ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Recursive Ant Colony Based ECOC: An Ensemble Learning Technique for Classifying Data

Deepak Rajak[1], Roopam Gupta[2], Sanjeev Sharma[3]

M. Tech Research Scholar, School of Information Technology, UTD, RGPV Bhopal, M.P, India[1]

Associate Professor, School of Information Technology, UIT, RGPV Bhopal, M.P, India[2]

Associate Professor, School of Information Technology, UTD, RGPV Bhopal, M.P, India[3]

ABSTRACT: Error correcting output code (ECOC) is one of the widely used classifier ensemble technique .That technique provide solution for the various multiclass classification problem by dividing multiclass problem into binary class classification problem. In this paper, a new enhanced heuristic coding method, based on ECOC, RACS-ECOC is proposed. To generate strong classifiers for the multiclass classification problem following three steps are used. The first step starts with layered clustering-based approach. The approach can split a hard to classify multiclass problem in to subclasses and construct multiple different strong binary class classifiers on a given binary-class problems, so that the heuristic training process will not be stopped by some difficult binary-class problems. Second a weight optimization technique is used to optimise new classifiers for that problem. For that purpose a recursive ant colony algorithm is used to optimize classifiers. It optimize weights in a way that ensures the non-increasing of the heuristic training process whenever a new classifiers added to the ECOC ensemble. In third step these classifiers are updated to the classifier field to generate better classifier for the classification problem. Which provides a simple and efficient way to optimize weight and a KDD CUP intrusion detection data set is used for data mining. A performance Comparison of the proposed Recursive ant colony System-ECOC technique with existing WOLC-ECOC over the parameters like precision and recall is presented in results section.

KEYWORDS*:* Classifier ensemble, error correcting output codes, multiple classifier systems, multiclass classification problem, KDD cup.

I. INTRODUCTION

In Machine learning Support vector machines (SVMs),neural networks, genetic algorithm etc have been used to perform the classification task but in recent time over the last decade, classifier ensembles (i.e. Multiple classifier systems), such as bagging, boosting, and their mutations, have been shown to be effective approaches for solving, learning problems like classification, regression, etc. For such projects, the success of the classifier ensembles relies strongly on a good choice of the base learners and a strong diversity among the base learners. One of the well-known classifier ensembles for solving multiclass problems is the error correcting output code (ECOC) [3]. ECOC decomposes a.multiclass problem to a serial binary-class problems.

Each binary-class problem is figured out by some binary-class classifier, such as Ada Boost and support vector machine. Given a P class problem with a set of labelled samples {(ρi, Yi)} n i=1 where ρi is a d dimensional sample, and Yi ∈ {1,2,..., P} is the label of ρi, the ECO Ctriesto use Q binary-class classifiers to address this problem. There lation between the divisions and the classifiers can be extracted by a code matrix M ∈{−1, 0,1} P×Q.ECOC consists of two phases – coding and decoding. In this instalment we explore the actual coding process,

ECOC tries to find a code matrix M for the classifier training, where the p-th row of M expresses the code word of the p-th class, denoted as cp, and the q-th column expresses the q-th classifiers, denoted as hq. If the entry mp, q = 0, it means that hq does not take the data of the Pth class into classifier training [5]. In the decoding process, conducting a test sample ρ into h1,..., HQ successively can get a test code word of ρ, denoted as x = [x1,..., xQ] T.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Given a decodingstrategy f(x, cp), thepredictionof ρ canbeformulated as a minimization problem mincp∈Mf(x, cp), where M = {cp} Q p=1 is the codeword set. For the coding stage, there are generally two research directions for the code word design. The first direction is the problem-independent coding, design, such as the well-known one-versus-all (1vsALL) and one-versus-one (1vs1). The second direction is the problem-dependent coding, design, which looks more promising and has drawn much attention.

A new ensemble learning technique called RACS-ECOC (Recursive ant colony System based ECOC) is proposed in this paper. That technique provides enhanced functionality to classify data. In existing WOLC-ECOC (Weight Optimization and Layer Clustering Error Correcting Output Codes) is used to classify the data but that technique not efficient to provide better solution for the various classification problems. In WOLC-ECOC a complex CPA(Cutting Plane Algorithm) based technique is used which not provide better optimization of the classifiers, recursive ant colony provides better performance to optimize classifiers into the classifier matrix. A vrief description over the proposed technique is presented in Proposed Methodology section.

Further this paper is organized as follows:-

II. RELATED WORK, III. PROPOSE METHODOLOGIES, IV RESULTS V. CONCLUSION.

II. RELATED WORK

In this paper Xiao-Lei Zhang [1], propose a CPA based WOLC-ECOC technique, it initialize with ECOC. There are two steps are performed in that algorithm in first step strong dichotomizes are generate for most confusing pair of binary class problem and in second step it updates these dechotomizers by optimized weight decoding algorithm and steps repeat until all classes classifies accurately. In this way it converges some of the drawback of other techniques and perform best among these technique

In this paper A. Rahman and B. Verma [2] propose a cluster generating technique in which clusters of different dechotomizers are generated. In this model different level of clustering is presented for datasets. In that way datasets are trained over those clusters dechotomizers to classify data and generate different level of clustering for that dataset.

It is easy to classify binary-class data but to organize multi-class data; much technique fails to generate suitable classifiers for multiclass data. For this purpose Dietterich and Bakiri propose [3] an ECOC framework technique to classify multi-class data properly. it has two parts one is encoding and another one is decoding , in encoding stage it generates different classifiers for each class based on different binary problems and in decoding stage makes decision for classification for given data based on different output.

In [4], Pujol et al. proposed a problem-dependent coding method called discriminative ECOC (DECOC). It uses a binary decision tree for coding purpose, where each node of the tree is a strong bipartisan of a hard to classify problem, but the decision tree suffers with a drawback that if any node of the tree predict wrongly then descending nodes of that node also predict wrong data.

To overcome this drawback, in [5], Pujol et al. presented an ECOC optimizing node embedding (ECOC-ONE) algorithm. It initialize with an traditional ECOC classifier ensemble and iteratively updates binary-class classifier that classifies the most confusing pair of classes to the ECOC ensemble until get the desired results of classification. This technique remove the drawback of decision tree based ECOC ensemble and directly classifying the most confusing pair. But seldom is the most confusing pair in the classes hard to classify, in that case ECOC-one not efficient to generate classifier for that confusing pair.

For this problem, in [8], Escalera et al. to resolve drawbacks of the ECOC-ONE authors presented a technique which splits hard to classify classes in to different subclasses, in that a hard to classify problem converted into several easily classifiable problems. That technique provide a new way to classify hard to classify problems but subclass-ECOC also uses decision tree to make subclasses but in that case it hard to control the scale of subclasses for large scale problems. That arose a question, there is a solution can be formed A subclass- ECOC without using decision tree.

Optimized Weighted Decoding:

In [4], Escalera et al. Presented that a good decoding strategy should get each category has the same decoding dynamic range and zero decoding dynamic range bias. Then, they proposed the LW decoding algorithm.

Layered Clustering-Based Approach:

The layered clustering-based (LC) approach [1] is a special dechotomizerclassifier ensemble. It first splits the attribute space into several various clusters by clustering, in that case classification can performed separately within a cluster and classification problem separately solved by classifiers within cluster. And so that process performed several times,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

each time the process runs it form a layer. The LC approach contains two complementary techniques. One is the clustering in each layer. It can identify overlapping patterns that are difficult to distinguish. But it does not include any mechanism to integrate the diversity.
The other one is the layered approach. It utilizes the mechanism of the bagging and boosting to achieve the diversity between any layers for the helplessness of the first property. The layered structure, as examined in [9], Williams proves the discriminability of a classifier ensemble on a binary-class problem.
Layered Clustering-Based ECOC:
As examined in the introduction section, in [8], Escalera et al. proposed the subclass technique that splits a difficult classification problem into several easier sub-problems. Each sub problem is resolved by an independent classifier. Ultimately, the difficult problem is resolved by a classifier ensemble. Nevertheless, they used a tree structure for the splitting. In parliamentary law to inherit the advantage of the subclass technique, and in the meantime, to avoid using the decision tree for the subclass splitting, we investigate the ensemble learning for the solution as follows:
The central idea of the ensemble learning is to make a strong, diversified among the base classifiers. Broadly speaking, the methods of constructing the diversity can be grouped into four types [9].
They are the methods of
1) Manipulating the training examples,
2) Manipulating the input features,
3) Manipulating the training parameters, and
4) Manipulating the output targets.
However, aside from manipulating the outputs of the binary classifiers which is the key idea of ECOC, the diversity has been seldom referred to the ECOC study yet. To our knowledge, only in [10], prior and Wind eat manipulated different parameter settings of the base classifiers (multi-layer perceptron).

III. PROPOSED METHODOLOGY

There are various technique of ECOC ensemble but WOLC-ECOC perform best among of these techniques, in this paper a recursive ant colony based ECOC is proposed, RACS-ECOC gives more accurate results then previous one.
RACS Algorithm:
In existing ant colony system it is quite capable to handle small scale data or it perform well in small scale data but in case of large scale data it suffers in performance e degradation. To resolve that performance degradation a recursive ant colony system scheme [13] is presented which is just similar to recursive merge sort a divide and conquer technique. Because ant colony has random nature it is hard to classify large scale random data so a technique is requires constructing classification decision about large datasets in this algorithm data set divided in to smaller data sets and performing classification for data. For implementation purpose first it divide problem into several sub classes and performed classification and merge these results to optimize the classification before it get inactive.
Recursive ant colony Based Optimized ECOC scheme:
To take advantage from the capabilities of Recursive Ant Colony System (RACS) algorithm a recursive ant colony based ECOC algorithm is proposed RACS-ECOC, which also use to improve the performance of Ant Colony System algorithm for large scale Data access problems by partitioning the feasible data obtained and applying the algorithm recursively on smaller pockets to find good solutions for the problems by improving the exploration efficiency of the ants. And use for propose work in this paper to enhance the efficiency of WOLC-ECOC.

Algorithm for the proposed technique
Input: Dataset
Output: classifier generated by the RACS-ECOC technique

1. Initialize the ECOC process, Weight matrix and load datasets.
2. Apply LC-ECOC technique.
3. Select most confusing pair of classes or stubborn classes.
4. Split these classes into subclasses.
5. APPly ECOC techique over these subclassees.
6. Generate classifier for each subclasses.
7. Aggregate these classifiers to generate a classifier for the stubborn class.
8. Use recursive ant colony optimization technique to update these classifiers into the ECOC repository.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

9. Repeat step 3.
10. End.

Flowdiagram of the Proposed technique contains following steps:
Initialization: load dataset and use LC-ECOC ensemble. Dataset are loaded from the database and perform the classification to classify the dataset.

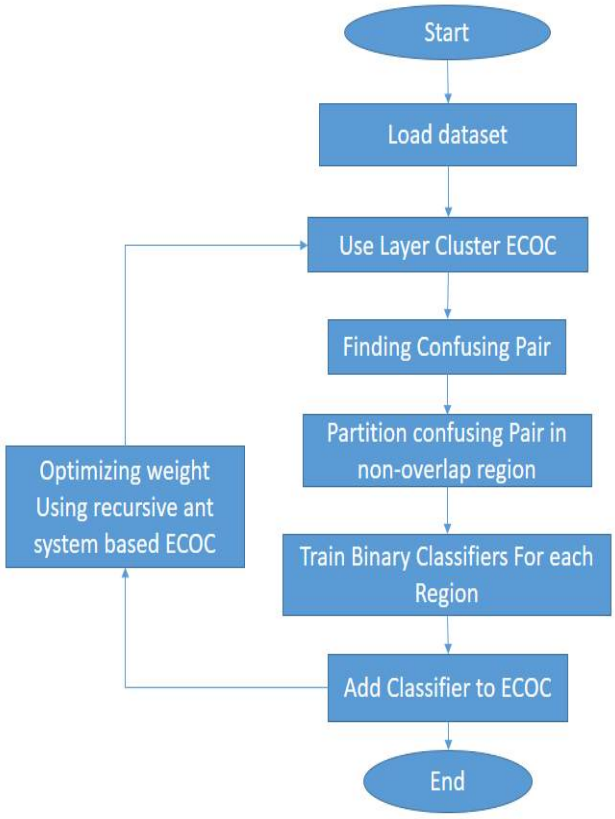

Figure 5.1: Flow Diagram for Proposed Technique

1. Start
2. Any valid ECOC (LC-ECOC): Layer Clustering ECOC an ensemble learning technique is applied to form various clusters of the datasets. There are different protocols are used to form clusters. Divide dataset into clusters of protocols like TCP, ICMP, and UDP.
3. Find confusing pair of classes: the process to form most confusing pair in the datasets, to classify these confusing pairs a technique is applied.
4. Partition confusing pair in non-overlapped regions: Partition that confusing pair into various non-overlapped regions and classification is performing on these partitions to classify the dataset.

5. Train a simple classifier for each region: training operation is performed over these regions to provide better classification for these datasets. By combining the classifiers of the each region a classifier for that confusing pair is generated.
6. Optimize weight using RACS system: by using RACS optimization technique these classifiers are updated into the database and for an updated classifier database for the classification.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

7. Repeat step 3 until risk converges: repeat the step 3 to classify most confusing pairs in the datasets.

IV. RESULTS

Evaluation of performance of the existing and proposed technique is presented in this section. To evaluate performance of the proposed technique Precision, Recall, Accuracy and detection rate are used as a parameters.

Evaluation parameters

Precision: - It is the ratio of no. of relevant records retrieved to the no of irrelevant records and total no of relevant records.

$$\text{Precision} = \frac{(Relevantdocument)(RetrievedDocument)}{RetrievedDocuments}$$

Recall: - It is theratio of total no. of relevant records is retrieved from the total records and relevant records.

$$\text{Recall} = \frac{(Relevantdocument)(RetrievedDocument)}{RelevantDocuments}$$

Accuracy: -It is the measure of how accurately the no of relevant records found from totalno. of records.

$$\text{Accuracy} = \frac{(Relevantdocument)(RetrievedDocument)}{Totaldocuments}$$

Detection rate: - It is the measure of how efficiently intrusion can be detected from the dataset**.**

Statistical comparison of the results

A Statistical comparison of the results is presented in Table 4.1. In that comparison of the performance of existing and proposed is presented. Which contains numerical values of accuracy, Detection rate; precision and recall compare the performance. That comparative analysis shows, proposed technique accurate results as compare to the existing technique. Provide accurate classification of the large scale datasets. That result analysis shows that RACS-ECOC is an accurate and efficient technique for classification of data.

Table 4.1: Statistical comparison of the proposed technique and existing technique.

Technique	Accuracy (in %)	Detection Rate (in %)	Precision (in %)	Recall (in %)
WOLC-ECOC	47.6	67.17	68.20	64.17
RACS-ECOC	62.38	78.15	76.24	74.14

In Figure 4.1, a graphical analysis is shown, curves of existing methods and proposed method shows that RACS-ECOC has the high detection rate, precision and recall. Accuracy is also high which shows that RACS-ECOC is an efficient and accurate technique to classify data.

Graphical analysis for the proposed technique

A graphical analysis for the proposed technique is shown in graph 7.1, which shows that proposed technique provides better results as compare to the existing technique.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

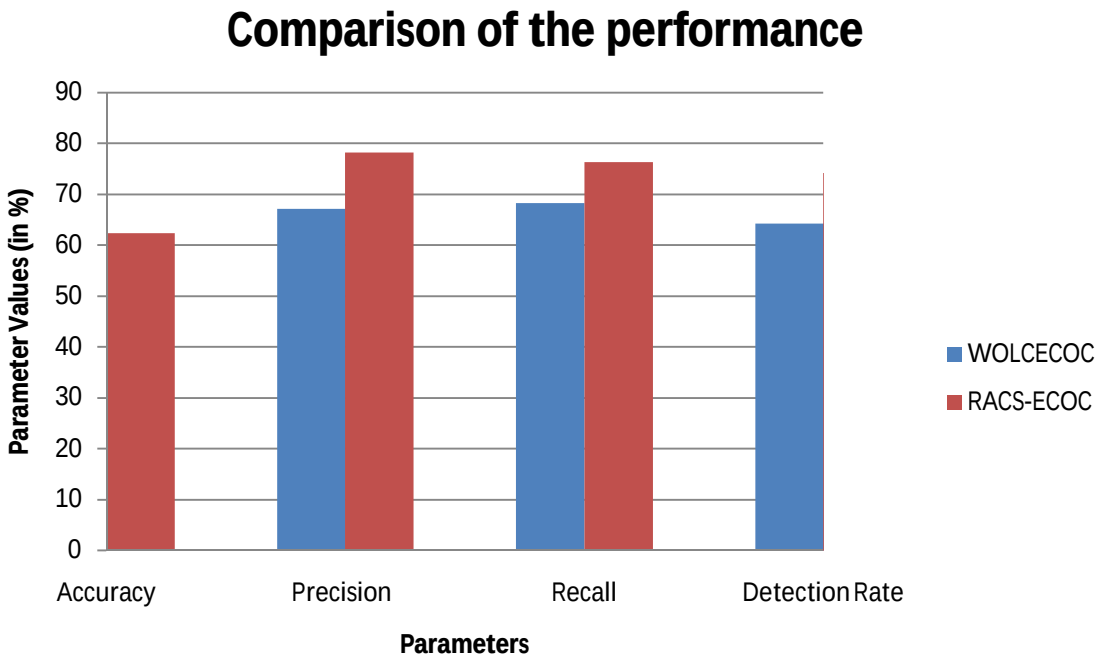

Figure 4.1: A Graphical comparison of Results for proposed method.

To analyse the performance of the existing and proposed technique accuracy, detection rate precision and recall are used as an evaluation parameter. Figure 4.1 shows a bar graph of the existing and proposed technique. In that X-axis shows the categories or parameter on which the performance of the graph is evaluated and Y- axis contains the value of the existing technique over the different category. % is used as a unit to analyse the performance of the technique. That comparison shows proposed technique provides advanced performance to classify the dataset

V. CONCLUSION AND FUTURE WORK

There are various algorithms like, DECOC [4] which takes a binary decision tree in coding phase, where each node of the tree is a strong classifier but it has an inherent demerit of decision tree, if any node of the tree is not classify correctly then there is possibility that successor nodes of that node also has wrong classifiers and there is no chance to correct it further. Subclass-ECOC [6], it divide classes in to subclasses and generate classifiers for those classes, it works on the principal that many small strong classifier is better than a single one but it also take binary decision tree for sub classes, it difficult to control the scale of subclasses. ECOC-ONE [8] it initialize with conventional ECOC classifiers ensemble and iteratively adds the binary class classifier that discriminates the most confusing pair of the classes to ECOC ensemble til the expected performance reached but this technique is not good to deal with hard to classify problems. WOLC-ECOC [1], in this technique author uses CPA based weight optimization which is time consuming and less accurate. This paper proposes a RACS-ECOC which uses Recursive Ant Colony System to optimize results. A comparative analysis is shown in results section which shows that propose method performs better in terms of accuracy, precision, recall and also have a high detection rate to classify data.

REFERENCES

1. Xiao Lie Zhang "Heuristic Ternary error correcting output codes via weighted optimization and layer clustering based approach"inIEEE Transactions on cybernetics, VOL. 45, NO. 2, February 2015
2. Rahman and B. Verma, "Novel layered clustering based approach for generating ensemble of classifiers," IEEE Trans. Neural Netw., vol. 22, no. 5, pp. 781–792, 2011.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

3. Thomas G. Dietterich, GhulamBakir"Solving Multi-class Learning Problem via Error Correcting Output Codes" Journal Of artifitialIntelligence Research 2(1995).
4. OriolPujol, PetiaRadeva, and JordiVitria, "Discriminant ecoc: A heuristic method for application dependent design of error correcting output codes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 6, pp. 1007– 1012, 2006.
5. O. Pujol, S. Escalera, and P. Radeva, "An incremental node embedding technique for error correcting output codes," Pattern Recogn., vol. 41, no. 2, pp. 713–725, 2008.
6. S. Escalera, D. M. J. Tax, O. Pujol, P. Radeva, and R. P. W. Duin, "Subclass problem-dependent design for error-correcting output codes," IEEE Trans. Pattern Anal.Mach.Intell.,vol.30,no.6,pp.1041–1054,2008.
7. X. L. Zhang, J. Wu, Z. P. Chen, and P. Lv, "Optimized weighted decoding for errorcorrecting output codes," in Proc.Int.Conf.Acoustic, Speech,SignalProcess.,2012, pp. 2101–2104.
8. T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," J. Artif. Intell. Res., vol. 2, pp. 263–286, 1995.
9. S. Escalera, O. Pujol, and P. Radeva, "On the decoding process in ternary error-correcting output codes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, pp. 120–134, 2010.
10. E.L.Allwein,R.E.Schapire,andY.Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," J. Mach. Learn. Res., vol. 1, pp. 113–141, 2001.
11. T. G. Dietterich, "Ensemble methods in machine learning," in Proc. Multiple Classifier Syst., 2000, pp. 1–15.
12. Prior and T. Windeatt, "Over-fitting in ensembles of neural network classifiers within ecoc frameworks," in Proc. Multiple Classifier Syst., 2005, pp. 834–844.