Multiclass classification using support vector machine and OligoIS Technique

Shubhangi Dusane#1, Er.Praveen Bhanodia*2, Er.Atish Mishra#3

#1 Shubhangi Dusane, Student M.Tech CSE, PCST Indore, RGPV Bhopal, India
#2 Er.Praveen Bhanodia, H.O.D( CSE), PCST Indore, RGPV Bhopal, India
#3Er. Atish Mishra, Assistant Prof.(CSE), PCST Indore, RGPV Bhopal, India

Abstract — In current era, data were collected from various sources like, public survey, web survey, data collected by analysis etc. So these collected data may be or may not be having balanced class. That time, this is typical, task to manage data. If, we want select some instances from these data, then this is a triodes job. Here proposed Hybrid approach for the large class problem in multiclass classification using SVM for solving large class problem. Hybrid OligoIS is an extension of OligoIS. Proposing hybrid approach of OligoIS and SVM can be used in large class problem and also it better perform when increases size of the database. On the Er.Praveen Bhanodia basis of Binary Decision Tree and Probabilistic output of Support Vector Machine here want to present Binary Decision Tree SVM (BDT) using Support Vector Machine(SVM) as an original approach to the multi-class classification problem. Instead of using a simple SVM classifier in each node, here propose SVM classifier associated with a threshold function of OligoIS to estimate the probability of membership to each sub-group in the node. Proposing 6 classifier architecture OligoIS-SVM takes the advantage of both the highest classification accuracy of SVM and the efficient computation of the tree architecture. OligoIS-SVM is based on recursively dividing the classes into two groups in every node of the Binary Decision Tree and training an SVM associate with threshold value (OligoIS Algorithm). Here we take three types of data set small, medium and large. There are three Dataset in small dataset category i.e. IRIS, WINE, ECOLI number of classes less than 10. and performed our algorithms on those dataset, also in medium dataset number of classes less than 50. We have taken dataset i.e. Movement Libra and finally in large number of dataset we have taken two dataset Thyroid, Yeast. Number of classes more than 100 and performed our functionality process.

Keywords— SVM , SVM -BDT

I. Introduction

In the current scenario data were collected from various sources. Than this data may or may not be class balanced data. A data set is called imbalanced if it contains many more samples from one class than from the rest of the classes. Data sets are unbalanced when at least one class is represented by only a small number of training examples (called the minority class) while other classes make up the majority. In this scenario, classifiers can have good accuracy on the majority class but very poor accuracy on the minority classes due to the influence that the larger majority class has on traditional training criteria. Most original classification algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of misclassification errors. In particular, they implicitly assume that all misclassification errors cost equally.

In many real-world applications, this assumption is not true. The differences between different misclassification errors can be quite large. For example, in medical diagnosis of a certain cancer, if the cancer is regarded as the positive class, and non-cancer (healthy) as negative, then missing a cancer (the patient is actually positive but is classified as negative; thus it is also called —false negative‖ ) is much more serious (thus expensive) than the false-positive error.

This is very interesting job to select useful instances from datasets. The instance selection is a method in which, we select useful instances from dataset. Unbalanced data set, a problem often found in real world application, can cause seriously negative effect on classification performance of machine learning algorithms. There have been many attempts at dealing with classification of unbalanced data sets. A two-class data set is said to be imbalanced (or skewed) when one of the minority classes is heavily under-represented in comparison to the other majority class. Traditionally, research
on this topic has mainly focused on a number of solutions both at the data and algorithmic levels. However, there have recently appeared other research lines within the general framework of class imbalance. These can be categorized into four groups:

- Resampling methods for balancing the data set.
- Modification of existing learning algorithms.
- Measuring the classifier performance in imbalanced domains.
- Relationship between class imbalance and other data complexity characteristics.

The most common techniques to deal with unbalanced data include resizing training datasets, cost-sensitive classifier, and snowball method. Recently, several methods have been proposed with good performance on unbalanced data. These approaches include modified SVMs, k nearest neighbour (kNN), neural networks, genetic programming, rough set based algorithms, probabilistic decision tree and learning methods. There are different types of sampling methods are available like, Oversampling and Under sampling. Oversampling is the simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a no heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, since this method replicates existing examples in the minority class, over fitting is more likely to occur. Under-sampling is an efficient method for class imbalance learning. This method uses a subset of the majority class to train the classifier. Since many majority class examples are ignored, the training set becomes more balanced and the training process becomes faster. The most common preprocessing technique is random majority under-sampling (RUS). In RUS, Instances of the majority class are randomly discarded from the dataset.

**II RELATED WORK**

Madzarov is presented a novel architecture of Support Vector Machine classifiers utilizing binary decision tree (SVM-BDT) for solving multiclass problem. Design the hierarchy of binary decision subtask using SVMs with clustering algorithm. The clustering model utilizes distance measures at the kernel space instead of input space. The experimental result indicates that the training phase of SVM-BDT is faster while comparing better accuracy with other SVM based approaches, ensembles of tree (Bagging and random forest) and neural network. During the testing phase, due to its logarithmic complexity, SVM-BDT is much faster than widely used multi-class SVM methods like OaO and OaA.

Platt present a method for extracting probabilities \( p(\text{class|input}) \) from SVM outputs, which is used for classification post-processing. Standard SVM does not provide such probabilities. After train an SVM, train the parameters of an additional sigmoid function that used to map the SVM output into probabilities. In the Experimental steps SVM+sigmoid yield probabilities are compared to the raw SVM.

Arun present an improved version of One-against-All (OAA) method for multiclass SVM classification based on decision tree. The proposed decision tree based OAA (DTOAA) aimed, to increasing the classification speed of OAA by use of posterior probability estimates of binary SVM output. DTOAA decreases the average number of binary SVM test that required in testing phase. It is compared to OAA and other multi class classification methods. Computational comparison indicates that the proposed method can achieve almost the same accuracy as OAA with 99.92% but much faster in decision making. When there are hundred or even thousands of classes, existing technique for mapping multiclass problem into simple binary classification problem give serious efficiency problem.

**IV PROBLEM DEFINITION**

A major challenge was to find good datasets that can be used for data mining. To gain a good understanding of the data and to create models with reasonable support we are in need of complete and noise-free datasets. Most available datasets are neither carefully selected nor up-to-date; hence, the task of predicting anything from this data will not yield good results.
During this thesis, we came across a lot of datasets that were either incomplete or simply not expressive enough to allow an accurate prediction. Hence, we argue, that further experiments on data mining from Semantic Web data could be greatly facilitated with the creation of common datasets for the evaluation and comparison of different approaches. The benefit of our approach is based on the expressiveness of the underlying ontologies. While ontologies with a deep inheritance hierarchy can outperform data mining without ontology support,

III PROPOSED STATEMENT

Previous works have shown that under sampling the majority class usually leads to better results than oversampling the minority class when oversampling is performed using sampling with replacement from the minority class. Furthermore, combining under sampling of the majority class with oversampling of the minority class has not yielded better results than under sampling of the majority class alone. One of the possible sources of the problematic performance of oversampling is the fact that no new information is introduced in the training set as oversampling must rely on adding new copies of minority class instances already in the data set. Sampling has proven a very efficient method of dealing with class-imbalanced data sets.

Removing instances only from the majority class, usually referred to as one-sided selection (OSS), has two major problems. First, reduction is limited by the number of instances of the minority class. Second, instances from the minority class are never removed, even when their contribution to the model’s performance is harmful. However, few attempts have been made to cope with class- imbalanced data sets using instance selection algorithms, which can remove instances from both the minority and majority classes. Standard widely used methods can be applied, but they do not achieve good results because their design bias is not focused on these kinds of problems. Evolutionary computation has been used with more success, but scalability is important and those methods cannot be applied to large and very large data sets.

There are various instance selection (IS) algorithms are available for reduction of data sets, but this traditional IS algorithm is not sufficient for reduction of imbalanced data sets. another problem with this type of data sets are space and time complexity. so here, we have proposed new methodology for reduction of this data sets. by applying this algorithm we reduced learning time for k-NN Classifier

IV. PROPOSED METHODOLOGY

In our proposed technique after analysing the SVM in detail we have proposed a new technique OligoIS SVM technique which we have used for the multiclass classification for the dataset which we have taken in different category for the working ahead purpose.

Oligarchic instance selection (OLIGOIS) is used for class imbalanced type of dataset. Subset used in it is always containing the same number of instances from both classes. The number of times that each instance has been selected to be kept in record after applying instance selection algorithm. The combination of classifier with its similarity in an ensemble by voting, call this record as number of votes. Repeat these processes for number of rounds with different random partition of data set. Threshold must be set at the final combination of the votes at the final output of the process where an instance must be selected.

We have used following Algorithm of Hybrid OligoIS

Data: A training set T = {(x1, y1), . . . , (xn, yn)}, subset Size s, and number of rounds r.
Result:

The set of selected instances \( S \subseteq T \).

for \( i = 1 \) to \( r \) do

1 Divide instances into \( n_s \) disjoint subsets \( D_i : U_i D_i = T \) of size \( s \) for \( j = 1 \) to \( n_s \) do

2 Apply instance selection algorithm to \( D_j \)

3 Store votes of selected instances from \( D_j \) end end

4 Obtain thresholds of votes to keep an instance from the minority, \( t^+ \), and the majority, \( t^- \), classes

5 \( S = \{ x_i \in T | (\text{votes}(x_i) \geq t^+ \text{ and } x_i \in C^+) \text{ or } (\text{votes}(x_i) \geq t^- \text{ and } x_i \in C^-) \} \)

6. Apply SVM classification technique.

7 end

V. IMPLEMENTATION AND RESULT ANALYSIS

On the basis of experimental setup we have taken various dataset of different number of classes and we have found the analysis on the basis of experiment performed on MATLAB tool available with us and got following explained result and we have found the algorithm proposed is providing or finding better results than the compare with the previous SVM or other multiclass classification technique available with us, the complete flow of the implementation is given below that what exactly we have performed with the implementation.

![Complete Work Flow Analysis](image)

We have taken three dataset in small dataset category and performed our algorithms on those dataset, also in medium number category we have taken two dataset and finally in large number of dataset we have taken two dataset and performed our functionality process. Here is the output screens and analysis result for our experiments.

We have calculated the results in the form of timing and accuracy and the results are compared presenting below:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training time</th>
<th>Testing time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligoi SVM</td>
<td>21.5325</td>
<td>1.7959</td>
<td>95.2489</td>
</tr>
<tr>
<td>BDT SVM</td>
<td>24.4090</td>
<td>0.0224</td>
<td>56.2083</td>
</tr>
<tr>
<td>DAG SVM</td>
<td>291.4891</td>
<td>0.0578</td>
<td>61.5000</td>
</tr>
<tr>
<td>OAO SVM</td>
<td>524.8090</td>
<td>0.1010</td>
<td>88.5833</td>
</tr>
</tbody>
</table>

Table: Comparison analysis of classification using different algorithm
the proposed technique contain a hybrid approach where we have used characteristics of the SVM and Oligois and performed the proposed algorithm which is efficient in the term of training time and the accuracy % which we have calculated according to our simulation result analysis which we have performed using MATLAB tool.

VI CONCLUSION:

The Classification and technique to retrieve data with large number of data is often taken long time for the training and testing and even after the long interval we can’t satisfy the requirement in accuracy phase and get less accuracy with different algorithm we have discussed in our project, thus we proposed a algorithm or approach where we are performing and taken the advantage of two different algorithm at their best and concluded the result with the existing techniques and the performance results are compatible in the timing and at the same time our approach is finding the best accuracy among the techniques we have discussed with the same sort of dataset and here we can observe in our technique the best results even with the large dataset and giving effective result without entity.

VII. REFERENCES

[01] Agarwal, S., Pandey, G. N., & Tiwari, M. D.; Data Mining in Education: Data Classification and Decision Tree Approach.(2012)


[3]. Dr.D.Ramya.chitra, nikandan,‖ Imbalanced DataSet Classification and solution:A Reviw‖ International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166 Volume 5 Issue 4 July 2014


