International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Design of Ensemble Classifier Selection Framework Based on Ant Colony Optimization for Sentiment Analysis and Opinion Mining

Sanjeev Kumar, Ravendra Singh

Abstract:Ensemble Classifier provides a promising way to improve the accuracy of classification for sentiment analysis and opinion mining. Ensemble classifier should combine with diverse base classifiers. However, establishing a connection between diversity and accuracy of ensemble classifier is tedious task because of sensitivity between diversity and accuracy. In this paper an Ensemble classifier selection (ECS) framework based on Ant Colony Optimization (ACO) algorithm is presented. The framework provides a subset of base classifiers from a given set of classifiers with maximum possible diversity and accuracy to design an ensemble classifier for sentiment analysis and opinion mining. This framework uses diversity measures and accuracy as selection criteria for classifier selection for ensemble creation. The experimental result shows that the ensemble classifiers provided by this framework presents an efficient way for sentiment analysis and opinion mining.

Keywords:Ant Colony Optimization, Ensemble Classifier, Sentiment Analysis and Opinion mining, Ensemble Classifier Selection

I. INTRODUCTION

The amount of text data available on the internet has increased so much in the last few years. This huge amount of available data is affecting not only IT industries but also various other sectors like business companies, public services. The government collects text generated by the public via feedback and complaint system to become aware of public opinions for policy establishment, business companies use market analysts to take advantage of product or services reviews for strategic analysis and commercial planning and e-learning systems read the student sentiment to adapt teaching resources and methodologies. Sentiment Analysis is a field of study within Natural Language Processing that identify the mood or opinion of subjective elements within a text. It determines the attitude or mood of a reviewer with respect to some topic or the overall polarity of the document. Extracting sentiments and determining the polarity of the text has become a well-known classification task for NLP researchers. The various classification algorithms are designed to achieve the best possible performance for the classification task at hand.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Sanjeev Kumar*, Department of CS & IT, M.J.P. Rohilkhand University, Bareilly, India. Email: sanjeev.kumar8988@gmail.com
Dr. Ravendra Singh, Department of CS & IT, M.J.P. Rohilkhand University, Bareilly, India. Email: r.singh@mjpru.ac.in

This led to the development of a different type of classification schemes. Such as SVM does the classification by separating the classes by finding the hyperplane in between them, Decision tree does the classification by creating a tree based on information gained by available features, Naïve Bayes algorithm tries to find the posterior probability using prior probabilities to assign a data item in a class etc. After experiment assessment of these different types of classification schemes one of the techniques had been selected which provides the best results. It is observed that different classifiers offered complementary information about the patterns which can be used to improve the performance of the selected classifier. These observations give a reason tocombine various kinds of classifiers and improve the performance of classifiers. The problems remain that how to select a subset of the classifier to improve the performance. This paper provides Ensemble classifier selection framework to solve this problem and finds an optimal subset of classifiers provides the best possible classification results.

The reason behind the ensemble classifier [1] is the theory in which it says that combining the answers of many different people can increase the accuracy of the final answer. Because of that an ensemble classifier measure a set of individual pattern classifier and then the final result is obtained by aggregating these classifier's results. The diversity [2] [3] between these individual classifiers plays an important role to improve the accuracy in the final result.

Sentiment analysis and Opinion mining [4] [5] is a special kind of task because it involves the analysis of reviews given by real individual persons. The text review data doesn't carry the same pattern in each review which makes this task more complicated. A single classifier is not sufficient to analyse these reviews because of that combining various different classifiers becomes a need to solve this problem. Finding the set of classifiers with the best accuracy becomes a combinatorial search problem. Various optimization techniques have been implemented to solve this problem. Harmony Search, Genetic Programming, Stepwise forward selections are some of the available algorithms that have been used to this optimization problem. In a different kind of Metaheuristic algorithms, ant colony optimization is also showing a promising way to this optimization problem. It is easy to implement and works very fast. Ant colony optimization algorithm based on the path search capability of ants. Ants use pheromone trail to efficiently find a short distance from colony to food.

Ants leave pheromone on the paths where they go and other ants follow these pheromone trails to find the path. Marco Dorigo, Mauro Birattari, and Thomas Stutzle [6] observe this special capability of ants and develop an optimization technique named as Ant System.

In this paper section II gives the background and related work done, section III give introduction about ACO. Section IV discusses the classifier ensemble selection framework. Section V give details about the integration of accuracy and diversity. The experiments and the results are discussed in section VI.

II. BACKGROUND AND RELATED WORK

There is a huge amount of work has been done on combining classifiers in the last decade. J. Kittler [7] provides a theoretical framework to combine classifiers which use distinct pattern representations and show that many existing schemes can be considered as special cases of compound classification where all the pattern representations are used jointly to make a decision.

E. Fersini, E.Messina, F.A.Pozzi [8] proposed a Bayesian Model Averaging based ensemble method where both uncertainty and reliability of each single model is taken into account. Paper works on ensemble learning to reduce the noise sensitivity related to language ambiguity. Author proposed greedy approach for classifier selection problem. This greedy approach calculates the contribution of each model with respect to the ensemble.

E. K. Tang, P. N. Suganthan · X. Yao [3] has done a deep analysis of diversity measures and presents a theoretical analysis on six existing diversity measures (namely disagreement measure, double fault measure, KW variance, inter-rater agreement, generalized diversity and measure of difficulty). This paper showed underlying relationships between various diversity measures, and relate them to the concept of margin.The concept of margin explicitly related to the success of ensemble learning algorithms.

M. Dorigo,Mauro Birattari, and Thomas Stutzle [6] analysed the behaviour of ants and develop an optimization technique named as Ant System based on path search technique used by ants to search the path. This algorithm uses the concept of pheromone released by ants to find the path. Ants deposit pheromone in order to mark some favourable path that should be followed by other members of the colony. Ant colony optimization exploits a similar mechanism for solving optimization problems.

Ahmed Al-Ani [9] has done a feature subset selection using a novel method that utilizes the ACO algorithm. In this hybrid evaluation measure is proposed for evaluating the overall performance of the subsets with local importance of features. It uses filter evaluation function Mutual Information Evaluation Function (MIEF) to measure the local importance of a given feature. Author made the comparison of this method with Stepwise and GA based feature selection method and shown that ACO based feature selection method works better then both the algorithms given.

Y. Chen,Man Leung Wong [10] proposed an ACO-Stacking ensemble approach. In this novel approach ACO is used to search the good configuration of stacking ensembles for specific datasets. In which the first layered base classifiers are selected by Ant Colony Optimization. To test the cumulative percent of improvement that ACO-Stacking can achieve comparing to bagging and boosting approaches, the Percent Improvement (PI) test is used.

R. Geetha, G. Umarani Srikanth [11] provided a study of research of ACO in various engineering applications related to computer science fields such as mobile and wireless networks, grid computing, P2P Computing, Pervasive computing, Data mining,Image processing, Software engineering, Database systems, sensor networks, Multicore Processing,Biomedical applications, Artificial intelligence and also other domains relevant to Electronics and Electrical Engineering fields.

X. Zeng,Derek F. Wong, and Lidia S. Chao [12] presented a weighted diversity and accuracy method to find the balance between diversity and accuracy that enhance the predictive ability of an ensemble for unknown dataFinal score is determined by computing the harmonic mean of accuracy and diversity for quality assessment of an ensemble, where two weight parameters are used to balance them.

Gang Yao et.al. [13] proposed a new ensemble subset evaluation method. The method applied diversity measures for classifier ensemble reduction. The approach used three conventional diversity algorithms and one new developed diversity measure method to calculate the diversity's merits for the reduction of classifier ensemble.

From these astounding works, it is clear that ACO provides a promising way to find the optimized set of classifiers. In this paper, we apply ACO to find the optimized set of classifiers on accuracy and diversity. Gang Yao et.al. [13] provides a metric Pairwise Quality Evaluation (PQE) to combine diversity and accuracy for the ensemble of classifiers. In this paper, we used this metric as optimization criteria for ACO to find the ensemble of classifiers which gives better accuracy then available ensemble classifier.

III. ANT COLONY OPTIMIZATION

The ant colony optimization algorithm (ACO) is a probabilistic technique which solves optimization problems by reducing the problem in a problem of finding good paths through graphs. Ants navigate from their nest to food source. They discover the shortest path via pheromone trails. First-time ants move at random and deposit pheromone on the path. Paths have more pheromone have more probability of being followed by other ants. Ensemble classifier selection is a combinatorial problem and ACO provides a way to solve this problem efficiently. Selecting an optimized subset of classifiers from a set of available classifiers with maximum diversity and accuracy reduced to a problem of finding the shortest path from available paths. Each classifier works as a vertex in the graph. To find the shortest path, those classifiers are selected which have maximum probability. The shortest path represents a subset of classifiers having maximum diversity and accuracy combining all classifiers in a subset.

Algorithm:

Set parameters, initilize pheromone trails
SCHEDULE_ACTIVITIES
ConstructAntSolutions
DaemonActions
UpdatePheromones
END_SCHEDULE_ACTIVITIES

IV. THE ENSEMBLE CLASSIFIER SELECTION FRAMEWORK USING ACO

This paper presents an ensemble classifier selection framework which is based on Ant Colony Optimization and provides ensemble classifiers with best results. This framework takes a set of base classifiers and dataset as input. After preprocessing of data, framework calculates results for each individual base classifier stored in it. These results are used to compute the diversity and joint accuracy between various pairs of classifiers. This diversity and joint accuracy further used in the calculation of PQE merit to select the best k ants. The problem of selection of optimal classifiers for best possible results becomes the combinatorial optimization problem. In this framework optimization problem is solved by ACO. ACO uses pheromone trail and accuracy to calculate the probability for selection of base classifiers in an ensemble. After n number of iterations, the ants provide the best possible set of the classifier with maximum accuracy. Figure 1 providesa block diagram of the Ensemble Classifier Selection Framework.

V. INTEGRATION OF CLASSIFIER DIVERSITY AND JOINT ACCURACY FOR ENSEMBLE CLASSIFIER SELECTION

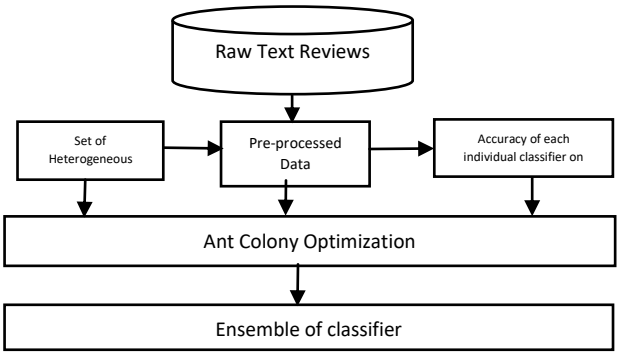

Fig. 1.The Ensemble Classifier Selection Framework

This paper focuses upon two points, first is how to incorporate the diversity and accuracy in ensemble classifier selection, and second is how to find optimized classifier subset using ant colony optimization. PQE merit is used in ACO algorithm for integration of classifier diversity and join accuracy. For calculation of accuracy and diversity between combinations of classifiers, the prediction combinations of each two classifiers are given as table 1. In this 'a' is the number of time when both classifiers are correct. 'b' is the number of time when A classifier is correct and B classifier is incorrect. 'c' is the number of times when B is correct but A is incorrect. At last, 'd' is the number of time when classifier A and B both are incorrect.

Table 1 The prediction combination of each two classifiers.

Result		*Classifier B*	
		Correct	*Incorrect*
Classifier A	*Correct*	*a*	*b*
	Incorrect	*c*	*d*

If the number of ensemble candidates is k, then the number of any two classifiers taken in combination, m, is obtained by:

$$m = \frac{k(k-1)}{2} \quad (1)$$

A. Precise

To calculate Pairwise Quality Evaluation (PQE) merit first it needs to calculate the Precise P, the average precision of each ensemble candidate. P is the entire ensemble candidate's precision rate, rather than a single classifier's precision rate. The joint accuracy between two classifier combination is given as:

$$p = \frac{a}{a+b+c+d} \quad (2)$$

The ensemble member's accurate P is obtained by equation 3.

$$P = \frac{\sum_{n=1}^{m} p_n}{m} \quad (3)$$

Where n denotes the nth pair of classifiers; m is determined by Eq. 1.

B. Diversity

In the next step, we need to calculate the Diversity q, of each classifier ensemble's candidate. There are two kinds of diversity calculation Pairwise and Non-Pairwise. Pairwise diversity is calculated by the summation of all the diversity pairs of each classifier combination, while Non-Pairwise measure is obtained directly from the candidate's diversity.

Pairwise Evaluation

The four basics pairwise diversity measures are used in this paper and these are given as:

q_{QS} is the Q statistics method, and is defined as:

$$q_{QS} = \frac{ad-bc}{ad+bc} \quad (4)$$

q_{DM} is the disagreement measure method:

$$q_{DM} = \frac{b+c}{a+b+c+d} \quad (5)$$

and q_{DFM} is the double-fault measure

$$q_{DFM} = \frac{d}{a+b+c+d} \quad (6)$$

and q_{CC} is the Correlation-Coefficient

$$q_{CC} = \frac{ad-bc}{\sqrt[2]{(a+b)(c+d)(a+c)(b+d)}} \quad (7)$$

The ensemble candidates' Pairwise diversity Q_x is obtained as follows:

$$Q_x = 2\frac{\sum_{n=1}^{m} q_n}{k(k-1)} \quad (8)$$

Non-Pairwise Evaluation

In this paper we are not using Non-Pairwise evaluation. Kohavi-wolpert variance measure, interrater agreement and entropy measure can be used if needed.

C. Merit Calculation

Now the step is to combine the P and Q_x to calculate each ensemble's PQE merit value. The PQE merit is given as:

$$merit_{PQE^x} = \frac{k * \mathrm{P}}{\sqrt{k+k(k-1)(1-Q_x)}} \quad (9)$$

VI. EXPERIMENTAL RESULTS

A. Dataset Used

The datasets contain product reviews and metadata from Amazon [18] [19]. Eleven different categories have been chosen for the experiment. The experiments are done on these eleven datasets. These datasets contain review text, rating with some other information about users and others. The ratings are used for labelling of each instance as a positive and negative opinion. The instances having rating 3 or more labelled as positive and instances having 1 or 2 rating is labelled as negative.

B. Data Pre-processing

During pre-processing steps, the datasets are found skewed having a large number of positive instances than negative instances. To solve this problem a sufficient number of instances of positive and negative instances have been selected from the dataset on the first come first serve basis. After selection, each dataset contains approximately 25000 total reviews in which 15000 positive reviews and 10000 negative reviews. To create the training and test dataset, a document term matrix has been created. During the creation of document term matrix stop-word removal techniques has been used. Using TF-IDF algorithm 500 text features have been generated and used in document term matrix creation. For feature selection, unigram and bigrams are considered.

C. Experimental Setup

9different kind of classification models are used to create ensemble classifier. The implementation of Ant Colony Optimization algorithm is done in Python3.0 and Scikit-learn library is used for classification model creation.

D. Results

The results found in this experiment are very promising and very motivational for use of this framework in further research work. To show that how this framework adopts in between different domains the results are evaluated on eleven different datasets having reviews for different kind of products. To evaluate the performance of the given framework the comparison done on three different levels. Firstly, the performance of this framework is compared with the individual performance of its base classifiers. The next comparison done with the popular ensemble classifiers techniques.

Figure 2 compares the performance of ECS based classifiers with its base classifiers. This comparison is required to show that the ensemble classifier designed by this framework outperform its base classifiers. From figure2 it is shown that the ensemble classifier designed by this framework outperforms each of its base classifiers with respect to accuracy. Comparing to all of the base classifiers in base classifier pool, Support Vector Machine and Logistic Regression performs better than all other base classifiers. So, the effect of these two classifiers is huge in each of the ensemble classifier developed by ACO based ECS framework. SVM and LR are the part of each ensemble classifier designed by this ECS framework.

Figure3, provides a comparison between ECS based Ensemble classifier with some other traditional ensemble classification algorithms. In traditional ensemble classifiers,

we have chosen Begging, Adaboost and Random forest Classifier. As we have already seen in figure2 that Support Vector machine and Logistic regression provides better results than any other base pool classifier. So, the begging was implemented twice with different base classifiers. The first time the performance of begging is evaluated with SVM

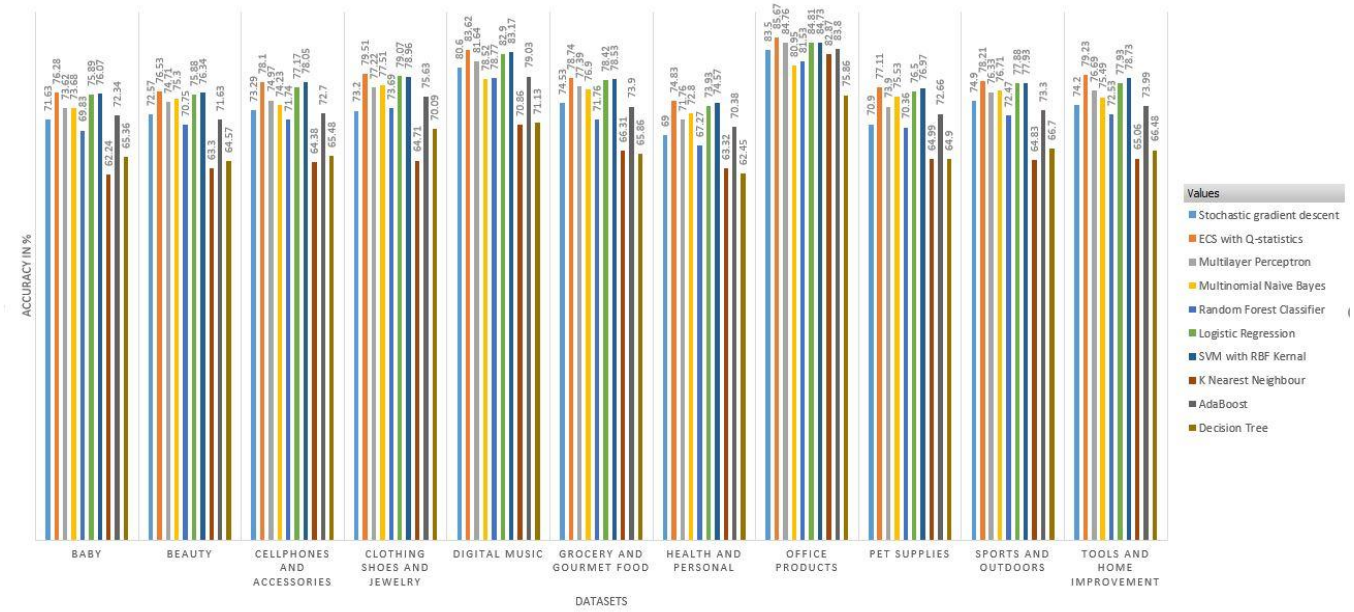

Fig. 2. Performance comparison between ECS based ensemble classifier and its base classifiers

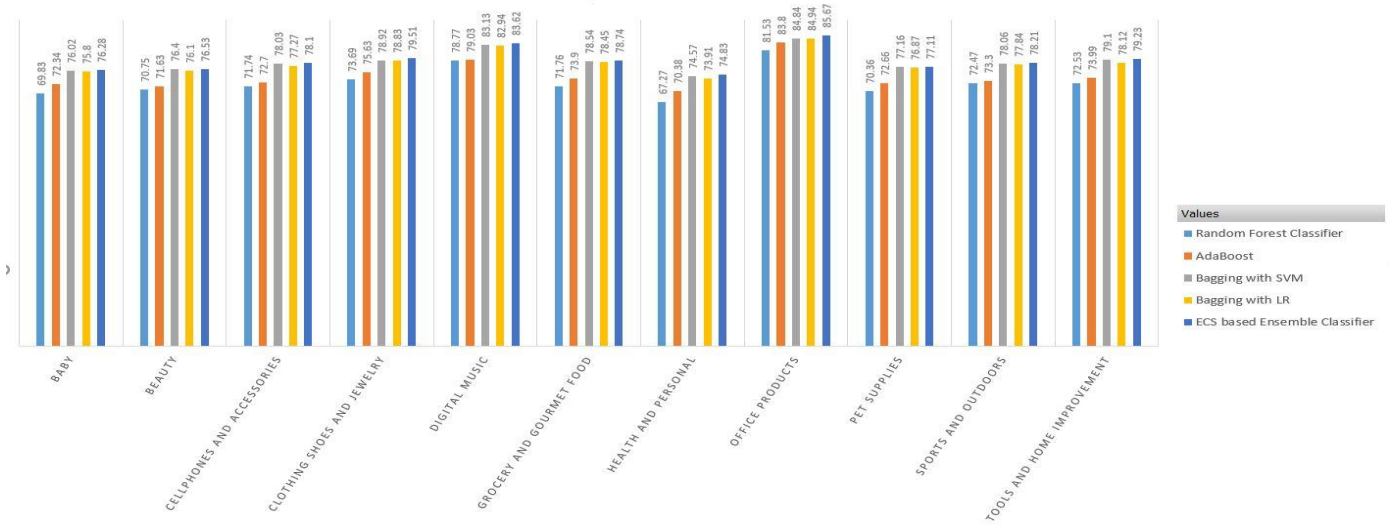

Fig. 3.Performance comparison between traditional ensemble classifiers and ECS based ensemble classifier

as a base classifier. And then begging again evaluated with logistic regression as a base classifier. AdaBoost isimplemented using Decision Tree Classifier as its base classifier and 50 estimators are used. Random forest classifier is implemented with 10 estimators and Gini index is used to measure the quality of a split.From figure 3 we can see that the ensemble classifier developed by ECS framework outperform all other ensemble classifiers in 10 out of 11 datasets implemented on. After all these results it is clear that ECS framework completes its task of developing an ensemble classifier with best possible performance very efficiently and successfully. This framework can also be used upon other datasets and results can be evaluated.

VII. CONCLUSION

The development of Ensemble classifier which provides the best possible results is a popular research area these days. A huge amount of research is going on these days for Ensemble classifier creation. On the other side due to the social media, online shopping like websites the amount of text data is available is huge. So, the researchers are trying their best to analyse these huge reviews and opinion text datasets. Due to the text data is so unstructured; analysing it becomes a tedious task.

In this paper, we present a framework which provides an ensemble classifier which gives the best possible results comparing to other available methods. Ant colony optimization technique searches between best possible ensembles very efficiently. ECS framework shows a promising way to solve the problem of ensemble classifier selection problem. In the future, non-pairwise diversity measures can also be used in this framework.

REFERENCES

1. Zhi-Hua Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press Taylor & Francis Group, 2012
2. Alexey Tsymbal a, MykolaPechenizkiy b, adraig Cunningham, Diversity in search strategies for ensemble feature selection, Information Fusion 6 (2005) 83–98
3. E. K. Tang · P. N. Suganthan · X. Yao, An analysis of diversity measures, Mach Learn (2006) 65:247–271
4. Liu, B. Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
5. Zhang L., Liu B. (2017) Sentiment Analysis and Opinion Mining. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA
6. Marco Dorigo, Mauro Birattari, Thomas Stutzle , Ant Colony Optimization Artificial: Ants as a Computational Intelligence Technique, IEEE Computational Intelligence Magazine · December 2006,28-39
7. J. Kittler, M. Hater, and R. P. W. Duin, "Combining classifiers," Proc. - Int. Conf. Pattern Recognit., vol. 2, no. 3, pp. 897–901, 1996.
8. E. Fersini ,E.Messina,F.A. Pozzi, Sentiment analysis: Bayesian Ensemble Learning, Decision Support Systems 68 (2014) 26–38
9. Ahmed Al-Ani, "Ant Colony Optimization for Feature Subset Selection",International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:4, 2007, pg. 999-1002
10. Yijun Chen, Man Leung Wong, An Ant Colony Optimization Approach for Stacking Ensemble, 2010 Second World Congress on Nature and Biologically Inspired Computing Dec. 15-17,2010, 146-151.
11. R. Geetha, G. Umarani Srikanth, Ant Colony Optimization in Diverse Engineering Applications: An Overview, International Journal of Computer Applications (0975 – 8887) Volume 49– No.17, July 2012
12. Xiaodong Zeng, Derek F.Wong, and Lidia S. Chao, "Constructing Better Classifier Ensemble Based onWeighted Accuracy and Diversity Measure", The Scientific World Journal, Volume 2014
13. G. Yao, H. Zeng, F. Chao et al. "Integration of classifier diversity measures for feature selection-based classifier ensemble reduction", Soft Comput (2016) 20: 2995.
14. Liying Yang, Classifiers selection for ensemble learning based on accuracy and diversity, Procedia Engineering 15 (2011) 4266 – 4270
15. R. He, J. McAuley , Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, WWW, 2016
16. J. McAuley, C. Targett, J. Shi, A. van den Hengel ,Image-based recommendations on styles and substitutes, SIGIR, 2015
17. Xu-Cheng Yin, Chun Yang, Hong-Wei Hao, "Learning to Diversify via Weighted Kernels forClassifier Ensemble, IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI),2014
18. Diao R, Chao F, Peng T, Snooke N, Shen Q (2014) Feature selection inspired classifier ensemble reduction. IEEE Trans Cybern 44(8):1259–1268
19. Nikunj C. Oza.,Ensemble Data Mining Methods, In Encyclopedia of Data Warehousing and Mining, pp. 448–453, Idea Group Reference, 2006

AUTHORS PROFILE

Mr. Sanjeev Kumar is currently working for his Ph.D. from M.J.P. Rohilkhand University, Bareilly, India. He received his Bachelor Degree in Science in 2007 at M.J.P. Rohilkhand University, Bareilly, India, and completed M.C.A in 2010 from Uttar Pradesh Technical University, Lucknow, India. Thereafter he has done M.Tech. in Computer Science and Information Technology from Invertis University, Bareilly, India. He has published several research papers in various conferences and journals. His research interest in Machine learning, Big Data, Artificial intelligence, Deep Learning, Algorithms and Natural Language Processing etc.

Dr. R. Singh is a **Professor** of Computer Science and Information Technologyat MJP Rohilkhand University Bareilly (India).Hereceived a Bachelor Degree in Electronics Engineering in 1991 and earned his Ph.D. in Computer Science and Engineering from Lucknow University, India. After completing his degree, he has worked in differentLaboratories/Divisions of U.P. State Government undertaking from 1991 to 1997. Thereafter Dr. Singh switched to academia and joined as an Assistant Professor in the CS & IT department of MJP Rohilkhand University, Bareilly (India) in 1997. At Present, he has been a Professor since 2011 in the same state Government University. He has been a doctoral supervisor at MJP Rohilkhand University, Bareilly, U.P. Technical University, Lucknow, etc. since 2005. His research interests lie in Wireless communication, mobile ad-hoc network architecture, wireless routing protocol, wireless medium access control, routing in under water and terrestrial sensor network, admission control schemes for Real time communication over the Internet, simulation and performance evaluation of wireless networks and Task allocation & routing in parallel systems and currently working in the field of Data Science. He has published more than 70 papers in journals and conference proceedings and he has also authored a number of books and chapters with prestigious publishers. He has been a Member of ISCA, USA, Computer Society of India and Institute of Electronics & Telecommunication Engineers, New Delhi (India).