ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Pervasive Utilization of PSO Techniques Over Twitter Data Streams.

Senthil Kumar N*, and Kauser Ahmed P.

School of Information Technology and Engineering, School of Computer Science and Engineering, VIT University, Vellore-632014, India.

ABSTRACT

Since the rapid growth of social media content has promulgated in digital space, the amount of data that has been generated via social media got multiplied every day. Moreover, the social media contents are very complicated and loosely coupled for integration. Besides it paves the tough path for capturing the data and analyzing the data for future decision making process. In order to facilitate this difficult operation, we have proposed here the clustering technique to effectively cluster the data in lesser amount of time. This paper implements the PSO algorithm with the modification required to match the Twitter data streams. The outcomes clearly depict the performance of the PSO algorithms and shows that there is an increase in number of particles in the larger selection of data streams. The proposed work has taken two different approaches to cluster the Twitter data streams using the modified PSO algorithm. First, it has used to select the centroids based on the number of clusters given by the user. Second, refine the cluster for giving the high convergence to the selected data sets.

Keywords: PSO algorithm, Twitter Streams, Feature Extraction, Clustering, Centroids.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

It has been observed that clustering the tweets is the crucial task of grouping identical classes of objects. The distance measured between the two objects in the cluster is always very less and indistinguishable. The performance of the clustering is ranked based on the similarity score and the intra-cluster similarity score should be very high and the inter-cluster similarity score must be low. According to the author [7], the scalability of good clustering technique is based upon the estimations of comparability. By the means of this clustering techniques, various sorts of uses like web search tools and programs utilized are promoted that clustering is an effective device for seeking .This clustering strategies gives great and complete standpoint of the report's data .Specific issue in information clustering where the different strategies for report grouping works: they are report's high dimensionality ,dataset of colossal size , capacity to comprehend the document clustering .There is additionally popularity of various clustering techniques leveled with the goal that report ought to be sought quick which is based upon the client's need and subjects of expanding specificity. In the paper [3], the author were discussed that PSO is a technique which tackle an issue by iteratively redesigning a hopeful arrangement .In PSO, the calculation issue is improved by number of competitor arrangements .Here various particles are utilized for finding the proper arrangement where each segments in the clustering have some speed and position .To discover the development of particles in the hunt space there is some particular numerical formulae to discover the speed and position of .Particles nearby best known position impact every particles development however, in the inquiry space it is likewise guided toward the best known positions, when better positions are found by different particles then this best position of molecule is overhauled . By upgrading the position swarm move toward the best arrangements

The pertinent need to cluster the tweets:

In the social space, users are permitted to post their thoughts and views on the social media sites and the messages posted on the Twitter [1] are termed as tweets. Tweets may incorporate recordings, photographs and 140 characters of content and connection and so forth. Tweets may connected with occasions like movie ,music , individual perspectives and thinking's .Recently tweets accomplished part of centrality on account of their ability to spread data quickly. The greater part of the acclaimed web indexes include these twitter messages in their query items since expansive no of tweets are including every day and contain valuable data. So breaking down small scale blogging framework is the zone where the majority of the scientist works. This short message gives next to no data about the particular subject. Tweets incorporates short structures, netslangs, abbreviations, grammer botches ,sections of sentences and so forth .So it is intricate to remove the data from tweets due to their casual style clustering of twitter dataset classify the tweets based upon the substance . By utilizing clustering it is anything but difficult to recognize the specific occasion or theme about which tweet is composed .To register the closeness of content words for a particular occasion information is spoken to by vector space model utilizing term recurrence and opposite term recurrence. So Particle swarm improvement system is the best technique to tackle bunching issue [5].

RELATED WORKS

The internet is generating enormous amount of short text data from social media web sites that need to be analyzed for getting fast and high quality tweets for the given Twitter data streams. Clustering of tweets data will help us to improve the retrieval process. The goal of tweet clustering is to partition the tweets into clusters according to their similarities of positive, negative and neutral tweets. Major steps of tweet mining include tweet preprocessing, feature extraction of tweets and clustering using Particle Swarm Optimization algorithm developed by [14]. The clustering problem is an optimization problem that locates the optimal centroids of the clusters. The drawback of K-means algorithm [18] is that it may end in local optima. Therefore PSO is the best choice since it is led to find an optimal or near optimal solution to a numerical and qualitative problem. The tweet clustering is evaluated using the average distance of tweet data to cluster centroid.

One of the important problems of tweet clustering is discussed by [13] using adaptive PSO technique and mapreduce. They implemented PSO algorithm for clustering twitter data using Hadoop's map-reduce framework. The outcome illustrates that parallel PSO performs very well compared to K-Means algorithm. Hence PSO is ideal choice for text clustering. Accuracy level of document mining can be improved by hybridization of swarm intelligence algorithms with machine learning algorithms. [12] proposed a SVM-PSO model for classification which shows that the accuracy level can be increased by hybridization process. [10]

introduced novel measures using particle swarm optimization techniques to improve the accuracy of term extraction results. Two steps are employed in their approach. First, terms are ranked to emphasize the most relevant from domain of input document; second, the score function is trained by the particle swarm optimization to obtain a suitable combination of feature weights. It achieves better precision than existing algorithms for text processing. [11] also analyzed SVM and PSO to classify the user opinions of tweets as positive, negative for the movie review dataset which could be used for better decisions. The work done in this research is only related to classification opinions into two classes, positive and negative class. In future work, a multiclass of sentiment classification such as positive, negative and neutral can be developed. [17] Developed two new approaches using PSO to cluster data. It is shown how PSO can be used to find the centroids of a user specified number of clusters. The algorithm is then extended to use K-means clustering to seed the initial swarm. This second algorithm basically uses PSO to refine the clusters formed by K-means. The new PSO algorithms are evaluated, and compared to the performance of K-means clustering. Results show that both PSO clustering techniques have more potential and have better convergence to lower quantization errors. [16] Provided a novel framework based on PSO-LA for text categorization systems that surpass previously introduced methods from three points of view: i) It works more efficiently over high dimensional datasets and its efficacy is not affected with the increase of features. ii) The PSO algorithm has fewer operators in comparison with other evolutionary algorithms and therefore the implementation is so simpler. iii) There is an information stream among the particles. The evaluation results indicate that the proposed method increased the accuracy of classifier in comparison with others. [8] presents an accurate and efficient algorithm for clustering Twitter streams. They break the clustering task into two distinctive stages: (1) batch clustering of user annotated data, and (2) online clustering of a stream of tweets. Results show that the algorithm presented is both accurate and efficient and can be easily used for large scale clustering of sparse messages.

Among many methods proposed for feature extraction, the PSO algorithms work more efficiently. These algorithms find the best solutions according to the knowledge obtained from previous iterations.

FEATURE EXTRACTION

Irrelevant and noisy data can be eliminated from original data set using preprocessing data mining techniques like feature selection. Based on the proposed work given in [6], it has been well argued that selection of best and relevant attributes or feature in large date sets can be done with the help of feature selection. Feature selection results in better classification of data and the reduction of processing time.

Feature Selection (FS) is the process of selecting a subset of relevant features from original data. FS can be done in two methods. Evaluation function and Generation procedure are used to evaluate and generate the candidate feature subset. When the evaluation function makes use of a classifier to evaluate the generated feature subsets, it is called as wrapper method. When a classifier is not involved and feature subsets are evaluated by looking into the intrinsic properties of data, it is known as Filter method.

Applications of Feature selection includes data classification, image classification, cluster analysis, data analysis, image retrieval, opinion mining, review analysis, etc. The messages on Twitter include reviews and opinions on certain topics such as movie, book, product, politic, and so on. Based on this reviews and opinion discussed in the research paper [2], this research attempts to use the messages of twitter to select relevant product using feature selection with the help of PSO. Feature extraction is done in two phases using wrapper method: In the first phase, twitter related data is extracted. The tweet is transformed into normal text in this phase. In the next phase, more features are extracted and added to feature vector. Each tweet in the training data is associated with class label. This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers [9]. So finally we get the tweets which are classified into the positive, negative and neutral. Wrapper method achieves superior results than filter methods. FS is seen as an optimization problem because obtaining optimal subset of relevant features from irrelevant and redundant data is very important. Many evolutionary algorithms have been used for optimizing the feature selection, which include genetic algorithms and swarm algorithms. Some of the swarm based optimization algorithms [15] for feature selection include, Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC).

Classification of Tweets:

Classifying the tweets by its domain topic is an all around examined issue. Regardless, ordering twitter messages by theme is troublesome in light of the fact that the messages are short and the components space for arrangement is exceptionally inadequate. We propose a technique to upgrade the content of the messages that contain joins with outside data, for example, the title of the website pages and with the most regular terms from these site pages. We demonstrate that the after effects of the characterization enhance considerably while including this outer data. At first, we utilize standard machine learning techniques in calculations from Weka [6] and straightforward elements to prepare classifiers. We tried different things with Support Vector Machines (SVM) since this calculation is known not great results on numerous orders assignments, with Naïve Bayes since it is known not well on content arrangement, and with Decision Trees (DT) in light of the fact that the model that is learnt is comprehensible. For a first examination, we utilized the words as a part of the preparation corpus as elements. This is known as a pack of-words representation. We disposed of the prevent words from the arrangement of components by utilizing the Python NLTK English stop words asset. There is a requirement for a calculation that can group the information in a lesser measure of time, in the event of information stream. Henceforth the need to utilize a parallel and appropriated environment utilizing map reduces system [13]. In like manner molecule swarm advancement methods are ideal for bunching issue, since it scales exceptionally well as information, measurements increment. K-Means clustering doesn't perform well with high dimensional information which is not the situation with PSO Clustering calculation. Thus it is perfect decision for content grouping. In future this bunching system can be utilized to perform ubiquity examination, assessment mining and nostalgic investigation and so forth.

Particle Swarm Optimization:

Swarm intelligence [10] is ability of such systems, to achieve a higher level of intelligence, which is absolutely unreachable for any of system units. For example, a flock of birds as a society has very complex behavior patterns, which is beyond the intelligence level of any of birds in the flock, of course. However, these complex patterns are created via simple and repetitive tasks, performed by any of members in the flock. Each particle will modify its current position and velocity according to the distance between its current position and pbest, and the distance between its current position and gbest.

Particle Swarm Optimization (PSO) is an intelligent optimization algorithm based on the Swarm Intelligence. It is based on a simple mathematical model, developed by Kennedy and Eberhart in 1995[14], to describe the social behavior of birds and fish. The model [12] relies mostly on the basic principles of self-organization which is used to describe the dynamics of complex systems. PSO utilizes a very simplified model of social behavior to solve the optimization problems, in a cooperative and intelligent framework. PSO is one of the most useful and famous metaheuristics and it is successfully applied to various optimization problems.

Various applications of particle swarm optimization include feature selection, image reduction, data mining, and cluster analysis.

PSO based tweets clustering:

The complete tweet data clustering process is categorized into four stages: a) Collection of twitter data b) Data preprocessing c) Feature extraction and d) Data clustering using PSO.

Collection of twitter data:

In first phase, Collection of twitter data is done with the help of Twitter API, we collect streaming tweets after every 1000ms. The Twitter API plays a very important role of enabling us to extract the definite category tweets as per our need. We are storing an individual tweet in the corresponding folder of its category, in a separate text file, we term as a document.

b) Tweet Preprocessing:

In Second phase, pre-processing of documents includes the following steps: (a) data cleansing and a corpus of data is created for cluster analysis.

ISSN: 0975-8585

c). Feature extraction:

In third phase, Tokenization is done to divide the tweets into words, phrases and symbols termed as tokens and Stemming is done to reduce the actual word to their base or root form. Stop word list is maintained to observe the common words in tweets.

d) Data clustering using PSO:

After pre-processing of tweets, we have applied PSO algorithm for clustering. The first step of PSO is initializing the number of particles. A particle is nothing but one of the possible solution for clustering the streaming tweets. Therefore, a swarm consists of collection of candidate clustering solutions of streaming tweets. Each particle is represented as X = (C1, C2, C3…Ci. Ck), where Ci represents the it cluster centroid vector and k is the number of clusters.

After initialization of the particles, for each particle, assign each tweet to its closest centroid vector. The fitness of each particle is computed by considering the average similarity between the cluster centroid and a tweet in the document vector space, belonging to that cluster using cosine correlation measure. Experimental results show that PSO clustering out performs over hierarchical and partitioning clustering techniques.

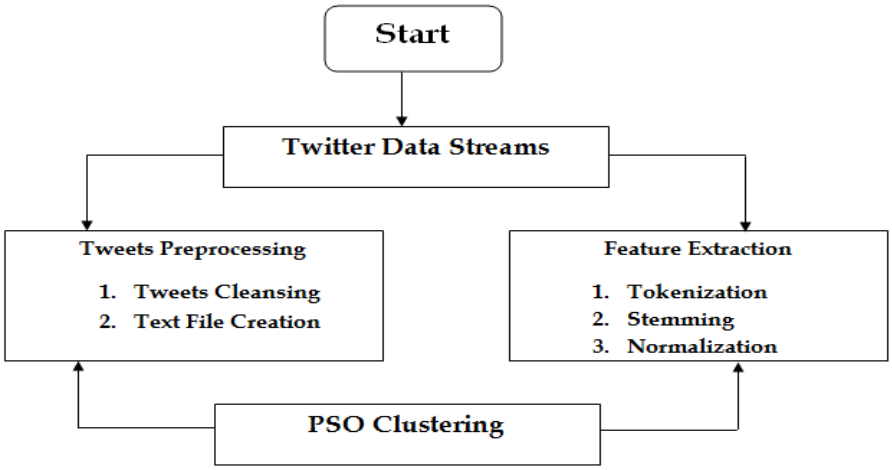

Figure 1: Proposed architecture for PSO clustering

Algorithm for PSO based tweets clustering:

Begin:
Step 1: Input tweet data rb_{IN}, & specify the size of population n, maximum iterations $tmax$ and t = 0.
Step 2: Initializing the position and velocity of each particle .
Step 3: For each particle $X\ ^{t}_{i}$, assign all observations in the data set to the nearest cluster which is measured by cosine correlation measure.
Step 4: Computing the fitness function values of X .
Step 5: $t = t + 1$.
Step 6: If $t < t_{max}$
Step 7: Update each particle's local best position p_i and the population's global best position p_g.
Step 8: Update each particle's velocity and position according to Eq.(1) and Eq.(2).
Step 9: Go to Step 3.
Step 10: End if.
Step 11: Out put the population's global best position p_g.

End

Empirical Analysis over Twitter Data Streams:

ISSN: 0975-8585

We have analyzed the results of PSO for the taken Twitter datasets and compared the quality of the clustering with respect to the following functional criteria:

In order to minimize the inter-cluster distances, we measured the distance path between the data vectors within the given cluster.
To maximize the cluster distance, we have evaluated the distance between the centroids of the cluster.

To the overall reports generated, we have undertaken almost 30 simulations and the proposed algorithm has taken 1000 functions to effectively evaluate the results. In particular, the PSO algorithm has been given the IO particles and the threshold values of PSO parameter will have U=0.68, cl = 1.33 and cz =1.45 respectively. Thus it has resulted in good convergence of parameter value selection and yields the expected results optimally. For the given Twitter datasets, when we take the inter or intra cluster distances for reference, then it has been observed that there is a trivial deviation in the cluster centroid. Hence, we have proposed the PSO algorithm to ensure that the compactness of the cluster must be intact and find the clusters which have larger separation in the cluster centroids.

```
Running Matlab Proposed PSO Version
  iter    phase       num              sum
     1        1       150          268.148
     2        1        29          150.133
     3        1        16          95.1727
     4        1         4          87.3855
     5        1         2          86.3902
Best total sum of distances = 86.3902
```

Table 1: Clustering Measure for the chosen datasets

The result has shown that the proposed algorithm is given the gradual improvement when the cluster has larger separations and the distance between the centroids has also been witnessed that it has improved with preciseness. The global fitness of the proposed PSO algorithm is 0.4827. In certain cases, when the quantization error is high, then the PSO algorithm has step down to slower convergence. In order to tackle this slower convergence rate for the Twitter datasets, we have included the optimal functional evaluations and standard PSO functions to overcome the glitches happened during the processing of datasets. It has been apparently seen in the Table 1 that the sum of distance for every phase of clustering is optimal and yield the results with good convergence.

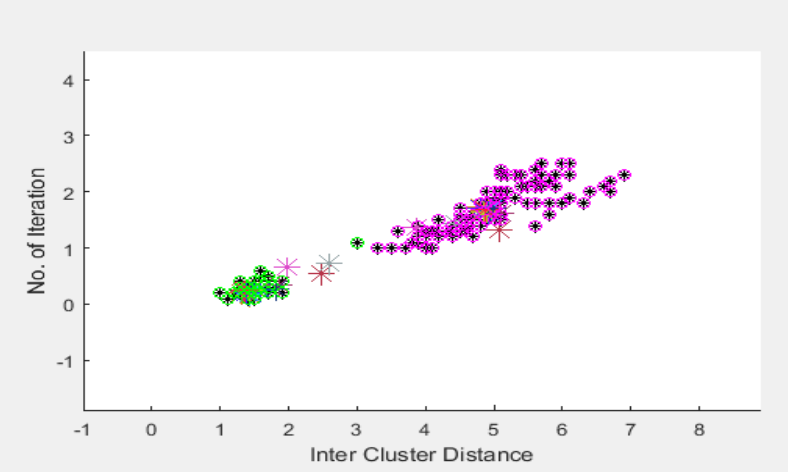

Figure 2: PSO based Clustering for Twitter Data Streams

Application of PSO:

The applications of particle swarm optimization spreads from science and engineering research to biological research [15]. Clustering analysis is a popular data analysis, data mining and knowledge acquisition technique generally used in engineering and biological research. The main purpose of clustering is to form groups of similar and dissimilar objects in separate classes based on the values of their attributes. It is observed that, swarm-based algorithms furnish fruitful results compared to conventional clustering analysis techniques. It is found that the hybridization of particle swarm optimization with other machine learning algorithms provides better cluster analysis. Various advantages of particle swarm optimization algorithm are listed below.

- The main advantage is particle swarm optimization can be applied to both scientific and engineering research and biological research.
- These algorithms are used to solve nonlinear optimization problems.
- The algorithm can easily be implemented since the global search of the algorithm is efficient.
- The algorithm runs faster since the dependency on the initial solution is smaller.
- Parameter selection and parameter tuning are easily done since the issues are already available in literature.
- It is easy to calculate and very simple.

Such algorithms have no overlapping and mutation. The implementation of such algorithms is easy since the velocity, position and memory are easily calculated.

CONCLUSION

The proposed work here investigated the PSO algorithm on Twitter data streams to effectively cluster the data sets and tested that it has yield the better convergence with lower quantization errors. In general, it has given the good convergence based on the two simpler principles that it has taken the larger inter-cluster distance for the Twitter streams and smaller intra-cluster distance for centroids in the Twitter data streams. As the Twitter data stream is very large in size, the proposed PSO algorithm has followed the global search mechanism to find the optimum results of the clustering. We have compared the proposed algorithm with other existing algorithms and found that this proposed work has given the optimum results with lesser time by taking less parameter for clustering.

REFERENCES

[1] M Omran. A P Engelbrecht and a Salman Particle Swarm Optimization Method for Image Clustering Int. J. Pattern Recognit Artif Intell 2005; 19: 297–321.
[2] K S Gaikwad and Dr. Manasi S Patwardhan. Tweets Clustering : Adaptive PSO. Annual IEEE India Conference (INDICON) 2014;
[3] S Liangtu and Z Xiaoming. Web Text Feature Extraction with Particle Swarm Optimization International Journal of Computer Science and Network Security.2007; 7: 132–136.
[4] H Patel and Nilesh Mali. Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data using Genetic Algorithm International Journal of Advance Engineering and Research.2015; 2: 367–370.
[5] T Nguyen. T Nguyen and Q Ha Applying Hidden Topics in Ranking Social Update Streams on Twitter IEEE RIVF International Conference on Computing & Communication Technologies -Research, Innovation, and Vision for the Future (RIVF)2013; 180–185.
[6] B Batrinca and P C Treleaven. Social media analytics: a survey of techniques, tools and platforms. Ai Soc.2014; 30: 89–116.
[7] R D W Perera. S Anand, KP Subbalakshmi and R Chandramouli, Twitter analytics: Architecture, tools and analysis. Military Communications conference.2010; 2186-2191
[8] O Tsur and A Rappoport. Efficient Clustering of Short Messages into General Domains Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.2013; 621–630.

[9] T L Wang and M Wang. Features extraction based on particle swarm optimization for high frequency financial data. IEEE International Conference on Granular Computing (GrC) 2011; 728-733
[10] M Syafrullah and N Salim. Improving Term Extraction Using Particle Swarm Optimization Techniques, Journal of Computing 2010; 2: 116–120
[11] K Umamaheswari. S P Rajamohana and G Aishwaryalakshmi. Opinion Mining using Hybrid Methods. International Conference on Innovations in Computing Techniques 2015; 18–21
[12] A S H Basari. B Hussin. I G P Ananta and J Zeniarja. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Eng.2013; 53: 453–462.
[13] A P Chunne. U Chandrasekhar and C Malhotra, Real time clustering of tweets using adaptive PSO technique and MapReduce Global Conference on Communication Technologies (GCCT).2015; 452-457
[14] James K and Russell E. Particle Swarm Optimization, IEEE International conference on Neural Networks. 1995; 1942-1948.
[15] D P Acharjya and Kauser Ahmed P. Swarm Intelligence in Solving Bio-Inspired Computing Problems - Reviews, Perspectives and Challenges. Handbook of Research on Swarm Intelligence in Engineering IGI Global Publishers, USA, 2015, p.74-98
[16] Mozhgan Rahimirad. Mohammad Mosleh and Amir Masoud Rahmani, Improving the Operation of Text Categorization Systems with Selecting Proper Features Based on PSO-LA. Journal of Advances in Computer Engineering and Technology, 2015; 1: 1-8
[17] DW van der Merwe and AP Engelhrecht Data Clustering using Particle Swarm Optimization, IEEE Congress on Evolutionary Computation 2003; 215-220
[18] Taher Niknam and Babak Amiri An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis Applied Soft Computing 2010;10: 183–197