ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

AN EFFICIENT MULTI-CLASS EXPLORATION OF ENSEMBLE CLASSIFICATION USING GENETIC ALGORITHM

GURURAMASENTHILVEL. P[*1], DR. G.GUNASEKARAN[*2]

[1]Research Scholar, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli - 627012, India, gurupandian@gmail.com

[2]Principal, JNN Institute of Engineering, Anna University, Kannigaipair, Chennai-601102, India, gunaguru@yahoo.com

Abstract: Pattern mining is the process of extracting and identifying patterns in the unstructured data from different sources. The existing methodologies are focused on the single class ensembles with individual parameters. The main objective of the proposed methods is to analyze and identify multiclass ensemble with different perspective of exploration. This multiclass ensemble is classified with multilayered approach in order to achieve high efficiency. The proposed method has been implemented by using evolutionary computing method. The genetic based classification performs more accurate result when compared to existing method of classification. There are various steps of genetic based method are considered such as selection, crossover and mutation process. The ensembles are classified in two layered approach namely local classification and global classification. In local classification the ensembles are maintained in separate clusters whereas global classification performs multiclass classification with single set. This method considered various parameters from diverse sources with high efficiency.

1. INTRODUCTION:

Evolutionary Algorithms (EA) are derived from biological which is used to map the problem with suitable solution domain in the computer science and its application. In any kind of research domain needs the analysis with the help of Genetic algorithm (GA) which is a subset of EA. The application of those algorithms is optimization and solution space based search. The main operation of the GA is a selection process, crossover process and mutation process. The better solutions are analyzed and identified based on the suitable problem which is identified from the sample space called population in GA.

The candidate solutions are selected from the sample set of population by achieving optimized solution. Suppose if there is any new properties added and already existing properties are removed from the sample set which will be done by using mutation process. Tradition al solutions are encoded using binary strings and also with other encoding methods with efficient manner. The generation of the population starts with individual candidate by using iterative manner. The individual are selected from the solution space called population is by calculating the fitness function in each step of generation process. This function produces the optimized solution form the values of objective function which is related to the problem domain.

The new generation is generated from the existing population in which the properties are modified or added. The currently generated solution set is a input to the another iterative process. The GA process gets stops if enough number of solutions (i.e., candidate) are produced from the population.

Normally the GA needs two types of domain they are solution domain and fitness value. The solution domain is the space which is called population and fitness value is the suitable candidate with solution. The representation of the sample solution is a array based and also use some other common notation. There are static and dynamic way of handling the candidate which will produced the accurate set by using the combination of the properties called crossover.

Normally the GA starts with the initial sample set with the representation of its properties using the calculated fitness value over its population. This process is looping procedure with the GA operations such as process of selection, process of mutation, crossover process and inversion process.

2. RELATED WORK:

2.1 INITIALIZATION

The problem of the real time area only decides the size of the population with different kinds of properties which are related to the solution set. Normally the initial set is selected with various algorithms and properties which suitable solution set or space. The main objective of the GA is to finding out the optimal solution from the problem area.

2.2 SELECTION

The individual is selected from the existing set and the suitable candidate is selected by calculating the fitness function which produces the suitable new candidate or solution. From the populations with fitness-based process, in which fitter solutions measures with function of selection. Some selection methods have a technique to the fitness value is calculated from the corresponding function for selecting and assessing solution which are related problem domain with best solution. Previous techniques of rating process are time consuming process with sample of the population. The quality and efficiency of the solution is represented by using the function and metrics which are related to genetic nature of the problem domain.

2.3 CROSSOVER AND MUTATION

The new generations of the candidates are produced by selecting and integrating the different attributes and its combination using process of mutation with crossover. The newly generated solution is mixed with already existing solution domains which are related to common or relevant problem domain. Various properties and attributes are shared among the candidates who are generated from the previous iterations by using the generic GA process.

The new head of the candidate called parent from newly create the candidates is selected with diverse sizes produces the dynamic population generation process. The same process is repeated for selecting more number of parents with high level cohesion in order to achieve high efficiency and performance during evaluation process. The different properties are identified from initial stage of generation for generating new set of properties called new population with related process. The fitness value of the newly formed population is analyzed and calculated the mean value which decides the overall efficiency of the generation process. The fitness values are categorized as most suitable and least suitable based on the threshold identified from the mean value. There are two set of populations are generated over the identified category.

The properties and attributes identification related to problem domain based on the probability of sample space with suitable size with the process of GA. The accuracy of the mutation process based on the rate of the recombination process is analyzed without any complications. There are two types of value is exists maximum and minimum. The maximum rate achieves the high accuracy in the GA whereas minimum rate targets the low level accuracy.

3. HEURISTICS

Most reliable way of handling GA operators and its processes are implemented with heuristics methods. The same generation with similar heuristics and different heuristics suffers a problem of candidate combination. The minimum convergence of the different sample space protects the early combination process.[6][7]

4. PROBLEM DEFINITION:

Nowadays the data are growing enormously from diverse sources with various kinds of devices. These data are stored in to the high data centre with maximum classification. The different data are secure and validated importantly by educational institutions, business companies and scientific organizations [1]. These data are handled and accessed by using the web based application. The data are categorized as social networking data, online database, and other text information. The vital problem of handling and retrieved the data using the pattern which are selected during the mining process. There are plenty of modern data mining tools are available which are not suitable for handling such type of raw data [2].

Pattern mining is a process of extracting the suitable data or information from various sources based on the region of interest. The computational intelligence are used for targeting the patterns with clustering and classification techniques [3]. The knowledge is extracted using various mining techniques such as document summarization, process of clustering and classification process [4]. The data are categorized in to semi-structured, unstructured, natural language processing [5]. These techniques are applied over the different real time applications such as web retrieval engine, CRM based application, spam filtering, recommendation systems intrusion system based detection, analysis with various categories [6].

5. ALGORITHM

5.1 Algorithm 1:

Algorithm for multi-class text or label of ensemble classification (MCEC)

```
begin
  data collected from various sources as DC; for each d ε DC do
begin
  preprocess the data using data cleaning; ensembles are identified from the preprocessed
                                                  data as EI
  for each e ε EI do
 begin
  select the ensembles as ES;
   for each es ε ES do
 begin
  classify the ensemble and form a classification  set CS;
  end; end;
  for each cs ε CS do
  begin
   select the ensemble from $CS_i$ for each ensemble ε $CS_i$ do
begin
  integrate all the ensembles as one global cluster as gc;
   end
   end
   call genetic based ensemble management (gc);
generate the report for all class of ensembles;
end
end
```

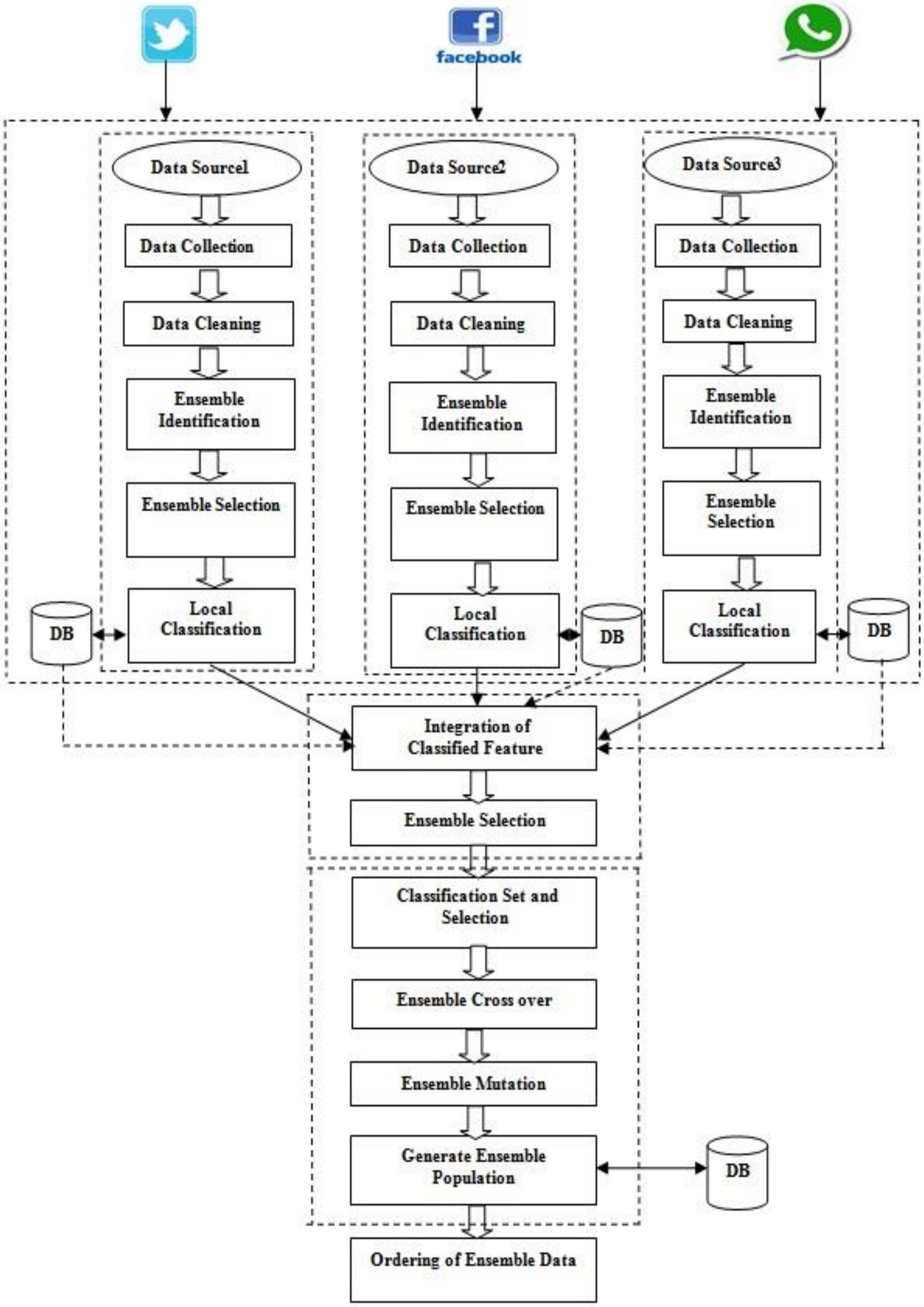
5.2 Algorithm 2:

Algorithm for genetic based ensemble management (Global Cluster GC)

```
begin
  for each gc ε GC do
begin
  select the ensemble from the set as ES;
  calculate the fitness function for the ensembles
  as f(x);
  for each e ε ES do
begin
  calculate the fitness function of all ensemble e; f(e) ;
if f(e) >= f (x) then
collect those ensemble as ensemble set as ESS; end
 for each es ε ES do
begin
 for each es ε ESS do
begin
  perform cross over create set as ESS; end
end
 generate new ensemble set with class Labels as
NES;
```

```
return NES;
end; end
```

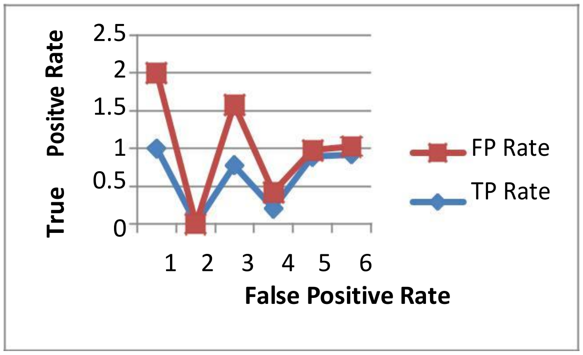
6. Figure 1: WORKING MODEL OF EGA



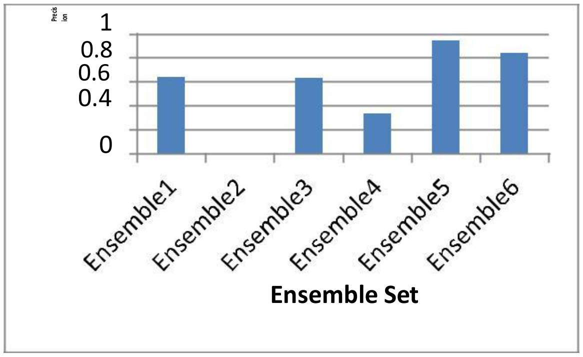
The data are collected from various social networks such as twitter, face book and whatsapp etc. The data collection process performs the operations with integrated data. The data cleaning process removes the irrelevant data from data collections. The ensembles are identified from the collection set. The ensemble set are preprocessed and categorized as a label and then performs the selection process. The local classification is done by suitable ensemble set. The individual set has different ensemble, which are stored in database. The integrated feature classification process has been carried out and maintained as a single set. This classification set selection process identified ensemble as a new population. The proposed system of classification follow genetic based method. The ensembles are selected with fitness functions. The selected ensembles are crossed over by generating new set. The mutation process also done by constructing new population. These ensembles are maintained in database of with ordering of ensemble data set. Figure 1 shows that the proposed method of ensemble classification.

7. PERFORMANCE EVALUATION

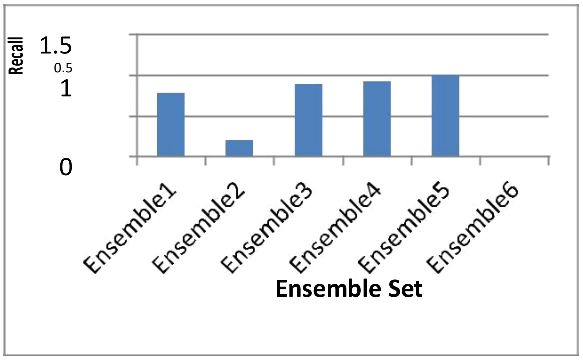
The performance evaluation is done based on the ensembles collected from various sources with difference ensemble set. The performance parameters like true positive rate, false positive rate, precision, recall and F-Measure are considered for evaluation. The comparison takes place in two levels namely local and integrated. In local level the classification are done and maintain only restrictive set because the data are collected from specific source. The performance is restricted to only particular data set. In integrated classification perform over different types of ensembles which are maintained during local classification process. The genetic algorithm based comparison also mapped with extraction process in order to achieve the maximum efficiency. Figure 2 a) shows that the comparison of True positive Rate and False Positive Rate. Figure 2 b) describes that the precision comparison. Figure 2 c) represented as recall comparison. 2 d) shows that the F-Measure comparison. The overall performance of local classification is shown in figure 2. The integrated classification with performance comparison is represented in figure 3.



a) True positive Rate Vs False Positive Rate

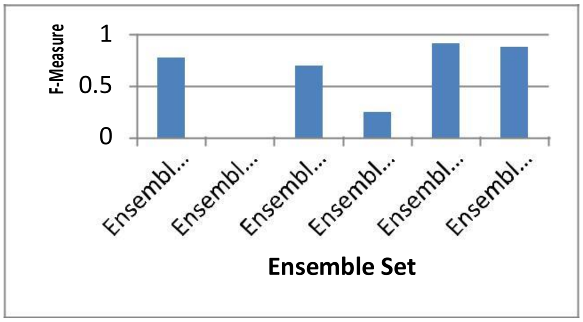


b) Precision comparison



c) Recall Comparison

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515



d) F-Measure Comparison

Figure 2. Local Classification Performance Comparison

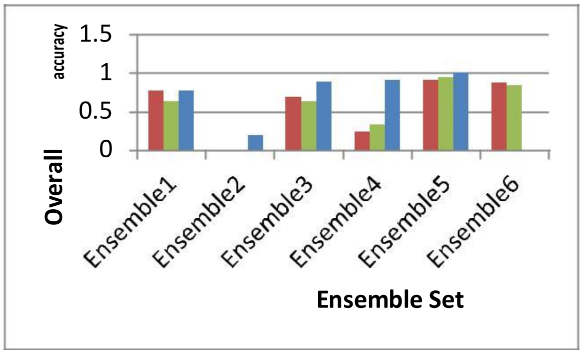


Figure 3. Integrated Classification Performance Comparison

8. CONCLUSION AND FUTURE WORK

Traditionally data are produced only particular source and most of the customer has consumed the data, so there is no problem during the data handling. Nowadays the data are generated by various sources with huge in number, so it needs proper handling with efficient Method. There is no reliable solution exists while handling single source data for prediction. The above problem is overcome by implementing proposed ensemble classification method. This performs two classification processes with genetics algorithm as a base local classification and integrated classification process. The new ensembles are identified and replace the existing ensembles by using the performance metrics. The main objective of the proposed method is to classify the ensembles more accurate without any false information to the customer. In future these methods can be extended to scientific and medical domain for generic solution.

REFERENCES:

[1] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

[2] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.

[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[6] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.

[7] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," Copy at http://j. mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper, vol. 456, 2014.

[8] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.

[9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius,ˇ and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.

[10] B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.

[11] C. P. Chen and C.-Y. Zhang, "Data- intensive applications, challenges, techniques and technologies: A survey on big data," Information Sciences, vol. 275, pp. 314–347, 2014.

[12] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," International Journal of Computer Applications, pp. 159–171, 2013.

[13] P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on. IEEE, 2015, pp. 634–638.

[14] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45.

[15] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp.