Survey on Swarm Search Feature Selection for Big Data Stream Mining

S. Meera
Asst Professor
Dept of Computer Science (PG)
PSGR Krishnammal College for Women
Coimbatore

B. Rosiline Jeetha, PhD
HOD, Professor
Dept of Computer Science
Dr. N.G.P College of Arts and Science
Coimbatore

ABSTRACT
Now days, there are more number of corporations are gathering a more number of information, frequently produced incessantly as a series of measures and approaching from different types of positions. Big data defines a knowledge used to record and execute the data set and it has the structured, semi structured and unstructured data that has to be mined for valuable data. On the other hand, mining through the high dimensional data the search space from which an optimal feature subset is determined and it is enhanced in size, guiding to a difficult stipulate in computation. With respect to handle the troubles, the research work is generally based on the high-dimensionality and streaming structure of data feeds in big data, a new inconsequential feature selection methodology that can be used to identify the feature selection methods in the big data. Some of the research work illustrates the different kinds of optimization methods for data stream mining would lead to tremendous changes in big data. This research work is focused on discussing various research methods that focus on finding the efficient feature selection methods which is used to avoid main challenges and produce optimal solutions. The previous methods are described with their advantages and disadvantages, consequently that the additional research works can be focused more. The tentative experiments were on the entire research works in Mat lab simulation surroundings and it is differentiated with everyone to identify the good methodologies beneath the different performance measures.

Keywords
Big Data, Feature Selection, Particle Swarm Optimization, Classification

1. INTRODUCTION
One major challenge facing researchers working with big data is high dimensionality, which occurs when a dataset has a large number of features (independent attributes). To resolve this problem, researchers often utilize a feature selection process to identify and remove features which are irrelevant (not useful in classifying the data) and redundant (provide the same information as other features). Selecting a subset of features that are most relevant to the class attribute is a necessary step to create a more meaningful and usable model [1]. Data streams are possibly a large amount in size and therefore it is unfeasible to execute more number of data mining methodologies. Classification methodologies are unsuccessful execution data streams for the reason that of the two factors: their irritable volume and their distinctive feature known as theory drift [2].

There are 3V challenges are available in this method. The velocity troubles that provide the large number of information to be tackled at a growing high speed. There is more number of troubles creates the data processing and integration complexity. In the view of the3V challenges, the conventional data mining methodologies which are based on the entire batch-mode learning may execute in minimum time on the suggestion of analytic efficiency. It means, the conventional data mining model construction methodologies need loading in the entire set of data, and after that the data are divided corresponding to some divide-and-conquer methodologies. There are two classical procedures are namely CART decision tree induction and Rough-set discrimination [4].

The efficient feature ranking and selection procedures [5] can lead us in decreasing the size of the data set by removing the features that are irrelevant, unimportant, or helpless. Consider some bio information or medical information data steps, for instance, the number of features can achieve the tens of thousands. This is partially owing to the explanation that more number of data sets created at present for proposed data mining purposes, without prior knowledge about what is to be particularly explored or determined from the data, it is probable have included measurable parameters that are essentially unimportant or inappropriate and certainly resulting in huge amount of helpless parameters that can be removed to significantly decrease the size of the datasets without any negative subsequence in data analytics or data mining.

High dimensional data in a real life problem poses a big challenge for data analysis in real time [6]. Feature subset selection helps the learning algorithm to perform efficiently by removing irrelevant and redundant information in the data. Feature subset selection requires an efficient measure for evaluation of feature set and an optimal search strategy for finding out the best feature subset from a large number of candidates. Although a huge amount of procedures of feature subset selection with different group formation of the estimation calculations with search methodologies are present. Still now the research is going on with respect to identify the good procedures with lower price. In this research work, a difficult fuzzy consistency based calculation has been implemented for the estimation of feature set. The calculation is grouped with genetic algorithm (GA) and particle swarm optimization (PSO) both respect to class of evolutionary computation, for the model of optimal feature subset selection procedures.

Feature selection is assumed as one of the important basic troubles in the region of machine learning. The basic objective of feature selection is to retain the discriminatory information needed for the recognition process while discarding irrelevant and inconsistent information to minimize the computational burden of the pattern classifier. There are two types of aspects are available in feature subset selection execution, the estimation of the better feature or subset of the feature and

analysis for the optimal feature subset from a huge amount of individuals all the way through the entire feature space [7].

Thus feature selection becomes a most prominent role in the big data stream mining environment. The high dimensionality and the selection of the feature is a important difficult troubles which is examined and estimated by the different types of researchers using various methodologies. In this proposed work, the research methods are described with respect to the algorithm and their specified process by the side of the different types of performance calculations. The advantage and disadvantages that occur in those methods also described.

2. EXISTING RESEARCH METHODOLOGIES

The Ensemble of Decision Tree Classifiers for Mining Web Data Streams is proposed by the Tani et.al (2012) [8]. This effective mining methodology is specifically used to acquire a correct set of procedure for the extracting technology from the huge number of web data streams. The system create a web server using Model 2 Structural Design to gather the web data streams and assigned the ensemble classifier for producing the decision rules using more number of tree learning models. The proposed system produced three training data sets with the unique number of instance using web data stream, and create the three decision tree designs using every training data sets. For categorizing the test or unobserved instance: counts the weighted votes for every decision tree and apply the class with the highest weighted vote for that example. These ensemble methodologies can enhance the classification rate of web data streams.

A Task Graph of Stream Mining Algorithms is modeled by Akioka et.al (2013) [9]. This system implements a task graphs produced from the real development of stream mining procedures with respect to give to a implement of efficient, and sensible scheduling procedures for stream mining procedures. The important involvement of this system is 1) the first one is the task graph for stream mining procedures, 2) the realistic and sensible workloads taken from the previous implementations, 3) task graphs as representations of the characteristics of stream mining procedures to open up unexplored troubles for traditional scheduling procedures, and 4) task graph as a benchmarking tool to accelerate the implementation of scheduling procedures for stream mining procedures.

A towards Scalable and Accurate Online Feature Selection method for Big Data is proposed by Yu et. al (2014) [10]. The proposed methodology tackles the challenges in online feature selection from extremely high dimensional data, and develops SAOLA, a Scalable and Accurate OnLine Approach for feature selection. Comparing to Fast-OSFS, SAOLA employs a k-greedy search strategy to filter out redundant features by checking feature subsets for each feature. More specifically, to process each new feature efficiently, the system has theoretical analysis to derive a low bound on pair wise correlations between features so that the system can filter out redundant features.

A Context Adaptive Big Data Stream Mining methodology is performed by Tekin et.al (2014) [11]. There is more number of realistic methods, the distribution through the data and labels may not know priory in changeable ways over time. The proposed system considers data streams that are characterized by their context information which can be used as meta-data to select which classifier should be used to create a particular prediction. The context adaptive learning procedures modeled learns online what are the best context, learner and classifier to use to execute a data stream. The learning structure is introduced which accomplished a sub linear lament with time order independent of the dimension of the context vector. Hence it can converge very fast to the marginally optimal benchmark, by learning the marginal accuracies of classifiers for each type of context.

Ruta et al (2014) [12] presented robust and generic feature selection method that appears to be particularly suitable for handling very large data sets of high-dimensional sparse features. The method employs highly diversified backward-forward search that is relatively fast yet allows to achieve deep features complementarity and very high and stable predictive performance. The system has drafted the journey leading to the development of the top model and included informative and comparative examples of other feature selection models some of which could be good candidates for specific predictive requirements. Specifically it has shown that greedy forward search could be a very good model for a very limited number of features. The feature selection method is optimized with Naive Bayes classifier.

A distributed Adaptive Model Procedures for Mining Big Data Streams are proposed by Vu et. al (2014). A distributed streaming procedure is initially proposed to learn the decision rules for regression tasks. The procedure is present in Scalable Advanced Massive Online Analysis (SAMOA), an open-source platform for mining big data streams. This process is mainly used a vertical and horizontal hybrid parallelism to share the Adaptive |Model Rules (AMRules) on a cluster. The decision rules created by the AMRules are understandable models, where the predecessor of a rule is a conjunction of conditions of the parameter values, and the subsequent is a linear combination of the parameter. A huge advantage of decision rules is comprehensibility, required in many business decision making applications. The system begin by pipelining the processing of each instance into two steps: the training and predicting and applying these process to the listener and design aggregator processors in AMR. This methodology has confirmed to enhance the throughput for the difficult data sets.

Fong et.al (2014) [14] implemented a new Scalable Data Stream Mining approach in big data. In this methodology also known as Stream based Holistic Analytics and Reasoning in Parallel (SHARP) is introduced. This methodologies working procedure is based on the incremental learning and it span across a characteristic data mining model process from the lightweight feature selection and the one pass incremental decision tree induction, and also the swarm optimization process. The entire elements in SHARP is modeled to execute simultaneously achieving at enhancing the classification or prediction performance to its best possible. The fine-tune process is used by the swarm optimizer and its attributes values including the chosen the optimal feature subset regularly. SHARP is scalable, that depends on the accessible computing resources for the period of the execution time, the elements can process in parallel, cooperatively improving the various aspect of the entire SHARP process for the mining data streams.

The ACO Swarm Search based Feature Selection for Data Stream Mining in Big Data is introduced by Harde et al (32015) [15]. An optimal feature subset which is determined by the mining through the high dimensional data search space grows exponentially in size which guides to an intractable suggestion in computation. with respect to overcome to these troubles, this process is executing based on the high dimensional and streaming structure in data feeds in big data,

the light weight feature selection is indicated which will specifically focused on the mining data on fly, by using ant colony optimization (ACO) kinds of swarm search which can accomplish the improved analytical correctness. The novel feature selection procedures will validate the big data from the high level of dimensionality and streaming structure.

The initialization method for Particle Swarm Optimization based FCM is introduced by Wang et .al (2015) [16] in Big Biomedical Data. The feature of random initialization in Particle Swarm Optimization (PSO) based Fuzzy c-means (FCM) methodology affects the computational performance specifically in big data. As the data points in high density regions are more adjacent cluster centroids, the system modeled novel procedures to lead the initialization corresponding to the data density patterns. The procedures are initialized by fusing the data features adjacent the cluster centers. More importantly, our approach achieved more obvious computational efficiency on bigger data.

Fong et.al (2016) [17] introduced a new feature selection procedures for Data Stream Mining Big Data. The Feature selection has been generally used to reduce the processing load in inducing a data mining model. On the other hand, while it comes to mining through the high dimensional data the search space from which an optimal feature subset is determined grows exponentially in size, guiding to an difficult request in computation with respect to handle this troubles and this troubles is based on the high dimensionality and streaming structure of data in Big Data, a new lightweight feature selection in introduced. The feature selection is modeled particularly for mining streaming data on the fly, by using the Accelerated Particle Swarm Optimization (APSO) kinds of swarm search that accomplishes the improved analytical correctness within the reasonable processing time. This methodology also fits a good manner with current applications anywhere their data reach in streams.

3. INFERENCE FROM THE EXISTING SYSTEM

With respect to simulate the adverse effects of Big Data in order to high dimensions (more number of features) and a huge amount of volume (more number of examples), five representative data sets from the different regions are downloaded from UCI archive 1 for experimentation namely “arcene”, “dexter”, “dorothea”, “gisette” and “madelon”. The lengthy sequence of repeated input variables are known as data set “arcane” from the mass spectrometric data and it is taken from cancer patients. The more number of numbers is known as “dexter” and it is represent the entire text words and it is generally called as bag of word. The drug discovery is known as “dorothea” and it has an particularly a huge amount of parameters, 100,000 in total. The set of digitized information is called as data set “gisetter” and it is present in a 2D matrix format and it displays either a digit 4 or 9. The synthetically produced the data set “madelon” and it is present a set of numeric data points clustered in 32 combinations and it is placed on the vertices of a 5D hypercube, and they are randomly tagged which values of positive 1 or negative 1. There are five dimensions has the informative features, which guide to 15 linear group of features.

The previous Task Graphs of Stream Mining Algorithms, SAOLA methodology, context adaptive learning procedures, Diverfied backward-forward search and Naïve Bayes classifier, and ACO methodologies are mainly used for mining data stream from big data. But these methodologies are trouble with difficult request in computation process while it is used to mining through the high dimensional data, and also it is difficult to model a distributed decision rules is based on the settings of multi target learning and structured learning. Therefore, it decreased the entire system performance based on the exactness, accuracy and recollect. In this literature, the Decision Tree Classifiers, SHARP methodology, APSO approach, PSO-FCM and Distribute AM Rules are estimated with respect to the exactness, accuracy and recollect parameters. The new APSO method is modeled specifically for the mining streaming data, which is mainly based on the high dimensionality. The combinatorial explosion is mainly concentrated by using the swarm search methodology assigned in incremental method. Compared to the other previous methodologies, the new system is accomplished an excellent performance for the duration of using the high dimensional feature data set.

4. COMPARISON ANALYSIS

S.No	Reference	Method	Merits	Demerits	Results
1	Tani et.al [8] , (2012)	Decision Tree Classifiers	This method performs a god decision making and calculating the value of the class value of the web data streams	The information on the web and the structure of the mining troubles are not considered	The ensemble methodology can enhance the classification rate of web data streams up to 98%
2	Akioka et.al [9], (2013)	Task Graphs of Stream Mining Algorithms	The validation of scheduling procedures is heavily impacted by the task graphs.	For the computational model some other good methods are required	The task graphs are proposed to represent actually a various features, and dependencies differentiated to the data intensive applications in HPC.
3	Yu et.al [10], (2014)	SAOLA approach	It is mainly used for the high dimensional data set.	It does not accomplish a improved analytical correctness with the proper mentioned processing time.	SAOLA is scalable on data sets of extremely high dimensionality, and has superior performance over the state-of the- art feature selection methods.

4	Tekin et.al [11], (2014)	context-adaptive learning algorithm	It accomplishes a sub linear regret with time sequence independent of the dimension of the context vector.	To create the exact prediction process the good classifier is required.	The outcome of this process demonstrates that the new structural design is enhance the performance of Big Data Systems
5	Ruta et al [12] , (2014)	Diversified backward-forward search and Naive Bayes classifier	It is simple and robust feature selection method. This system achieved deep features complementarity and very high and stable predictive performance.	Better classifier is needed to optimize the feature selection.	The scored the second location with the predictive value is 96%
6	Vu et.al [13], (2014)	Distribute Adaptive Model Rules (AMRules)	It has proved to increase the throughput for "complex" datasets. A huge advantage of decision rules is comprehensibility, required in many business decision making applications.	The distributed decision rules is too complex and it is challenge the more settings of multi target learning and structured learning	On a small commodity Samza cluster of 9 nodes, it can tackle a rate if more than 30000 instances per second, and accomplishes a speed up to 4:7x through the series versions.
7	Fong et.al [14] , (2014)	SHARP method	It improves the classification/prediction performance. SHARP is scalable	It does not integrate and test all the component to produce the best possible performance	The experimental results achieve better accuracy and reduced tree size.
8	Harde et.al [15] , (2015)	Ant colony optimization (ACO) method	The high scale of data dimensionality is handled by this method `	While it reaches to mining through the high dimensional data and the search space form the optimal feature subset is determined and it increased exponentially in size, guiding to a difficult computation process.	The ACO kinds of swarm search and it accomplish the improved analytical correctness.
9	Wang et.al (2015) [16]	Particle Swarm Optimization (PSO) based Fuzzy c-means (FCM) methods	The designed methodology can enhance the computational performance of PSO based Fuzzy clustering methodologies, for the duration of preserving the similar performance of the clustering method	Better mechanism needed for further improvement of computational efficiency with less computational time	The results show that the proposed method achieves better clustering accuracy and computing time.
10	Fong et.al [17] , (2016)	Accelerated Particle Swarm Optimization (APSO)	The improved analytical accuracy within reasonable processing time is accomplished	For the data stream mining process the better feature selection methodology is required For the optimize the feature selection the good classifier method is required	The proposed system achieves accuracy of 97.8 and F-measure of 0.978

4.1 Accuracy

The exact positive and the negatives total is described as the accuracy and it partitioned by the total number of classification attributes ($T_p + T_n + F_p + F_n$)

$$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}}$$

Where,

Tp -True positive

Tn – Ture negative

Fp -False positive

Fn – False negative

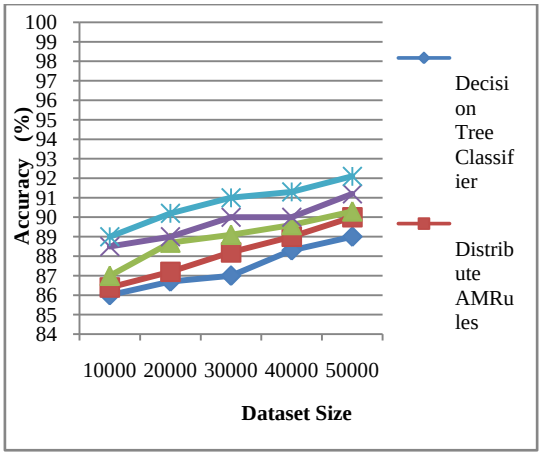

Fig1.Accuracy comparison

The differentiation of the Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO with respect to the accuracy is demonstrated in figure 1. The size of dataset is taken as X axis and in y axis accuracy is taken. For dataset size 50000, Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and APSO achieves accuracy result of 89%, 90%, 90.3 %, 91.2 % and 92.1%. Finally, the APSO approach reaches the high accuracy for the entire size of the data set.

4.2 Precision

The proportion of exact positives in opposition to both the exact positive and inaccurate positives results for intrusion and the real characteristics is described as Precision. It is described as follows

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

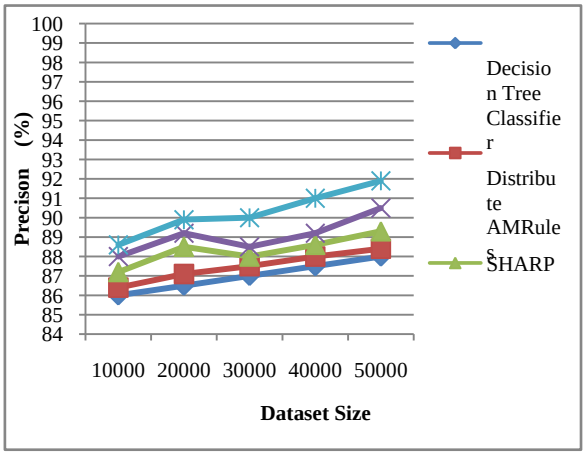

Fig 2.Precision comparison

The differentiation of the Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO with respect to the precision is demonstrated in figure 2. In X axis the size of the data set is represented and the precision is represent in the Y axis. For dataset size is 50000 of Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO accomplishes a precision outcome of 88%, 88.4%, 89.3 %, 90.5 % and 91.9 % correspondingly. From the graph it has been find out the APSO methodology outperforms than that of the other designs and results in precision values.

4.3 Recall

It measures the proportion of positives that are correctly identified

$$\text{Recall} = \frac{T_p}{T_p + F_n}$$

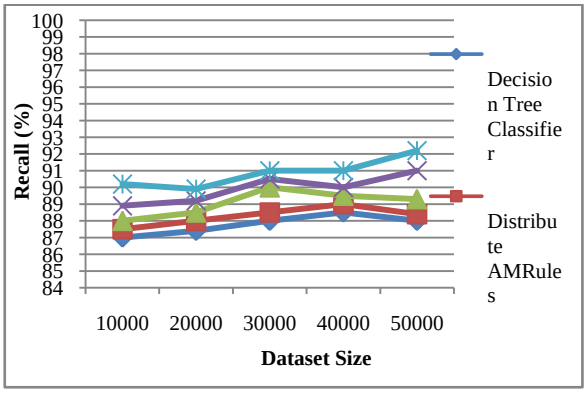

Fig3. Recall comparison

The differentiation of the Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO with respect to the recall is demonstrated in figure 3. For dataset size is 50000 of Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO accomplishes a recall outcome of 88%, 88.4%, 89.3 %, 91 % and 92.2 % correspondingly. Finally, that the APSO methodology has demonstrated the high recall value for the entire size of dataset.

4.4 Overall accuracy, precision and recall comparison

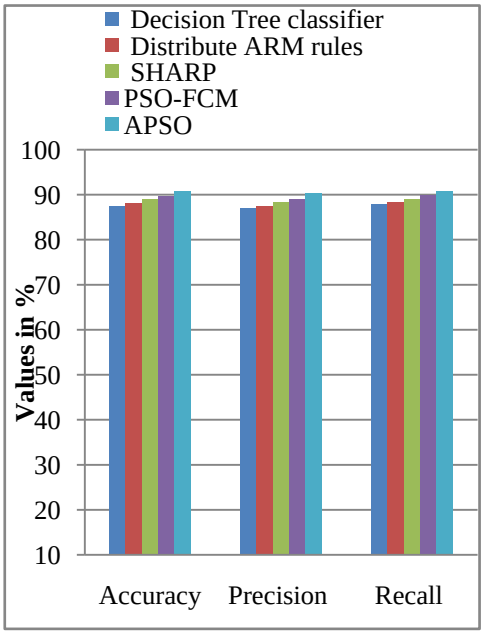

Fig 4. Performance Comparison

The new APSO methodology and the previous Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM are estimated with respect to the accuracy, precision and recall.

The differentiation of the Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM and proposed APSO with respect to the accuracy, precision and recall is demonstrated in figure 4.

According to this result, the proposed APSO methodology has higher performance outcomes while differentiated with the other previous Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM methods. This proposed search methodologies search the space of an aim of the process by the adjusting the entire agents. With respect to the search process, a speed up is developed in the new systems by incorporating a speed up in the initialization step in the Swarm Search. Therefore, it is mainly used to speedup the entire process of the system. at this point, it is focused that, the new APSO methodology, the exact result is acquired for the entire data set size like 10000, 20000,30000,40000,50000 is 90.72%,which is 3.1%,2.1%, 1.8% and 0.9 % higher than existing Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM approaches respectively. The proposed APSO approach achieves precision as 90.28% which is 3.9%, 3.5 %, 2.6 %, and 1.4 % higher than existing Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM approaches respectively. The APSO approach achieves recall as 90.88% which is 4.2%, 3.8 %, 2.9%, and 1.2% higher than existing Decision Tree Classifiers, Distribute AMRules, SHARP method, PSO-FCM approaches respectively. Finally, the APSO methodology is efficiently mine the streaming data from this graph.

5. CONCLUSION

The high dimensionality and the streaming nature of the input data aggravate the computational process in data mining in Big Data Analytics process. Big Data grows continually with fresh data are being generated at all times; hence it requires a computation approach which is able to monitor large scale of data dynamically .This research work analyses the various data stream mining techniques and risk factor analysis methodologies by different authors has been discussed. Those research methodologies are discussed along with their benefits and drawbacks in the detailed manner to find the effectiveness of every algorithm. The research works has been compared with each other based on their resultant metrics to find the better approach. The APSO is designed particularly for mining streaming data, which is mainly based on the high-dimensionality. The combinatorial explosion is addressed by using swarm search approach applied in incremental manner. The APSO based system achieves better performance compared to other existing system while using high dimensional feature dataset. In future, Min-max normalization and Acceleration-Artificial Bee Colony Optimization (AABC) algorithms are used for reduce the high dimensional dataset and effective feature selection respectively.

6. REFERENCES

[1] Alelyani, S., Zhao, Z and Liu, H., 2011. “A dilemma in assessing stability of feature selection algorithms”, in IEEE 13th International Conference on High Performance Computing and Communications (HPCC), 701–707.

[2] Minku, L.L., White A.P and X. Yao, 2010. “The impact of Diversity on online ensemble learning in the presence of concept drift”, 22(5):730–742.

[3] Fong and Simon, 2014. “A Scalable data stream mining methodology: stream-based holistic analytics and reasoning in parallel”, Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium.

[4] Ping-Feng Pai and Tai-Chi Chen, 2009. Rough set theory with discriminant analysis in analyzing electricity loads", Expert Systems with Applications 36:8799–880.

[5] Guyon, I and Elisseeff, A., 2003. “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, 3: 1157- 1182.

[6] Chakraborty and Basabi, 2014. “Rough fuzzy consistency measure with evolutionary algorithm for attribute reduction”, 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

[7] Bishop, C.M., 2006. Pattern Recognition and Machine Learning, Springer.

[8] Tani, Fauzia Yasmeen, Dewan Md Farid and Mohammad Zahidur Rahman, 2012. “Ensemble of Decision Tree Classifiers for Mining Web Data Streams”, International Journal of Applied Information Systems, 30-36 .

[9] Akioka and Sayaka, 2013. “Task Graphs of Stream Mining Algorithms”.

[10] Yu, Kui, 2014. “Towards scalable and accurate online feature selection for big data”, 2014 IEEE International Conference on Data Mining.

[11] Tekin, Cem, Luca Canzian and Mihaela Van Der Schaar, 2014. “Context-adaptive big data stream mining”, Communication, Control, and Computing (Allerton).

[12] Ruta and Dymitr, 2014, “Robust method of sparse feature selection for multi-label classification with Naive Bayes”, Computer Science and Information Systems (FedCSIS).

[13] Vu and Anh Thu, 2014. “Distributed adaptive model rules for mining big data streams”, Big Data (Big Data).

[14] Fong and Simon, 2014. "A Scalable data stream mining methodology: stream-based holistic analytics and reasoning in parallel”, Computational and Business Intelligence (ISCBI).

[15] Shivani Harde and Vaishali Sahare, 2015. “ACO Swarm Search Feature Selection for Data stream Mining in Big Data”, International Journal of Innovative Research in Computer and Communication Engineering, 3(12).

[16] Wang and Chanpaul, J., 2015. “A novel initialization method for particle swarm optimization-based FCM in big biomedical data”.

[17] Fong, Simon, Raymond Wong and Athanasios V. Vasilakos, 2016. “Accelerated PSO swarm search feature selection for data stream mining big data”, IEEE Transactions on Services Computing, 33-45.