ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2017.0217



# ENRICHMENT OF ENSEMBLE LEARNING USING K-MODES RANDOM SAMPLING

**Balamurugan Mahalingam[1], S. Kannan[2] and Vairaprakash Gurusamy[3]**
*Department of Computer Applications, Madurai Kamaraj University, India*

***Abstract***

***Ensemble of classifiers combines the more than one prediction models of classifiers into single model for classifying the new instances. Unbiased samples could help the ensemble classifiers to build the efficient prediction model. Existing sampling techniques fails to give the unbiased samples. To overcome this problem, the paper introduces a k-modes random sample technique which combines the k-modes cluster algorithm and simple random sampling technique to take the sample from the dataset. In this paper, the impact of random sampling technique in the Ensemble learning algorithm is shown. Random selection was done properly by using k-modes random sampling technique. Hence, sample will reflect the characteristics of entire dataset.***

***Keywords:***

***Sampling, Ensemble Classifiers, Cluster Random Sample***

## 1. INTRODUCTION

Ensemble classifiers aggregate the classification models into single model to perform the prediction on new instances. Usually ensemble is constructed by either dependent ensemble framework or independent ensemble framework [1]. Classifiers are sequentially processed in dependent framework, that is output of one classifier is determined by output of previous classifier. In Independent framework, all classifiers are processed in parallel manner. In proposed work, the boosting algorithm from dependent framework and random forest from independent framework is chosen. Sample is taken as input for the ensemble algorithm to build the classification model. Quality of samples is used to build a good prediction model [15]. Simple random, Systematic and Stratified are traditional sampling techniques which uses the randomization concept to perform the sampling process. These techniques are failed to give a quality of sample. Because there is a possibility of choosing biased samples.

The new k-mode random sample technique overcomes the problem of traditional sampling techniques. It is the combination of k-modes cluster algorithm and simple random sampling technique [2] [3]. In this technique, first grouping of the class is done. The groups are then divided into clusters by applying the k-modes algorithm on each group.

The Next step is to extract the equal number of sample from each cluster by applying the simple random sampling technique on each cluster. Finally merge the all samples into single set. That single set is the training dataset for the ensemble classifiers to build the better classification model [10]. Remaining instances of dataset considered as the testing dataset. This testing dataset is used by the prediction model of ensemble classifiers to perform the classification based on known class labels.

The rest of this paper is organized as follows: In section 2, Ensemble algorithms are explained. Section 3 describes the Random sampling techniques. Section 4 defines the k-modes cluster algorithm. In section 5, the process of k-modes random sampling technique is explained. Section 6 shows the experimentation of k-modes random sampling technique with discussion. Section 7 concludes the paper with future enhancement.

## 2. ENSEMBLE ALGORITHMS

Two algorithms are taken for this research work. They are, boosting from dependent framework and random forest from independent framework.

### 2.1 BOOSTING

Boosting is one of the algorithm in dependent ensemble framework [4]. Given dataset is divided into training dataset and testing dataset. Whole training dataset (sample) is taken as input for ensemble classifier to build the prediction model. Initially null weight is assigned to each classifier. In this algorithm, only one classifier can build the model in each iteration [14]. Each classifier is trained based on building the classification model by classifier in previous iteration. If the classifier performs poor classification, assign the larger weight to the classifier. Otherwise classifier gets the small weight. Finally, choose the classification model of classifier which has the smallest weight among the all classifiers.

### 2.2 RANDOM FOREST

Random forest algorithm has the independent ensemble framework [5]. Sample is taken from the original dataset. This sample is the training dataset. Bootstrap samples are taken from the training dataset by applying randomization concept. Each sample has the random instances. This bootstrap samples feed to each classifiers to perform classification [7]. While growing the forest, each classifier chooses the subset of features from bootstrap sample features to split the node in tree. After building the prediction model, choose a classification model which has the more number of votes by classifiers among the classification models.

## 3. RANDOM SAMPLING TECHNIQUES

Generally random sampling is the process of extracting the sample from dataset in random manner [9]. Sample has the random number of instances. That is, all instances has equal probability to select. So that, bias may be avoided in the sample. There are three main random sampling techniques are,

- *Simple Random Sampling:* Samples are randomly extracted from the dataset based on the given size. Each instance of the original dataset is having the equal probability.

- *Systematic Random Sampling:* Randomly choose one instance from the original dataset. After that, choose the consecutive fixed interval of elements next to the random instance.
- *Stratified Random Sampling:* Dataset is divided into groups based on the class labels of target attribute. Groups are called as strata. Samples are extracted from each stratum by applying the simple random sampling technique on the each stratum. Finally merge the all samples of strata [8].

## 4. K-MODES CLUSTER ALGORITHM

K-modes cluster is updated algorithm of K-means cluster algorithm for clustering the categorical data [11]-[13]. It uses the frequency based method to update the mode. Initially, set the different attribute value of an instance as mode for k number of cluster. Then the value of selected instance is compared with value of all other instances by using simple matching distance. If the one attribute value of instance is equal to attribute value of another instance, frequency is counted.

Each and every iteration mode is updated as which value has more number of frequencies and also allocate the instances whose frequency is nearest to the frequency of mode. This process is repeated until clusters are properly partitioned. In the output, clusters are externally heterogeneous and internally homogeneous.

**Advantages:**

- K-means cannot cluster the categorical data. But k-modes can do it.
- It is faster than other cluster algorithms because, k-mode cluster algorithm is partitioning the clusters with less iteration.

## 5. K-MODES RANDOM SAMPLING

K-modes random sampling is the new sampling technique used to extract the samples from the original data. It combines the k-modes cluster algorithm and simple random sampling technique. In this sampling, the proposed technique divides the original dataset into groups based on the number of classes in the target attribute. Each group may have homogenous records. But it fails to give exact homogeneous records. Hence, Groups are split into subgroups (clusters) by applying the k-modes cluster algorithm on each group. Dividing of groups depends upon the type of dataset. The reason behind this is to avoid the bias when forming the clusters.

While choosing the dataset, check whether the dataset is balanced or not. If the dataset is balanced, then the clusters are equally partitioned. If the dataset is unbalanced, then form the more number of clusters for which class label contains the many instances than others. Then equal numbers of samples are gathered from clusters by applying the simple random sampling technique on each cluster. After that, concatenate the samples from each cluster into one, which is named as training dataset. Remaining sample is named as testing dataset. The Fig.1 shows the workflow of cluster random sampling technique.

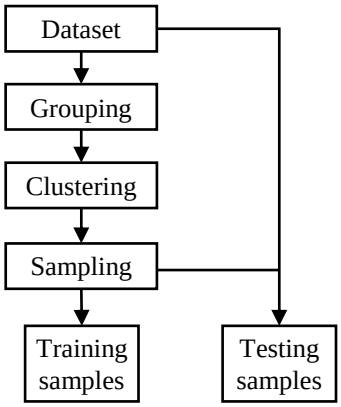


Fig.1. Workflow diagram of Overall process of K-modes random sampling

***Algorithm:***

**Step 1:** Read the dataset

**Step 2:** Split the dataset into a number of groups based on the class label.

**Step 3:** Again split each of the groups into clusters by applying the k-modes cluster algorithm on each group

**Step 4:** Take fifty percent of samples from each cluster by applying the simple random sample technique on each cluster.

**Step 5:** Merge all samples of clusters into one (training dataset) and set remaining samples as testing dataset.

**Step 6:** Give the training dataset as input to ensemble method

**Step 7:** Process the input and produce predictive model

**Step 8:** Give the testing dataset and predictive model as input to predict function and then it produces the result

**Step 9:** Finally, give the result of predict function and target classes of attribute as input to confusion matrix and it produce the accuracy.

## 6. RESULTS AND DISCUSSION

Samples of the dataset is trained and tested by ensemble classification. Based on the number of class distribution in dataset, there are two types of datasets commonly available. They are balanced and unbalanced dataset. All classes are equal in the balanced dataset. In unbalanced dataset, classes are unequal.

The Table.1 shows the two datasets in detail, one from balanced type and one from unbalanced dataset. In this work, extract the 80% of training samples and 20% testing samples from dataset by using the k-modes random sampling technique. Then give this training sample as input to random forest and boosting algorithm. In these algorithms, fixed number of classifiers is trained in different manner.

For random forest, classifiers are trained in parallel. It is then combined with the prediction model which has majority vote by classifiers. In boosting algorithm, classifiers are trained in serial manner and the prediction model is chose which correctly performs the classification among the classifiers. After the trained process, prediction model classifies the new instances of testing

samples. Finally performance is measured by Confusion matrix that produces the accuracy of the prediction model.

Table.1. Datasets

| Datasets | Classes | Description |
|---|---|---|
| Carseats (400 instances and 11 attributes) | High – 164<br>Low – 236 | It contains the instances about sales of child car seats at 400 different stores. |
| Iris (150 instances and 5 attributes) | Setosa – 50<br>Virginica – 50<br>Versicolor - 50 | It consists of species of iris plant. |

Confusion matrix is a two dimensional table with actual and predicted values. This contains the 4 features. They are number of false positives, false negatives, true positives and true negatives. For example, true positive means actual result is true and predicted decision is also true. In this experiment, one time execution does not define the accurate performance. Because, sampling process is done as random. So each time there may be a possible change in input. To overcome this problem, each algorithm is experimented ten times. Their accuracies are averaged. The result of average is taken as the final output. In this work result is defined by average of ten time experiments

Table.2. List of Accuracy values of ensemble algorithms for unbalanced dataset (Carseats)

| Sampling | Random Forest | Boosting |
|---|---|---|
| Simple random (SRS) | 79.7 | 82.1 |
| Systematic (SYS) | 75 | 83.5 |
| Stratified (SS) | 80.2 | 85 |
| K-modes CS | 85 | 89.3 |

Table.3. List of Accuracy values of ensemble algorithms for balanced dataset (Iris)

| Sampling | Random Forest | Boosting |
|---|---|---|
| Simple random (SRS) | 86 | 84.5 |
| Systematic (SYS) | 84.3 | 82.3 |
| Stratified (SS) | 89.7 | 91 |
| K-modes CS | 94 | 92.3 |

In the existing analysis, none of the algorithm is suitable for all type of dataset and ensemble algorithms [8], i.e. stratified sampling technique gives its better performance only for unbalanced dataset and simple random sampling for balanced dataset. Because of inaccurate homogeneous samples are generated by stratified sampling. So these techniques may introduce bias in samples of unbalanced dataset. If bias occurs during sample selection, sample fails to reflect the original dataset.

For this reason, the K-modes random sample technique is introduced for getting the exact samples from the dataset. K-modes algorithm has less iteration than other cluster algorithms. So those cluster sampling uses the k-modes for partitioning the data. While clustering the data, bias can be avoided.

The Table.2 has shown the performance of random forest and boosting with various sampling techniques on unbalanced dataset. From the Table.2, the new sampling gives the best performance than other sampling techniques. In this unbalanced dataset, class label low sales having the large number of instances than high sales. So that, three clusters are formed for low sales group and two for high sales group. This is to reduce the bias in selection of samples.

The Table.3 depicts the accuracy values of random forest and boosting with different sampling techniques for balanced dataset. In Table.3, K-modes random sample technique has the highest peak value than other sampling techniques for both random forest and boosting ensemble algorithms. In the balanced dataset, groups are equal to one another. There is no need to partition the group into different number of clusters. All groups are split into equal number of clusters by using k-mode algorithm.

For both balanced and unbalanced dataset, k-modes random sampling technique can extract the good quality of sample. Because, K-mode algorithm is partitions the exact similar and dissimilar clusters with related instances. From the discussion, K-modes random sample provides the flexibility to both balanced and unbalanced dataset. It can reduce the bias in sample.

## 7. CONCLUSION

Quality of sample could improve the accuracy and performance of ensemble classifiers. Existing random sampling techniques fails to extract the accurate samples from both balanced and unbalanced dataset. That is, stratified gives better performance for unbalanced dataset and simple random sampling technique for balanced dataset. So there may be bias in sample selection. K-modes random sampling technique is introduced to overcome the problem of existing sampling techniques. It is the combination of k-mode cluster algorithm and simple random sampling technique. In this theoretical and experimental research, k-modes random sampling technique gives better quality of samples to ensemble classifiers than other sampling techniques. The present work can only improve the sampling process in ensemble classifiers. Still there is in need of improve the prediction performance of ensemble algorithms. Moreover there are various ways to improve the accuracy and performance of ensemble classifiers. In future, research may be focused in diversity generation and ensemble selection.

## REFERENCES

[1] Lior Rokach, "Ensemble-based Classifiers", *Artificial Intelligence Review*, Vol. 33, No. 1-2, pp. 1-39, 2010.

[2] Robi Polikar, "Ensemble based Systems in Decision Making", *IEEE Circuits and Systems Magazine*, Vol. 6, No. 3, pp. 21-45, 2006.

[3] Robert E. Schapire, "The Strength of Weak Learnability", *Machine Learning*, Vol. 5, No. 2, pp. 197-227, 1990.

[4] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *Annals of Statistics*, Vol. 29, No. 5, pp. 1189-1232, 2001.

[5] Tin Kam Ho, "Random Decision Forests", *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1, pp. 1-6, 1995.

[6] Vrushali Y. Kulkarni and Pradeep K. Sinha, "Random Forest Classifiers: A Survey and Future Research Directions", *International Journal of Advanced Computer Technology*, Vol. 36, No. 1, pp. 1144-1153, 2013.

[7] Leo Breiman, "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[8] M. Balamurugan and S. Kannan, "Analyse the Performance of Ensemble Classifiers using Sampling Techniques", *ICTACT Journal on Soft Computing*, Vol. 6, No. 4, pp. 1293-1296, 2016.

[9] William G. Cochran, "*Sampling Techniques*", John Wiley and Sons, 2007.

[10] Iain A. Macdonald, "Comparison of Sampling Techniques on the Performance of Monte-Carlo based Sensitivity Analysis", *Proceedings of 11th International Building Performance Simulation Association Conference*, pp. 992-999, 2009.

[11] James D. Nelson and Robert C. Ward, "Statistical Considerations and Sampling Techniques for Ground-Water Quality Monitoring", *Ground Water*, Vol. 19, No. 6, pp. 617-626, 1981.

[12] Zhexue Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values", *Proceedings of 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 1-14, 1997.

[13] Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", *Proceedings of International Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 1-8, 1997.

[14] Guojun Gan, Zijiang Yang and Jianhong Wu, "A Genetic K-Modes Algorithm for Clustering Categorical Data", *Proceedings of International Conference on Advanced Data Mining and Applications*, pp. 195-202, 2005

[15] Rushi Longadge and Snehalata Dongre, "Class Imbalance Problem in Data Mining Review", *International Journal of Computer Science and Network*, Vol. 2, No. 1, pp. 1-6, 2013.