Product Sentiment Analysis Using Particle Swarm Optimization Based Feature Selection in a Large-Scale Cloud

P.Vasudevan[1*] and K.P.Kaliyamurthie[2]
{vasudevan62@gmail.com[1*] , kpkaliyamurthie@gmail.com[2]}

Research Scholar, CSE, Bharath Institute of Higher Education and Research, Chennai, India.[1],
Professor/CSE, Bharath Institute of Higher Education and Research, Chennai, India[2]

Abstract. Cloud computing is evolving by shifting its services out of the applications of the firewall, storage, services, and applications that are accessible on the web. After this, the services will be used with the help of the Internet and paid according to the user's/customer's needs. Big data can be efficiently and economically analysed using Cloud computing. Sentiment Analysis deals with study of opinions, and is based on the emotions, attitudes, and opinions of this entity. The objective of the proposed work sentiment analysis using Particle Swarm Optimization (PSO) algorithm based on new Feature Selection (FS) method. FS method is quite complex and a demanding task in terms of computation more so for a high dimension dataset. Swarm Intelligence (SI) is techniques capable of solving computational problems that are NP-hard (Non-Deterministic Polynomial time). It is gaining plenty of popularity to solve the problems of optimization and FS. Particle Swarm Optimization (PSO) is used widely for solving problems of optimization and also the problems of FS. The Support Vector Machine (SVM) analyses data and further identify patterns utilized for classification purposes. In this work, a PSO-based FS is proposed for product sentiment analysis. The classification accuracy achieved by PSO based FS is higher by 5.93% and by 6.91% for 20% training when compared to IG and GA based FS, respectively. Similarly, classification accuracy achieved by PSO based FS is higher by 3.65% and by 0.89% for 80% training when compared to IG and GA based FS, respectively.

Keywords: Sentiment Analysis, Feature Selection (FS), Swarm Intelligence (SI), NP-hard, Particle Swarm Optimization (PSO).

1 Introduction

Today, corporations are getting more information produced frequently and incessantly online, which needs to be analysed for various measures for the Expansion of their business. Big Data can be defined as the process used to record and further analyse huge datasets that were structured, semi-structured, or unstructured. Distribution of information on various systems can be a challenge that is important while processing large amounts of data within a reasonable time frame. Aside from this, deployment, accessibility, security, and privacy were the other issues the decision-makers had to take into consideration before there were major benefits among Big Data for gaining a competitive advantage [1].

Another primary challenge faced by the researchers that work with Big Data will be its high dimensionality occurring when the dataset has a very high number of attributes. For

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308639

reducing the number of attributes (or features) a process for identifying and removing features that were irrelevant (which cannot be used to classify data) is applied. The selection of feature subsets relevant to a class attribute was an important step in creating a usable and meaningful model [2].

Today, social media has a major effect in the digital media and communications. It was evident from the increase in the number of consumer opinions, as well as reviews, regarding the products of smartphones writing on different types of social media. It can be identified that different sentiments of the product can be positive or negative and sometimes neutral. Sentiment analysis can be identified as a study of computation on the opinion study, emotion, and behaviour of people, topics, or events. These topics will normally be reviews of various datasets, among which one is a product review. A smartphone product review is re-examined as a product review. A re-examination of smartphone product review is made means of a classification of either positive or negative classes which can be a way in which the response of consumers can be found out properly and quickly.

The meta-heuristic algorithms will follow procedures of the search for solving problems of optimization. The common algorithms include the Particle Swarm Optimization (PSO) and the Genetic Algorithm (GA). The PSO is employed widely for solving problems of optimization and also for solving problems of FS. For the PSO, every solution has been viewed as a particle, and an algorithm will also search for ideal solutions by taking into consideration the experience of the particles [3].

The Support Vector Machine (SVMs) is an effective technique applied for the problem of binary classification [3]. They are able to achieve optimal classifications for the cases that are linear and separable. The power of generalization is very remarkable, and the main advantages are it not having a local minimum, solution’s sparseness, and their capability of optimizing the margin. The SVM can identify and separate hyperplane to maximize the actual margin between two of the classes.

In order to overcome this drawback, a method of PSO-based FS used for text mining Sentiment Analysis has been proposed. The rest of the paper contains related work available in the literature in Section 2. The techniques used in the investigation are explained in Section 3. Simulation results were explained in Section 4, and the Section 5 concludes the paper.

2 Literature Survey

Developing IOT technologies were massively admired and also accepted as applications and tools of social media. Applying Opinion mining with Sentiment Analysis (OMSA) in Big Data is employed in such a way in which the opinion is categorized into various sentiments and evaluation of the public mood. Furthermore, there are various OMSA techniques that were developed in different periods that were applied to the experimental settings. Shayaa et al. [4] had proposed a review of literature that was systematic and comprehensive. This aims to discuss the OMSA, its types and techniques, and the non-technical aspect as discussed using areas of application. The manuscript also highlights the technical characteristics of the OMSA as challenges in developing the techniques and the non-technical issues based on application.

Meera and Jeetha [2] focused on the different research methods to find methods of effective selection of features to produce optimal feature subset. The earlier approaches were presented along with their advantages, as well as disadvantages, and there were some additional areas of research to focus. All simulations were carried out on MATLAB simulation, which was differentiated for identifying the ideal methodologies under various measures of performance.

As comprehensive surveys made in this field were minimal, the main objective was to fill the existing gap in covering the SI algorithms for the FS. Brezočnik et al. [6] had made a review of the literature of the SI algorithms to provide an overview of about 64 SI algorithms used for the FS grouped into various approaches that were explained along with their settings and used for a variety of aspects of the FS. Several datasets were frequently used for evaluating the SI algorithms used in the FS that were presented along with the common areas of application. There were guidelines on ways in which the SI approaches used for the FS were developed and were also provided for research support to analysts for tasks of data mining and for endeavours in which open questions and issues were discussed. This way, with the proposed framework and given explanations, it is possible to design the SI approach for a particular problem of the FS.

Big Data consists of challenges that stem from communities of academic research, and IT deployment made commercially that are the root sources in Big Data found on the curse of dimensionality and data streams. This is called data that is sourced from the streams of data that collect continuously, thus the traditional algorithms of batch-based model induction that may not be feasible for real-time data mining. According to Joshi et al. [7], this concept of FS permits data choice by using its features to filter data based on features. Where there is a large dataset coming, FS may get complicated. And for cases of a high dimension of data, filters will also be increased accordingly.

An SVM can identify a separated hyperplane to maximize the margin of two different datasets belonging to other classes. But the SVM lacks suitable features or parameters, and election features along with setting parameters at the SVM can affect the features and their classification accuracy. Thus, in this type of research, the method of merger for the election of features such as the PSO is employed to effectively increase the classification accuracy of the SVM. Wahyudi and Kristiyanti [8] had produced the text of classifications that were either positive or negative for the reviews of smartphone. The evaluation was made using a 10-Fold Cross-Validation. Accuracy evaluation was made using the ROC curve and the Confusion Matrix[9]. Results proved that there was an increase in the accuracy of the SVM from 82.00% to 94.50%.

3 Methodology

In this section, a review of the dataset of Amazon products used for evaluation is presented. The proposed FS based on the PSO, and the SVM employed is detailed.

Dataset

Amazon is a large site of E-Commerce, with numerous reviews seen online. The Amazon product dataset is created by the researchers, which is employed for investigations **[9].** This was an unlabelled dataset to be used as a model of supervised learning for labelling of data. The website proves to be a forum that has assorted opinions. The websites that are similar to amazon.com will promote the users to write a review. Such reviews may be grouped into three categories: either positive or negative and sometimes neutral. For all the chosen categories, three among them are the reviews of Electronics products, accessories, and Cell phones that constitute about 48500 product reviews. Among them, 21600 reviews are for mobiles, 24352 for electronics, and 2548 for musical instruments.

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

TF-IDF can be termed as a numerical statistic to mirror the rank of a word to a document in a collection. The method is utilized in the form of a weighting factor in text mining. The TF-IDF is chiefly used for stopping the filtering of words in the summarization of text and application categorization. Conventionally, the value of the TF-IDF will increase in proportion as the frequency of word occurring in one document and is further offset by the frequency of the same word found in the collection. This assist in controlling the statistic that certain words tend to be more common than that of the others. A frequency term denotes the actual raw frequency of a term found in a document. Furthermore, the word with regards to the frequency of an inverse document will be a degree of whether this word is rare or common among the reviews got by dividing the total reviews by the number of words [10].

TF-IDF denotes the numeric measure applied to score the word and its significance in the document based on its frequency. The intuition here was that the word will need an additional high score. If the word continues to appear in several documents that are not unique as an intender, it will be given a score that is lower. The TF-IDF is formulated as below (1):

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

Here, t signifies all terms; d is the document; D is its document collection.

Proposed Particle Swarm Optimization (PSO) based Feature Selection (FS)

PSO [11, 12] is a technique of metaheuristics which simulate the bird and its movements for finding food. Every particle within the swarm will represent a new candidate solution that flies through a search space that is multi-dimensional. The particle will make use of the best position which is explored and the neighbours will shift nearer to their optimum solution. The particle and its fitness (the actual goodness) of the particle to its global minimum will be computed in accordance with the fitness function that is pre-defined. If the search space is found to be D-dimensional with m particles in the swarm, every particle will be placed in the position which is $X_i = \left[x_{i1},\ x_{i2},\ \ldots,\ x_{iD}\right]$ that has a velocity $V_i = \left[v_{i1},\ \mathrm{v}_{i2},\ \ldots,\ \mathrm{v}_{iD}\right]$, wherein i=1, 2, …, m. In the case of a PSO algorithm, every particle will move closer to its best position (pbest) represented as $P_{best_i} = \left[p_{best_{i1}},\ p_{best_{i2}}, \ldots,\ p_{best_{iD}}\right]$ with the best position for the entire swarm (the gbest) shown as $G_{best} = \left[g_{best_1},\ \mathrm{g}_{best_2}, \ldots,\ \mathrm{g}_{best_D}\right]$. Now every particle will change its actual position in accordance with the velocity and this is generated randomly to the pbest as well as the gbest positions. For every particle, i and its dimension s, there is a new velocity which is vis and a position xis that is computed as per equation (2):

$$\begin{aligned} v_{is}^{t} &= wv_{i}^{t-1} + c_1 b_1 (pbest_{is}^{t-1} - x_{is}^{t-1}) + c_2 b_1 (gbest_{s}^{t-1} - x_{is}^{t-1}) \\ x_{is}^{t} &= x_{is}^{t-1} + v_{is}^{t} \end{aligned} \quad (2)$$

Wherein t denotes the iteration number. Inertial weight w will be utilized for controlling velocity and the balance of both exploration, as well as exploitation of the algorithm. A larger value of that of w will keep the particles within a high velocity thus preventing them from getting caught inside the local optima. There may be a smaller value of w that sustains the particles at a low velocity and further boosts them to exploit the search area. Constants c1 and c2 will denote the coefficients of acceleration to regulate if the particles will want to shift nearer to their pbest or their gbest positions. Both b1 and b2 will denote independent random numbers that are distributed uniformly between 0 and 1. The termination criterion for the PSO[11] will include the actual maximum number of generations and the designated value for pbest, without enhancement in pbest. PSO can be implemented with a few parameters escaping from local minima [12]. In this work, maximum number of iteration is taken as the termination criterion [13].

Normally, in the case of a continuous PSO being applied to difficulties of FS, the search space, and its dimensionality will be n and this represents the total features available within the dataset. Every particle found in the swarm will be encoded by making use of a vector of the n real numbers. The actual position of a particle i which is in the dth dimension, the xid will be in the interval [0, 1]. For the purpose of determining if the feature is chosen on not, threshold $0 < \theta < 1$ will be required for relating with real numbers found in the position vector. In case xid> θ, the consistent feature d is chosen.

The goal of a FS was to increase the accuracy of classification in a dataset. Thus, while making an evaluation of the particle and its fitness value, a training set is divided equally into 10 folds the actual fitness value of the particle, also cross-validation is run on a training set that has a supervised scheme of machine learning with the feature subset represented by the particle. The fitness value of this particle denotes the actual average accuracy of the 10 runs.

Support Vector Machine (SVM) Classifier

A technique of classification that is powerful and is also a sample of supervised learning working on the principle of the lowest structural risk is the SVM. At the time of training, the algorithm will create a new hyperplane to separate the samples into positive or negative. After this, the new samples are classified by means of specifying the point on the hyperplane in which every sample has to be placed. The SVM is a method of supervised learning which can analyse data and further identify the patterns employed in classification. The SVM[14] also has the major benefit of identifying separate hyperplanes to maximize the margin existing between two different classes [15].

The SVM was based on the theory of statistical learning to increase the property of generalization. It uses training instances to predict newer instances using two class labels −1 and 1. In figure 1, a hyperplane is $w^T x + b = 0$, wherein $w \in R^n$ denotes an orthogonal to a hyperplane and $b \in R^n$ the constant as in equation (3):

$$D = \left\{ (\vec{x}_i, \vec{y}_i) \mid \vec{x}_i \in R^m, \vec{y}_i \in \{-1, +1\} \right\}_{i=1}^{n} \quad (3)$$

Wherein, x_i denotes a m-dimensional real vector, y_i the class of an input vector x_i either −1 or +1. It aims at searching a hyperplane to maximize the margin between two sample classes D with minimal empirical risk [15].

$$y_i(\vec{w}^T \vec{x} + b) \geq 1 \quad (4)$$

The SVM also attempts to increase the distance between two hyperplanes. One computes the distance between them using $\frac{1}{\|\vec{w}\|}$. SVM training for a case that is non-separable is solved with a quadratic optimization as per equation (5):

$$\min imize: \ \mathrm{P}(\vec{\mathrm{w}},\mathrm{b},\xi) = \frac{1}{2} \|\vec{w}\| + C \sum_{i=1}^{n} \xi_i$$
$$subjectto: \ \mathrm{y}(\vec{\mathrm{w}}.\phi(\vec{\mathrm{x}}) + \mathrm{b}) \geq 1 - \xi_i \quad, \xi_i \geq 0 \quad (5)$$

3 Results And Discussion

The proposed PSO based FS methods are evaluated and compared with IG and GA FS for varying training percentage. The performance metrics used are classification accuracy, recall, precision and average f-measure. As shown in tables 1 to 4 values obtained figures 1 to 4 respectively.

Table 1 Classification Accuracy for PSO- FS

Training Percentage	IG based FS	GA based FS	PSO based FS
20%	63.33	62.71	67.2
40%	67.92	68.79	69.42
60%	78.33	75.46	79.94
80%	81.25	79.04	81.98

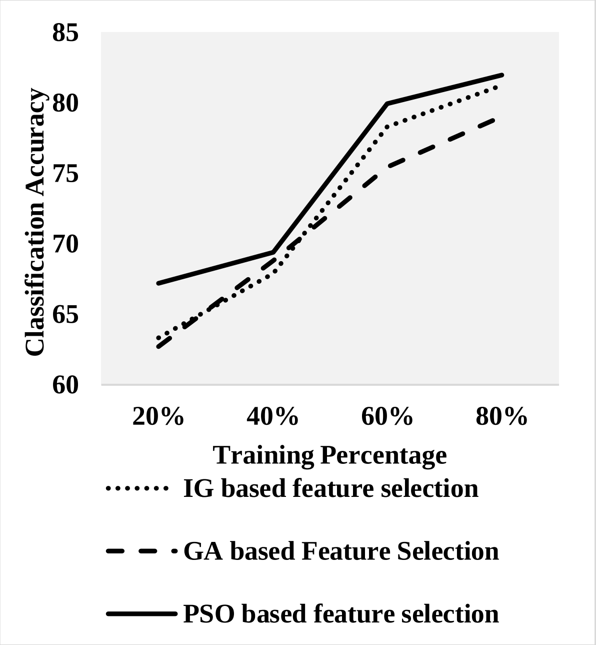

Fig 1 Classification Accuracy for PSO – FS

From figure 1 and Table 1, it can be observed that the classification accuracy achieved by PSO based FS is higher by 5.93% and by 6.91% for 20% training when compared to IG and GA based FS, respectively. Similarly, classification accuracy achieved by PSO based FS is higher by 3.65% and by 0.89% for 80% training when compared to IG and GA based FS, respectively.

Table 2 Recall for PSO - FS

Training Percentage	Positive	Negative	Neutral
IG based FS			
20%	0.6	0.675	0.6286
40%	0.6222	0.7	0.7286
60%	0.7444	0.8	0.8143
80%	0.7991	0.817	0.8152
GA based FS			
20%	0.6522	0.6163	0.6071
40%	0.7067	0.6825	0.67
60%	0.8	0.745	0.7071
80%	0.8667	0.745	0.7443
PSO based FS			
20%	0.6861	0.6589	0.6688
40%	0.7053	0.6797	0.6966
60%	0.7917	0.805	0.8029
80%	0.7778	0.825	0.8429

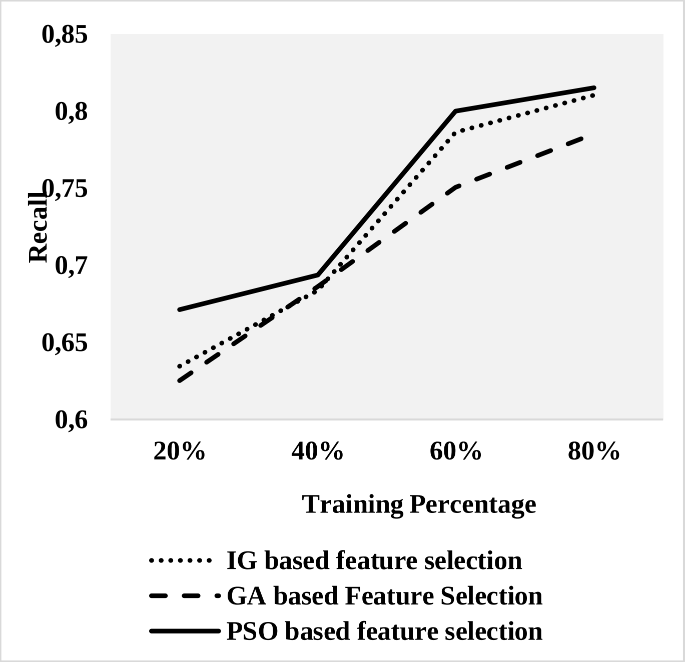

Fig 2 Average Recall for PSO - FS

From figure 2 and Table 2, it is seen that recall achieved by PSO based fs is higher by 5.62% and by 7.11% for 20% training when compared to IG and GA based fs, respectively. Similarly, recall achieved by PSO based fs is higher by 1.72% and by 6.34% for 60% training when compared to IG and GA based fs, respectively.

Table 3 Precision for PSO - FS

Training Percentage	Positive	Negative	Neutral
IG based FS			
20%	0.6879	0.6353	0.5752
40%	0.732	0.6829	0.6258
60%	0.8121	0.7805	0.755
80%	0.8652	0.7868	0.7732
GA based FS			
20%	0.6732	0.6178	0.5822
40%	0.7106	0.6877	0.6596
60%	0.7843	0.7385	0.7333
80%	0.7975	0.8043	0.7651
PSO based FS			
20%	0.7102	0.706	0.5972
40%	0.7346	0.7343	0.6131
60%	0.849	0.7781	0.7667
80%	0.8434	0.8049	0.7867

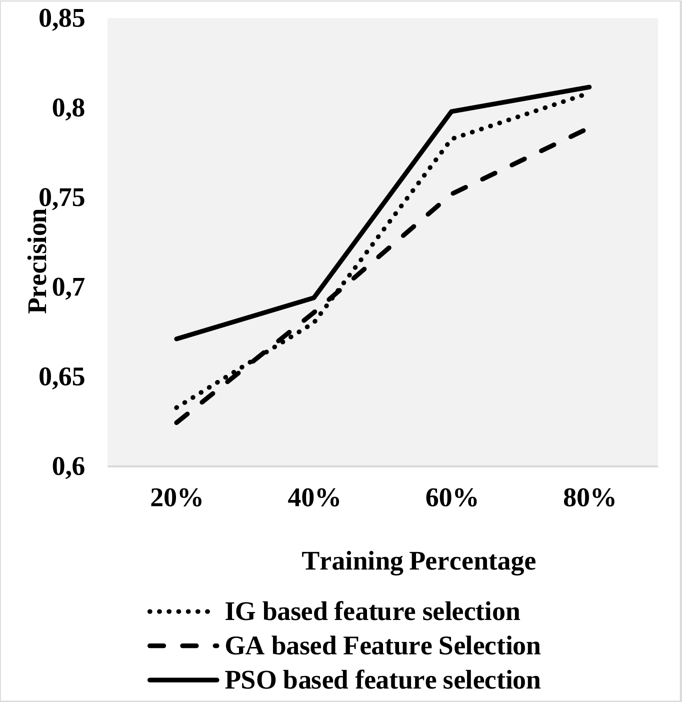

Fig 3Average Precision for PSO - FS

From figure 3 and Table 3, it seen that the precision achieved by PSO based fs is higher by 5.9% and by 7.21% for 20% training when compared to IG and GA based FS, respectively. Similarly, the precision achieved by PSO based fs is higher0.4% and by 2.43% for 80% training when compared to IG and GA based FS, respectively.

Table 4 F Measure for PSO - FS

Training Percentage	Positive	Negative	Neutral
IG based FS			
20%	0.641	0.6545	0.6007
40%	0.6726	0.6913	0.6733
60%	0.7768	0.7901	0.7835
80%	0.8308	0.8016	0.7936
GA based FS			
20%	0.6625	0.617	0.5944
40%	0.7086	0.6851	0.6648
60%	0.7921	0.7417	0.72
80%	0.8307	0.7735	0.7546
PSO based FS			
20%	0.6979	0.6816	0.631
40%	0.7197	0.7059	0.6522

60%	0.8193	0.7913	0.7844
80%	0.8093	0.8148	0.8138

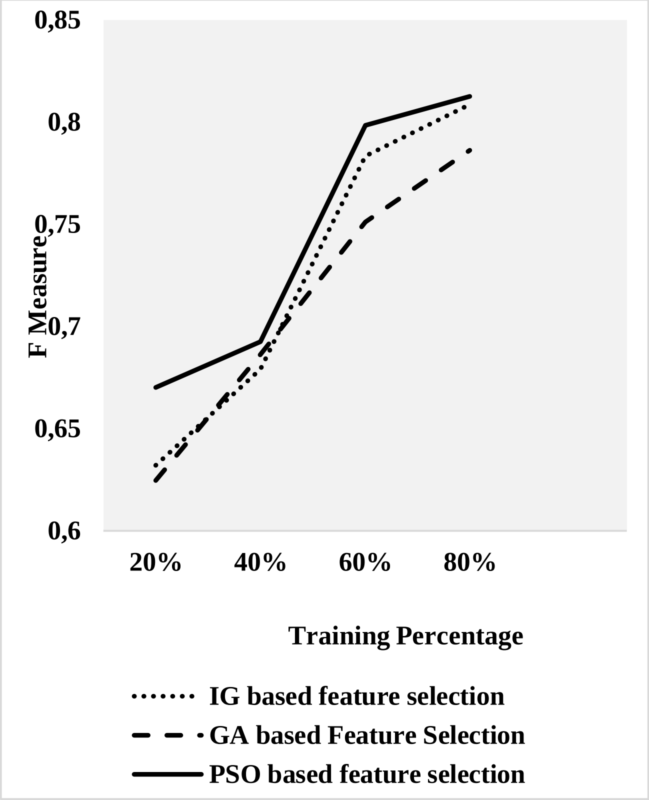

Figure 4 Average F Measure for PSO - FS

From Figure 4 and Table 4, it seen that the f-measure achieved by PSO based FS is higher by 5.85% and by 7.03% for 20% training when compared to IG and GA based FS, respectively. Similarly, the f-measure achieved by PSO based FS is higher by 0.49% and by 2.81% for 80% training when compared to IG and GA based FS, respectively.

4 Conclusion and future work

Sentiment Analysis will be performed for checking the positive, negative or neutral opinions of users. There are several issues connected to the classification of the sentence that is solved using machine learning. The PSO is employed for optimal parameters in the region of prediction and can also be employed for overcoming the local optimum problem. Results have proved that the classification accuracy achieved by PSO based FS is higher by 5.93% and by 6.91% for 20% training when compared to IG and GA based FS, respectively. Similarly, classification accuracy achieved by PSO based FS is higher by 3.65% and by 0.89% for 80% training when compared to IG and GA based FS, respectively.

Conflicts of interest

The authors should declare any conflicts of interest exist. If no conflict exists, the authors should state: the authors have no conflicts of interest to declare.

References

[1] A. Misra, A. Sharma, P. Gulia, and A. Bana. Big data: Challenges and opportunities. Int. J. Innov. Technol. Exploring Eng. 2014; 4(2): 41–42.

[2] Meera, S., &Jeetha, B. R. Survey on Swarm Search Feature Selection for Big Data Stream Mining. International Journal of Computer Applications. 2017; 158(1).

[3] Madhusudhanan, B., Sumathi, P., Karpagam, N. S., Mahesh, A., &Suhi, P. A. P.. An hybrid metaheuristic approach for efficient feature selection. Cluster Computing. 2019; 22(6). 14541-14549.

[4] R. Batuwita and V. Palade. Class Imbalance Learning Methods for Support Vector Machines. In Imbalanced Learning: Foundations. Algorithms and Applications. Haibo He and Yunqian Ma Ma (Eds.). Wiley. 2013.

[5] Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. Sentiment analysis of big data: Methods, applications, and open challenges. IEEE Access. 2018; 6. 37807-37827.

[6] Brezočnik, L., Fister, I., &Podgorelec, V. (2018). Swarm intelligence algorithms for feature selection: a review. Applied Sciences. 8(9). 1521.

[7] Joshi., R, B and Rode., S, V. "Particle Swarm Optimization Feature Selection for Data Stream Mining". International Journal of Innovative Research in Computer and Communication Engineering. 2016; 4(11).

[8] Wahyudi, M., &Kristiyanti, D. A. Sentiment Analysis Of Smartphone Product Review Using Support Vector Machine Algorithm-Based Particle Swarm Optimization. Journal of Theoretical & Applied Information Technology. 2016; 91(1).

[9] He, Ruining, and Julian McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with oneclass collaborative filtering." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee. 2016.

[10] Munot, N., &Govilkar, S. S. Comparative study of text summarization methods. International Journal of Computer Applications. 2014;102(12). 33-37.

[11] Amoozegar, M., &Minaei-Bidgoli, B. Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism. Expert Systems with Applications. 2018;113. 499-514.

[12] Marinakis, Y., Marinaki, M., & Dounias, G. A hybrid particle swarm optimization algorithm for the vehicle routing problem. Engineering Applications of Artificial Intelligence. 2010; 23(4). 463-472.

[13] Xue, B., Zhang, M., & Browne, W. N. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Applied soft computing. 2014; 18. 261-276.

[14] Kiran Kumar, T.V.U., Karthik, B., Improving network life time using static cluster routing for wireless sensor networks, Indian Journal of Science and Technology, 2013, 6(SUPPL5), pp. 4642–4647

[15] Thamarai, P., Karthik, B., Kumaran, E.B., Optimizing 2:1 MUX for low power design using adiabatic logic, Middle - East Journal of Scientific Research, 2014, 20(10), pp. 1322–1326

[16] Istia, S. S., &Purnomo, H. D. Sentiment Analysis of Law Enforcement Performance Using Support Vector Machine and K-Nearest Neighbor. In 2018 3rd International Conference on Information Technology. Information System and Electrical Engineering (ICITISEE) IEEE 2018; 84-89.

[17] Aruna, S. K., Sindhanaiselvan, K., &Madhusudhanan, B. Computerized grading of brain tumors supplemented by artificial intelligence. Soft Computing. 2019;1-7.

Author Bibliography

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