO algorithm and its hybridization with K-Means algorithm
Jakob R. Olesen, Jorge	*AutoCPB*	Artificial dataset,	QEF metric, ID metric, number of	To identify a rule to minimize local optima, to apply to other domains such

Cordero H., and Yifeng Zeng [40]		Iris, Wine, Pima, Haberman, Breast Cancer,	clusters and elapsed times	as attribute clustering, more specific analysis of parameter setting
J.Hyma, Y.Jhansi and S.Anuradha [41]	Hybrid PSO and Genetic algorithm	Document dataset	Intra Cluster distance	To extend this work to deal with other sorts of documents.
Swagatam Das, Ajith Abraham, Amit Konar [45]	Kernel_MEPSO (Multi-Elitist PSO) algorithm	Synthetic dataset, Glass, Wine ,Breast cancer, Image , segmentation and Japanese vowel	Mean and standard deviation of the clustering accuracy, Mean and standard deviation of the number of clusters, unpaired t-tests, execution time, Mean and standard	Improve the performance of the algorithm over high dimensional datasets by incorporating some feature selection mechanism in it. Automatic clustering in lower dimensional subspaces with MEPSO may also be a worthy topic of further research
Fun Ye and Ching-Yi Chen [46]	Alternative KPSO-clustering (AKPSO)	Artificial datasets and iris dataset	Cluster center location, distortion measure	-------
K. Premalatha and A.M. Natarajan[42]	Hybrid PSO and Genetic algorithm	Library Science, Information Science and	Fitness value	-------
Alireza Ahmadyfard and Hamidreza Modares[[27]	PSO-Kmeans	synthetic data sets (SET I, SET II and SET III), Iris	Error rate and Mean square error	-------

Yannis Marinakis, Magdalene Marinaki, and Nikolaos Matsatsinis[29]	Hybrid PSO-GRASP(Greedy Randomized Adaptive Search Procedure)	Australian Credit Breast Cancer Wisconsin 1 (BCW1) Breast Cancer Wisconsin 2 (BCW2) Heart Disease (HD) Hepatitis 1 (Hep1)	Feature selection	Use of different algorithms for both feature selection phase and clustering algorithm phase.
Wensheng Yi1, Min Yao and Zhiwei Jiang [30]	fuzzy particle swarm optimization clustering algorithm.	Iris, Wine, Ionosphere, sunset/sunrise images, beach and grassland	Entropy and cluster purity	To extend the clustering algorithm to video stream, Improvement of the speed of the algorithms
Ryan K. Johnson, Ferat Sahin[48]	PSO (Interia methods, Inertia with predator prey option, and Constriction with predator prey option)	Iris, Breast Cancer, Wine,E. Coli, Segmentation	Quantization Error (mean), Quantization Error (std. dev.)	Clustering of dynamic data
Jen-Ing G. Hwang, Chia-Jung Huang [44]	Hybrid scheme of differential evolution based PSO and fuzzy *c*-means(EDPSO)	Iris, Breast Cancer, Wine	Effect of perturbed velocity, determine an appropriate number of Clusters, Jaccard index	-------
Mahamed G. H. Omran , Ayed Salman and Andries P. Engelbrecht[31]	Dynamic clustering algorithm based on PSO (DCPSO)	Synthetic images,Lenna, mandrill, jet, peppers, MRI and Lake Tahoe	Mean and Standard deviation	Application of the DCPSO algorithm to general data, to investigate the effect of high dimensionality on the performance of the DCPSO, use of other clustering algorithms such as FCM and KHM to refine the cluster centroids, incorporation of
Xiaohui Cui, Thomas E. Potok [43]	Hybrid Particle Swarm Optimization (PSO) and K-means	TREC-5, TREC-6, and TREC-7	Average distance between documents and the cluster	-------

Sridevi.U. K., Nagaveni. N. [47]	PSO clustering using ontology similarity	NewsGroups	Sum of squared error, Precision, Recall, F-measure, Time in	Fuzzy ontology based methodology for clustering knowledge and personalized searching method
Ching-Yi Chen and Fun Ye[17]	PSO clustering algorithm	Artificial data set	Object function and Cluster centre	
Shi M. Shan, Gui S. Deng, Ying H. He[49]	Hybridization of Clustering Based on Grid and Density with PSO (HCBGDPSO	Artificial dataset	Shape of clusters	To devise an application and finding a way of adaptively tuning the parameters in HCBGDPSO

III. CONCLUSION

APSO is designed to be used for data mining of data streams on the fly. The combinatorial explosion is addressed by used swarm search approach applied in incremental manner. This approach also fits better with real-world. applications where their data arrive in streams. In addition, an incremental data mining approach is likely to meet the demand of big data problem in service computing.

REFERENCES

[1] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, Senior Member, IEEE "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data" IEEE TRANSACTIONS.

[2] Guoyin, W; Jun, H; Qinghua, Z; Xiangquan, L; Jiaqing, Z (2008). "Granular computing based data mining in the view of rough set and fuzzy set". In International conference on Granular computing. Proceedings in IEEE GRC. pp 67–67

[3] Shanli.W(2008) Research on a new effective data mining method based on neural networks. In International symposium on electronic commerce and security. pp 195–198

[4] Frigui, H, Krishnapuram, R (1999). "A robust competitive clustering algorithm with applications in computer vision," IEEE Trans. Pattern Anal. Mach Intell., vol. 21, no. 5, pp. 450–465.

[5] Leung, Y; Zhang, J; Xu,Z (2000). "Clustering by scale- space filtering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1396–1410.

[6] Jain, A. K ; Murty, M. N. Flynn, P. J. (1999). "Data clustering: a review. ACM Computing Survey31(3):264–323

[7] Steinbach, M; Karypis, G; Kumar, V. (2000). A Comparison of Document Clustering Techniques. TextMining Workshop, KDD.44

[8] Zhao. Y; Karypis G(2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, Machine Learning, 55 (3): pp.311-331.

[9] B. Pfahringer, G. Holmes, and R. Kirkby, "New Options for Hoeffding Trees", Proc. in Australian Conference on Artificial Intelligence, 2007, pp.90-99.

[10] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, pp.108-114, 1995

[11] Bifet A. and Gavalda R. "Learning from time-changing data with adaptive windowing". In Proc. of SIAM International Conference on Data Mining, 2007,pp. 443–448. [13]Wu, K.L, Yang, M.S (2002) Alternative C-meansClustering Algorithms. Pattern Recognition, 35, 2267- 2278

[12] Simon Fong, Suash Deb, Xin-She Yang, Jinyan Li, "Metaheuristic Swarm Search for Feature Selection in Life Science Classification", IEEE IT Professional Magazine, August 2014, Volume 16, Issue 4, pp.24-29.

[13] IT Professional Magazine, August 2014, Volume 16, Issue 4, pp.24-29. [15]Kennedy, J; Eberhart, RC (1995) Particle swarm optimization. In: Proceedings of IEEE conference on neural networks, Perth, Australia, pp 1942–1948.

[14]]Kennedy, J (1997). Minds and cultures: Particle swarm implications. Socially Intelligent Agents. AAAI Fall Symposium. Technical Report FS-97-02, Menlo Park, CA: AAAI Press, 67-72.

[15] Paterlini, S; Krink, T (2006) Differential evolution and particle swarm optimization in partitional clustering. Comput Stat Data Anal 50:1220–1247

[16] Chen, CY; Ye, F (2004). Particle swarm optimization algorithm and its application to clustering analysis. In: Proceedings of the IEEE international conference on networking, sensing and control. Taipei, Taiwan, pp789–794

[17] Niu, Y; Shen, L (2006) An adaptive multi-objectiveparticle swarm optimization for color image fusion. Lecture notes in computer science, LNCS. pp 473–480

[18] Silva, A; Neves, A; Costa, E (2002). Chasing the swarm: a predator pray appoach to function optimization. In: Proceedinge of the MENDEL, international conference on soft computing.

[19] Senthil, MA; Rao, MVC; Chandramohan, A (2005). Competitive approaches to PSO algorithms via new acceleration co-efficient variant with mutation operators. In: Proceedings of the fifth international conference on computational intelligence and multimedia applications

[20]]Hu, X; Shi, Y; Eberhart, RC (2004) Recent Advances in Particle Swarm, In Proceedings of Congress on evolutionary Computation (CEC), Portland, Oregon, 90- 97

[21] Shi, Y; Eberhart, RC (1998). A modified particle swarm optimizer. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Piscataway, NJ. 69-73

[22] Boeringer, D-W; Werner, DH(2004).”Particle swarm optimization versus genetic algorithm for phased array synthesis”. IEEE Trans Antennas Propag 52(3):771-779
[23] Junliang, L; Xinping, X (2008). Multi-swarm and multi- best particle swarm optimization algorithm. In: IEEE*International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, March 2013*world congress on intelligent control and automation. pp6281–6286
[24] Rana, S; Jasola, S; Kumar, R, “A review on Particle Swarm Optimization Algorithms and Applications to data clustering”. Springer Link Artificial Intelligence Review vol.35, issue 3:211–222.2011
[25] Van der Merwe ,DW; Engelhrecht, AP (2003) Data clustering using particle swarm optimization. In: Conference of evolutionary computation CEC’03, vol 1. pp 215–220
[26] Ahmadyfard, A; Modares, H (2008) Combining PSO and k-means to enhance data clustering. In: International symposium on telecommunications. pp 688–691
[27] Ghali, NI; Dessouki,NE ; Mervat A. N; Bakrawi, L(2008) Exponential Particle Swarm Optimization Approach for Improving Data Clustering. World Academy of Science, Engineering and Technology 42 .
[28] Marinakis, Y; Marinaki, M; and Matsatsinis, N (2007). A Hybrid Particle Swarm Optimization Algorithm for Clustering Analysis .DaWaK 2007, Lecture notes in computer science, LNCS 4654, pp. 241–250
[29] Yi, W; Yaoand, M; Jiang, Z(2006).Fuzzy Particle Swarm Optimization Clustering and Its Application to Image Clustering.
[30] Omran, M; Salman, A; Engelbrecht AP (2006). Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Anal Appl8:332–344
[31] Srinoy, S; Kurutach, W (2006).Combination Artificial Ant Clustering and K-PSO Clustering Approach to Network Security Model. International Conference on Hybrid Information Technology (ICHIT'06)
[32] Izakian,H ; Abraham, A; Snášel V(2009)Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle SwarmOptimization. World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)
[33] Mehdizadeh, E (2009) A fuzzy clustering PSO algorithm for supplier base management. International Journal of Management Science and Engineering Management Vol. 4 (2009) No. 4, pp. 311-320
[34] Niknam,T; Nayeripour, M; Firouzi, BB(2008). Application of a New Hybrid optimization Algorithm on Cluster Analysis Data clustering. World Academy of Science, Engineering and Technology
[35] Premalatha, K and Natarajan, AM(2008) A New Approach for Data Clustering Based on PSO with Local Search. International Journal of Computer and Information Science , vol 1, No. 4, 139
[36] Xiao, X; Dow,ER; Eberhart, R; Miled , ZB ;Oppelt, RJ (2003). Gene clustering using self-organizing maps and particle swarm optimization. Proceedings of International Symposium on Parallel and Distributed Processing.
[37] Abdul Latiff, N.M.; Tsimenidis, C.C.; Sharif, B.S.; Ladha, C.(2008). Dynamic clustering using binarymulti-objective Particle Swarm Optimization for wireless sensor networks. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposiumpp 1 - 5
[38] Rana,S; Jasola,S; Kumar, R(2010). A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm.International Journal of Engineering, Science and Technology.Vol. 2, No. 6, pp.167-176
[39] Olesen,J.R.; Cordero H.,J; Zeng, Y(2009). Auto- Clustering Using Particle Swarm Optimization and Bacterial Foraging.Lecture Notes in Computer Science, LNCS 5680, pp. 69–83
[40] Hyma, J; Jhansi, Y; Anuradha, S(2010). A new hybridized approach of PSO & GA for document clustering. International Journal of Engineering Science and Technology Vol. 2(5), 1221-1226
[41] Premalatha, K and Natarajan, AM(2010). Hybrid PSO and GA Models for Document Clustering. Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 3,
[42] Cui, X ; Potok, TE, (2005), Document Clustering Analysis Based on Hybrid PSO+Kmeans Algorithm, Journal of Computer Sciences (Special Issue), ISSN1549-3636, pp. 27-33.
[43] Hwang, J.-I.G.; Huang, C-J.(2010) Evolutionary dynamic particle swarm optimization for data clustering. In: International Conference on Machine Learning and Cybernetics (ICMLC) *International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, March 2013*
[44] Das S, Abraham A, Konar A (2008) Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. Pattern Recognit Lett 29:688–699
[45] Fun, Y. and Chen, C. Y.(2005). Alternative KPSO- clustering algorithm. Tamkang J. Sci. Eng., 8, 165–174.
[46] Sridevi.U. K. and Nagaveni. N.(2011) Semantically Enhanced Document Clustering Based on PSO Algorithm. European Journal of Scientific Research Vol.57 No.3 (2011), pp.485-493
[47] Johnson Ryan, K; Sachin, Ferat (2009) Particle swarm optimization methods for data clustering. In: IEEE fifth international conference soft computing with words and perceptions in system analysis, decision and control. Pp 1-6