Building Large Scale Cloud System for Product Sentiment Analysis using Hybrid Group Search Optimization Based Feature Selection

P. Vasudevan, K.P.Kaliyamurthie

Abstract: ***A very powerful technology that performs complex computing in a massive scale is known as Cloud computing. There has been a massive growth that has been observed in the data scale which may also be big data which is generated by means of cloud computing which is observed. Sentiment Analysis, on the other hand, denotes the opinion extraction of users from the documents used for review. A sentiment classification that makes use of methods of Machine Learning (ML) can face problems in high dimensionality for a feature vector. Thus, the method of feature selection is needed for the elimination of all noisy and irrelevant features from a feature vector for efficiently working the ML algorithms. All chosen features will be sub-optimal owing to a Non-Deterministic Polynomial (NP) hard type of technique that was used. The Group Search Optimization (GSO) based algorithm which was on the basis of a method of feature selection will find some optimal feature subsets through the elimination of all redundant features. For this work, the method of feature selection based on the GSO was applied to the sentiment classification. There was also a method of feature selection which was hybrid and based on the GSO and Local Beam Search (LBS) that has been proposed for a sentiment classification. The methods proposed were evaluated based on the product review dataset of Amazon. The results of the experiment proved that this method of a hybrid feature selection can outperform all other methods of feature selection for a sentiment classification.***

Keywords: ***Feature Selection, Group Search Optimization (GSO), Local Beam Search (LBS) and Support Vector Machine (SVM).***

I. INTRODUCTION

The Big data can manage the capacity of data, the speed which is besides an inconsistency by the streamlining which implements the redressing of fresh desires and this will refer to this as a big gauge information and its structural design. The cleaning procedures of deploying data that was on the data by means of traditional databases where data has been assumed as arranged in rows and also in columns at the time of the growth of the volumes of data over a time period and it normally lacks the ability to manage a big gauge data. The database or the warehousing systems used traditionally had been designed with a smaller capacity and an information assembly that had an expectable update aside from an unswerving functioning of information assembly that was on a lone server resulting in an expenditure on the operation that has an increase in the capacity of data [1].

Recently, the rapid growth and the popularity along with the technologies of storage along with the success of that of the Internet, the resources of computing are now cheaper and also more powerful and will also be ubiquitously available compared to the past. This trend of technology is called cloud computing and this has resulted in a manner which is evaluative for providing an answer to the current and the future Information and Communication Technology (ICT) needs. Cloud computing has given an adaptable environment online that encourages the capacity to be able to handle the work of an expanded volume without it affecting the execution of this framework. The advent of the Cloud has an increase in the number of providers with various services that are difficult for a researcher and also poses several challenges that need to be coped with. In the last few years, the researchers have been working all over the world for enabling the technology towards a wider opportunity of business and also in certain other areas in Information Technology (IT), using cloud computing-based services and their mechanism [2].

The Sentiment analysis has been identified to be a Natural Language Processing (NLP) based type for tracking the public mood as regards a particular topic or product. The sentiment analysis is also known as opinion mining and this involves the building of the system which was for the collection and the examining of opinions regarding the product in the comments, reviews, tweets or blog posts. This sentiment analysis can be used in several ways. For instance, it can judge the success of any ad campaign or a product launch in the marketing of the

Revised Manuscript Received on July 22, 2019.

P. Vasudevan, Research Scholar, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India.

Dr.K.P.Kaliyamurthie, Professor & Dean, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India

product that can determine which particular product or service is popular and also identify the likes or dislikes for specific features as according to the category of demographics. The first one was an opinion word which was taken to be positive in a particular situation and is negative in that of another. The next challenge was that the people may be able to express their opinions in different ways. A conventional method of text processing was based on the limited differences that are identified among two of the text pieces and this will not change its meaning [3].

The selection of features is made on the basis of certain specific measurements and aims at simplifying the concept training and further reducing the time taken for training. The feature has been taken to be relevant to of a particular class and in case its existence in a certain class which is quite high on being compared to the existence of this in yet another class. Text opinion mining has some critical tasks that need consideration and this includes the assigning of the sentiment polarities in an accurate manner (such as positive or negative or neutral) and the intensities (such as high or low) which is a challenge until now. There are four different steps in feature selection and they are: generation of subsets, evaluation of subsets, stopping criterion and validation of results. The feature selection has two different categories which are the filters and the wrappers. The former will rank the features using a criterion of valuation that retains only those features that have a value which is above a certain threshold. The latter will look out for some optimal subsets in a particular classifier. The performance measures were all attached to the subsets that were based on a performance using a particular classifier of learning data [5].

The traditional techniques of feature selection like the shared data, the data pick up and the chi-square is found to be diminishing the corpus but do have some problems as regards accuracy. Additionally, the choice of some irrelevant features can lead to a Non-Deterministic Polynomial (NP) hard type of a problem, that can also affect the system’s efficiency. This way, for solving the issues and for enhancing the execution of classification, the focus has been shifted to the systems of survey for scaling up and also for providing their outcomes. There were some intelligent algorithms which were fruitful in looking out for some good answers for some of these complex issues and also measuring the type of classification. The algorithm, for instance, the Nature-Inspired Algorithms (NIA), the Simulated Annealing (SA) and the Genetic Algorithms (GA) were investigated in an enhanced research classification [5].

At the time of a process of classification, the rate of error, the rate of misclassification and the issues related to accuracy were handled by means of an algorithm which was inspired by nature and a Swarm Intelligent Algorithm (SIA), a physics-chemistry algorithm and a bio-inspired algorithm. Additionally, the techniques of Particle Swarm Optimization (PSO) have been used in various problems that are complex like the optimization of power and a Traveling Salesman Problem (TSP) that result in better solutions as the PSO has been a process which was continuous and not well-suited to the problems which are combinatorial. The main drawbacks of the PSO are its outdated memory and loss of diversity. For overcoming this particular drawback, a hybrid GSO-LBS method of feature selection in its cloud system for a sentiment analysis was proposed. The remaining part of the investigation has been organized thus. All related work was discussed in Section 2. The various methods employed in this work was made in Section 3. Section 4 discussed all experimental results and the work was concluded in Section 5.

II RELATED WORKS

Quan and Ren [6] had introduced another novel term-based similarity measure which was a Point-wise Mutual Information (PMI)–Term Frequency-Inverse Document Frequency (TFIDF) for evaluating the connection of candidate features along with the domain entities. The results prove the approach to be outperforming all the other methods that were state-of-the-art. There were only external resources employed and were compared to the domain corpora thus making it unsupervised and generic. On being compared to the other traditional methods like the PMI and the PMI–TFIDF had shown a distinction ability which was better. The author further proposed an opinion that was feature-oriented based on the extraction of a feature-opinion pair and lexicon generation which was a feature-oriented opinion.

Tan et al., [7] had made a new presentation of another feature scaling scheme used for the feature selection that was of an ultrahigh dimension and later reformulated this to be a convex and problem of Semi-Infinite Programming (SIP). For addressing this SIP, the work had proposed another efficient method in the paradigm of feature generation. There were various approaches that were gradient-based conducting an optimization. This can solve an entire sequence of the sub-problems in Multiple Kernel Learning (MKL). For increasing the speed of the work, the MKL sub-problems which were in their primal forms by means of a proximal gradient approach was employed. Owing to this scheme of optimization, there were certain efficient cache techniques that were developed.

Yu et al., [8] had developed another Scalable and the Accurate on Line Approach (SAOLA) used for the feature selection. A theoretical analysis made on the low bound on correlations that were pair-wise among the features in their chosen feature subsets. There was an empirical work with a series of real datasets that showed the SAOLA was quite scalable on the datasets of a very high dimensionality with a superior performance compared to all methods that were state-of-the-art.

Agarwal and Mittal [9] had made a presentation of both unigrams and some bi-grams which were extracted from a text with certain composite features that were created by employing them. There was an extraction of adverbs and adjectives based on the Part of Speech (POS). Both Minimum Redundancy Maximum Relevancy (MRMR) and Information Gain (IG) methods of feature selection had been used for extraction of prominent features. The effects of all the categories of these features were investigated based on four different standard datasets such as a movie or a product review dataset. The results of the experiment proved that all composite features that were created from unigram and bi-gram features had performed better compared to all other features in the classification of sentiment.

Priyadarshini [10] had made a proposal of a Map Reduce SVM for a data of a large scale. This model works on different frameworks such as Hadoop Twister. Here in this work, the impact of the kernel and the penalty parameters were observed. The results of the experiment proved that there were several support vectors along with some predictive accuracy belonging to the SVM which was affected by these parameters. The results of the experiment further analysed the time taken for computation by an SVM along with a cluster that was multi-node was lower compared to a one node cluster for a large dataset.

Ahmad et al., [11] had made a presentation of other types of feature selection in the sentiment analysis which was based on both NLP and some modern methods like the Rough Set Theory (RST) or the GA. Feature selection has been a crucial step in a sentiment analysis owing to the fact a suitable selection of features were able to identify all the actual product features that were criticized or even discussed by the consumers.

There is a conclusion that all algorithms can have the potential of being implemented in the research of sentiment analysis which was produced to be an optimal feature subset by the elimination of features which are redundant or irrelevant.

Ahmad et al., [12] had proposed yet another hybrid Ant Colony Optimization (ACO) along with the K-Nearest Neighbour (KNN) based algorithms like the feature selections used for choosing all relevant features from the datasets of customer review. The Information Gain (IG), the GA, and finally the Rough Set Attribute Reduction (RSAR) had been employed to be the baseline ones in a performance evaluation and comparison. This process of evaluation has proven statistically that the ACO-KNN algorithm was improved to a significant level on being compared to the baseline algorithms. Additionally, the results of the experiment proved that an ACO-KNN was used as a technique of feature selection in a sentiment analysis for obtaining the quality and an optimal feature subset which represents actual data found in a customer review based data.

The data which was mobile-centric is significant and has been identified across many different applications of commercial services. But there is, however some uncertainty for the volume of the big data which solicits some appropriate analytics and ability of decision making that is inferred from these sources of data. The source of data and its analytics will be selected from the idea of an adaptive and intelligent technique.

Banerjee and Badr [13] had elaborated this solution at the time of deploying a rough set capable of being able to handle all uncertain and imprecise contexts of Big data. Additionally, the pheromone deposition of an ant colony and its evaporation process will assist in an optimal mechanism of selection. The model proposed was supported by hazard event case study propagated by means of a mobile data which was derived from the social network. Data was represented in the form of posts and social tweets. This has been analysed by means of an ant colony based on a rough set.

Alarifi et al., [5] had made an introduction of a new big data with a technique of machine learning for the evaluation of the process of sentiment analysis for overcoming the problem. All data collected for large dataset volumes are helpful in the system of effective analysis. The noise found in data was eliminated with a concept of data mining in pre-processing. From this type of a cleaned sentiment data, all effective features that had been selected using a greedy approach with another optimal classifier known as the Cat Swarm Optimization-based Long Short-Term Memory Neural Network (CSO-LSTMNN). All classifiers had analysed the features that were sentiment-related in accordance with the cat behaviour, the minimization of the rate of error and the technique also helps in improving the efficiency of the system and also analysed the results of the experiment, accuracy, recall, and precision.

III METHODOLOGY

For this section, the review of Amazon product dataset was used. A feature selection with the GSO and the GSO-LBS were proposed, and the SVM classifier was employed

A Dataset

The dataset also consists of some product reviews along with a metadata from Amazon which also had 142.8 million reviews that were spanning between May 1996 and July 2014. The dataset also includes the reviews (the ratings, helpfulness votes and the text), a product metadata (the image features, the brand, price, category information and their description), and the links (the viewed and the bought graphs). For this, a subset of an Amazon sentiment dataset (with a total of 45000 Positive, 40000 negative and around 35000 neutral) were employed.

B Feature Extraction Using Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF is a technique of information retrieval weighing the Term's Frequency (TF) with an Inverse Document Frequency (IDF). Every word or a term will have its own IDF and TF score. Their product scores for a term was referred to as the TF*IDF weight of the same term. This can also state that if the TF*IDF score (or weight) was higher, then the term was rarer. The TF for a word is its frequency. The IDF is the measure of the significance of the term which is throughout a corpus. If the words have a TF*IDF weight that is high in content, this content will be among the top in search results and so people can 1. Stop their worry regarding the usage of stop-words, 2. Identify words having higher volumes of search with a lower competition [14].

C Group Search Optimization (GSO) Based Feature Selection

The GSO is a metaheuristic inspired swarm intelligence technique by the social searching behaviour of animals and their theory of group living. The GSO further employed a model of Producer-Scrounger (PS) to be its framework [15]. In the GSO, the actual population G of the S individuals was known as the group and in a search space that was n-dimensional, its ith member at its t-th searching iteration will have its current position $\vec{X}_i^t \in \Re^n$ along with a head angle $\vec{\alpha}_i^t \in \Re^{n-1}$.This direction of the search of that of the i-th member being a vector $\vec{D}_i^t(\vec{\alpha}_i^t) = (d_{i1}^t, \ldots, d_{in}^t)$ is computed from the $\vec{\alpha}_i^t$ through a polar to a transformation of Cartesian coordinate as in (1):

$$d_{i1}^k = \prod_{q=1}^{n-1} \cos\left(\varphi_{iq}^k\right)$$

$$d_{ij}^k = \sin\left(\varphi_{i(j-1)}^k\right) \cdot \prod_{q=j}^{n-1} \cos\left(\varphi_{iq}^k\right) \quad \left(j = 2, \ldots, n-1\right)$$

$$d_{in}^k = \sin\left(\varphi_{i(n-1)}^k\right) \cdot \tag{1}$$

The group GSO contains three different members who are: the producers, the scroungers and the dispersed members (or the rangers). In a GSO search iteration, the member of a group with the best value of fitness until now (the most promising one) is selected to be the producer ($\vec{X}_p$), and all the other members will either be scroungers or the rangers. A producer will employ a strategy of scanning (producing) that is based on the vision field. In the GSO, at its t-th iteration a producer $\vec{X}_p^t$ scans in a lateral manner by means of a random sampling of three different points found in its scanning field: one was at the zero degree (2), one located in its hypercube on the right-hand side (3) and one on the hypercube on the left-hand side (4).

$$\vec{X}_z = \vec{X}_p^t + r_1 l_{\max} \vec{D}_p^t\left(\vec{\alpha}_p^t\right) \tag{2}$$

$$\vec{X}_r = \vec{X}_p^t + r_1 l_{\max} \vec{D}_p^t\left(\vec{\alpha}_p^t + \frac{\vec{r}_2 \theta \max}{2}\right) \tag{3}$$

$$\vec{X}_l = \vec{X}_p^t + r_1 l_{\max} \vec{D}_p^t\left(\vec{\alpha}_p^t - \frac{\vec{r}_2 \theta \max}{2}\right) \tag{4}$$

Where $r_1 \in \Re$ was distributed normally in a random number (with a mean 0 and a standard deviation as 1), $\vec{r}_2 \in \Re^{n-1}$ which is a random sequence uniformly distributed within a range of (0, 1), $\theta_{\max} \in \Re^{n-1}$ which is a maximum pursuit angle with the $l_{\max} \in \Re$ and is also a maximum pursuit distance given by (5):

$$l_{\max} = \left\|\vec{U} - \vec{V}\right\| = \sqrt{\sum_{k=1}^{n} (U_k - L_k)^2} \tag{5}$$

Wherein the U_k and L_k are the upper and the lower bounds for that of the k-th dimension.

In case the producer can find a resource which is better compared to its present position [16], it tends to fly to a point; in case there is no other point which is found, the producer stays in its present position and this turns its head to an angle which is newly generated (6).

$$\vec{\alpha}_p^{t+1} = \vec{\alpha}_p^t + \vec{r}_2 \beta_{\max} \tag{6}$$

Wherein $\beta_{\max} \in \Re$ will be its maximum turning angle. In case after the ∈ℵ iterations, a producer will not be able to find an area which is better and this turns and heads back to the zero degree as per (7):

$$\vec{\alpha}_p^{t+a} = \vec{\alpha}_p^t \tag{7}$$

All the scroungers join a resource which is found by a producer and a scrounging strategy is performed as per (8).

$$\vec{X}_i^{t+1} = \vec{X}_i^t + \vec{r}_3 \text{o} \left(\vec{X}_p^t - \vec{X}_i^k\right) \tag{8}$$

Wherein $\vec{r}_3 \in \Re^n$ denotes a new and uniform random sequence which is in a range (0, 1) and the ◦ is a Hadamard product and at times a Schur product, that computes an entry-wise product for two vectors.

Rangers will then perform certain random walks in the problem space (ranging) as per (9 and 10).

$$\vec{X}_i^{t+1} = \vec{X}_i^t + l_i \vec{D}_p^t\left(\vec{\alpha}_i^{t+1}\right) \tag{9}$$

Wherein,

$$l_i = a r_1 l_{\max} \tag{10}$$

At the time a member is able to escape from the bounds of its search space, it turns back to the earlier position within the search space. The GSO scrounging operator will focus a search performed by groups in some of the promising areas from their problem space when producing operators and ranging operators will be the primary mechanisms that were employed by the GSO for the purpose of escaping from the local minima.

The GSO will have to be extended for dealing with a feature selection. The position of the PS will be taken into consideration as the binary bit strings and each bit will represent a certain feature; there is a bit value 1 which represents a chosen feature and the bit value 0 will represent another feature that is non-selected. Every position in this will denote a feature subset.

D Proposed GSO-LBS Based Feature Selection

Every feature subset is then taken to be a pint within the feature space. An optimal point was the subset with its least length along with an accuracy of classification which was high. There was an initial swarm that was distributed in a random manner over the entire search space. The PS aims at flying to its best position.

They communicate with one another and change their position and finally search around for the local and the global best. Lastly, they will not converge on the good, the possibly optimal and the positions with an ability of exploration equipping them for performing the selection of features and also discover some subsets which are optimal.

The steps in a hybrid GSO-LBS are as below:

Step 1. Initialize randomly all members of a population and also generate their respective fitness values [17].

Step 2. When the terminal conditions have not been met.

Step 3. For all members i in the group do.

a. Finding all producers in the group.

b. Creating of a roulette wheel for the selection of producers from among a group of scroungers.

c. A Beam search [18]

1: Initialization: To set B = {n_0} and C =∅.

2: Branching: For every node in B:

i Branch this node and generate its corresponding children. ii Calculate a new upper bound based on a value of an optimal solution for every child node.

ii Select β as the most promising among child nodes and add then to C.

3: Selection of Node: Choose β as the promising nodes in the C adding them to B.

4: In case the nodes found in B are the leaf, choose the node having the lowest among the total cost to be one of the best sequences that were found and then stop. Else move to Step 2.

d. The remaining members (the rangers) perform the ranging.

e. Calculation of a fitness value for every member.

Step 4. End for

Step 5. End while

E Support vector machine (SVM) Classifier

The SVM is an algorithm which is non-probabilistic and used for the separation of the data either linearly or nonlinearly. The dataset D = {X_i, y_i} wherein the X_i denotes a set of its tuples and the y_i the class label for the tuples. The class labels were -1 and +1 which was for the no and the yes category. The SVM aims at a separate negative along with a positive example of training by means of finding another n-1 hyper plane. The problem of Quadratic Programming (QP) is required for being solved in a new linear data [19]. The problem has been transformed by making use of the theory Lagrange Multipliers along with the Lagrange coefficients sets that are optimal which are obtained. The separating the hyper plane is then written as in (11):

$$W * X + b = 0 \tag{11}$$

Wherein the W = {w_1, w_2, w_3… w_n}, w_n was a weight vector of the n attributes where the b denotes the bias. The actual distance from the separating hyper plane for a certain point on the H1 is 1/|W| and the Distance from a separating of the hyper plane to a particular point on the H2 is 1/|W| so that the maximum margin will be 2/|W|. An MMH will be rewritten to be a decision boundary as according to its Lagrangian formulation as in (12).

$$D(X^T) = \sum_{i=1}^{1} y_i a_i X_i X^T + b_0 \tag{12}$$

Wherein the X^T denotes a test tuple, a_i and the b_0 the numeric parameters, y_i denotes the class label of the support vector X_i. Thus, if the sign which is positive for the MMH equation the XT will come in a category which is positive. In case the sign which is negative of the MMH equation, the XT will come in its negative category. The SVM formula of classifier has been defined as per (13):

$$f(x) = \sum_{i=1}^{n} a_i k(x, x_i) + b \tag{13}$$

IV RESULTS AND DISCUSSION

In this section, the GSO based feature selection and GSO-LBS feature selection methods are used. Experiments are evaluated using 20 to 80 training percentage. The classification accuracy, average recall, average precision and average f measure as shown in tables 1 to 4 and figures 1 to 4.

Table 1 Classification Accuracy for GSO-LBS Feature Selection

Training Percentage	GSO based feature selection	GSO-LBS feature selection
20%	65.83	67.46
40%	69.58	70.79
60%	82.29	85.04
80%	87.29	90.55

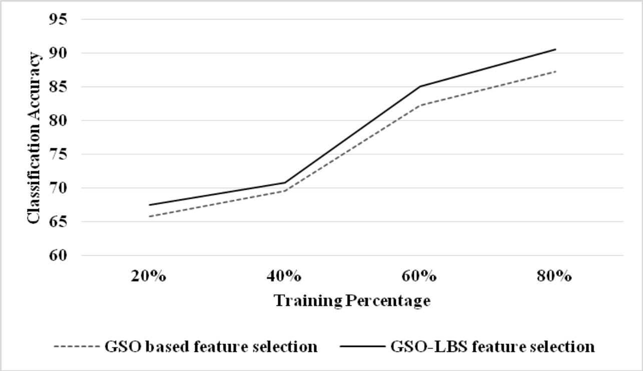

Fig 1 Classification Accuracy for GSO-LBS Feature Selection

From the figure 1, it can be observed that the GSO-LBS feature selection has higher classification accuracy by 2.44% for 20 training percentage, by 1.72% for 40 training percentage, by 3.28% for 60 training percentage and by 3.66% for 80 training percentage when compared with GSO based feature selection.

Table 2 Recall for GSO-LBS Feature Selection

Training Percentage	GSO based feature selection	GSO-LBS feature selection
20%	0.660967	0.673267
40%	0.700967	0.7061
60%	0.823567	0.849433
80%	0.8738	0.903767

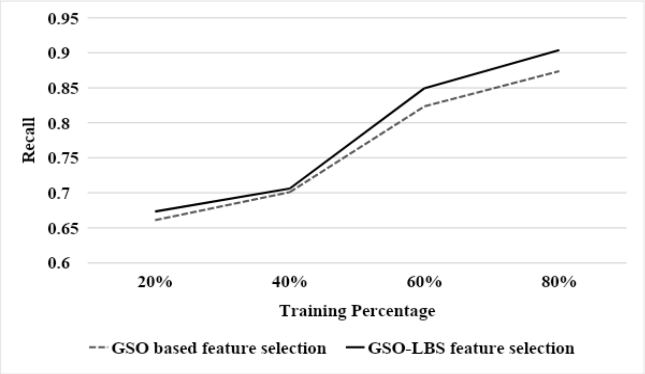

Fig 2 Recall for GSO-LBS Feature Selection

From the figure 2, it can be observed that the GSO-LBS feature selection has higher average recall by 1.84% for 20 training percentage, by 0.72% for 40 training percentage, by 3.09% for 60 training percentage and by 3.37% for 80 training percentage when compared with GSO based feature selection.

Table 3 Precision for GSO-LBS Feature Selection

Training Percentage	GSO based feature selection	GSO-LBS feature selection
20%	0.659267	0.672967
40%	0.697633	0.705867
60%	0.8213	0.850033
80%	0.8715	0.9039

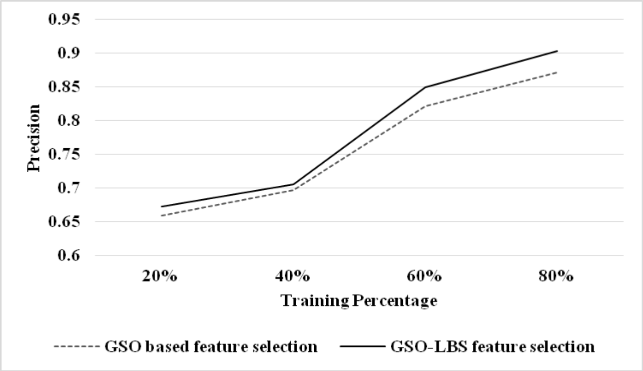

Fig 3 Precision for GSO-LBS Feature Selection

From the figure 3, it can be observed that the GSO-LBS feature selection has higher average precision by 2.05% for 20 training percentage, by 1.17% for 40 training percentage, by 3.43% for 60 training percentage and by 3.64% for 80 training percentage when compared with GSO based feature selection.

Table 4 F Measure for GSO-LBS Feature Selection

Training Percentage	GSO based feature selection	GSO-LBS feature selection
20%	0.657867	0.673033
40%	0.6959	0.705933
60%	0.821967	0.849433
80%	0.8724	0.903767

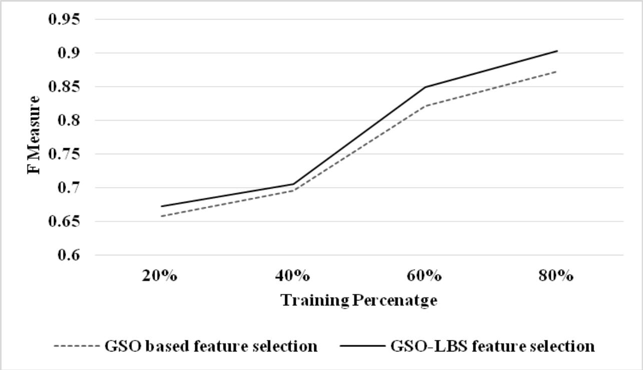

Fig 4 F Measure for GSO-LBS Feature Selection

From the figure 4, it can be observed that the GSO-LBS feature selection has higher average f measure by 2.27% for 20 training percentage, by 1.43% for 40 training percentage, by 3.28% for 60 training percentage and by 3.53% for 80 training percentage when compared with GSO based feature selection.

V CONCLUSION

The data size will be present and is huge continuing to increase each data. Using the cloud services for storing, processing and further analysing data has now changed the IT context. The text classification based on sentiment will be different from the classification of a topical text as it also involves opinion mining. For the purpose of evaluation, there is an Amazon dataset which is used. The sentiment analysis and their features from the text had been extracted by making use of the TF-IDF, and were classified by providing the sentiments and opinions regarding the documents, data or texts by means of the SVM. An optimal selection of a feature is used for the reduction of a feature subset and the complexity of computation thus increasing the accuracy of classification.

A GSO algorithm is a powerful technique of optimization used to solve problems which are combinatorial. The Hybrid GSO-LBS algorithm which is based on the feature selection has been proposed for improving cloud services of a large scale. The results have proved that the feature selection in a GSO-LBS has an accuracy of classification by about 2.44% for the training percentage of 20, by about 1.72% for a training percentage of 40, by about 3.28% for a training percentage of 60, and finally by about 3.66% for a training percentage of 80 on being compared to the feature selection based on the GSO.

REFERENCES

1. Manikandan, R. P. S., & Kalpana, A. M. (2017). Feature selection using fish swarm optimization in big data. Cluster Computing, 1-13.
2. Puthal, D., Sahoo, B. P. S., Mishra, S., & Swain, S. (2015, January). Cloud computing features, issues, and challenges: a big picture. In Computational Intelligence and Networks (CINE), 2015 International Conference on (pp. 116-123). IEEE.
3. Jotheeswaran, J., & Kumaraswamy, Y. S. (2013). Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure. Journal of Theoretical & Applied Information Technology, 58(1).
4. Sumathi, T., Karthik, S., & Marikkannan, M. (2014). Artificial Bee Colony Optimization For Feature Selection In Opinion Mining. Journal of Theoretical & Applied Information Technology, 66(1).
5. Alarifi, A., Tolba, A., Al-Makhadmeh, Z., & Said, W. (2018). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. The Journal of Supercomputing, 1-16.
6. Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. Information Sciences, 272, 16-28.
7. Tan, M., Tsang, I. W., & Wang, L. (2014). Towards ultrahigh dimensional feature selection for big data. The Journal of Machine Learning Research, 15(1), 1371-1429.
8. Yu, K., Wu, X., Ding, W., & Pei, J. (2014, December). Towards scalable and accurate online feature selection for big data. In Data Mining (ICDM), 2014 IEEE International Conference on (pp. 660-669). IEEE.
9. Agarwal, B., & Mittal, N. (2013, March). Optimal feature selection for sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 13-24). Springer Berlin Heidelberg.
10. Priyadarshini, A. (2015). A map reduce based support vector machine for big data classification. International Journal of Database Theory and Application, 8(5), 77-98.
11. Ahmad, S. R., Bakar, A. A., & Yaakub, M. R. (2015, July). Metaheuristic algorithms for feature selection in sentiment analysis. In Science and Information Conference (SAI), 2015 (pp. 222-226). IEEE.
12. Ahmad, S. R., Yusop, N. M. M., Bakar, A. A., & Yaakub, M. R. (2017, October). Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis. In AIP Conference Proceedings (Vol. 1891, No. 1, p. 020018). AIP Publishing.
13. Banerjee, S., & Badr, Y. (2018). Evaluating Decision Analytics from Mobile Big Data using Rough Set Based Ant Colony. In Mobile Big Data (pp. 217-231). Springer, Cham.
14. Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In Innovative Research and Development (ICIRD), 2018 IEEE International Conference on (pp. 1-6). IEEE.
15. He, S., Wu, Q. H., & Saunders, J. R. (2009). Group search optimizer: an optimization algorithm inspired by animal searching behavior. IEEE transactions on evolutionary computation, 13(5), 973-990.
16. Pacifico, L. D., & Ludermir, T. B. (2016, October). Data Clustering Using Group Search Optimization with Alternative Fitness Functions. In Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on (pp. 301-306). IEEE.
17. Raj, S. J., Sathya, S. S., & Ghosh, S. (2015). Flavours of group search optimizer: A survey. International Journal of Computer Technology and Applications, 6(2), 352-358.
18. Valente, J. M. (2010). Beam search heuristics for quadratic earliness and tardiness scheduling. Journal of the Operational Research Society, 61(4), 620-631.
19. Jadav, B. M., & Vaghela, V. B. (2016). Sentiment analysis using support vector machine based on feature selection and semantic analysis. International Journal of Computer Applications, 146(13).

AUTHORS PROFILE

Prof. P. Vasudevan received his Master of Computer Science Engineering degree from Sathiyabama Institute of Science and Technology, Chennai. Currently, he is working as Associate Professor in Department of CSE, Mookambigai College of Engineering since 2006 and doing research in the field of Data Mining in Bharath University. His area of interests includes Data Mining and Warehousing, Cloud Computing and Artificial Intelligence. His e-mail address is vasudevan62@gmail.com

Dr.K.P.Kaliyamurthie is self-directed, enthusiastic educator with a commitment on student development. He is with Bharath University, Chennai, Tamil Nadu, India as Professor and Dean of Computer Science and Engineering. He has over 29 years of rich experience in teaching along with student administration. He has guided more than 300 UG, PG projects and organized various national level conferences. He served as Senior Chair, Technical advisor in various national level conferences and Technical Committee member in International Conferences. He is an active member in CSI, IEEE, ISTE, ACM etc., His area of interests includes Computer Networks, Cloud Computing, Networks and Software Engineering. He can be contacted through Email:kpkaliyamurthie@gmail.com