

Improving Feature Selection Algorithm And Support Multiclass Problem In Oligois

Author: Lavanya.S¹; Dr.Palanisami.S²; Veeralakshmi.R³

Department of CSE, Anna University Regional Centre, Coimbatore¹; Principal, Government College of Engineering, Bodinayakanur, Tamilnadu, ²; PG Scholar, Anna University Regional Centre, Coimbatore³

Abstract

In current research, an enormous amount of information is constantly being produced, which poses a challenge for data mining algorithms. Many of the problems in extremely active research areas, such as bioinformatics, security and intrusion detection, or text mining, share the following two features: large data sets and class-imbalanced distribution of samples. Although many methods have been proposed for dealing with class-imbalanced data sets, most of these methods are not scalable to the very large data sets common to those research fields. In existing work OligoIS is dealing with the class-imbalance problem that is scalable to data sets with many millions of instances and hundreds of features. But in that system using single class values, this method is not suitable for multi class value. The goal of multi-class supervised classification is to develop a rule that accurately predicts the class membership of new samples when the number of classes is larger than two. By consider high-dimensional class-imbalanced data: the number of variables greatly exceeds the number of samples and the number of samples in each class is not equal. By using on Friedman's one-versus-one approach for three-class problems can show how its class probabilities depend on the class probabilities from the binary classification sub-problems. It explores its performance using diagonal linear discriminant analysis (DLDA) as a base classifier and compares its performance with multi-class DLDA, using simulated and real data. The class-imbalance has a significant effect on the classification results: the classification is biased towards the majority class as in the two-class problems and the problem is magnified when the number of variables is large. The amount of the bias depends also, jointly, on the magnitude of the differences between the classes and on the sample size: the bias diminishes when the difference between the classes is larger or the sample size is increased. Also variable selection plays an important role in the class-imbalance problem and the most effective strategy depends on the type of differences that exist between classes. DLDA seems to be among the least sensible classifiers to class-imbalance and its use is recommended also for multi-class problems. Whenever possible the experiments should be planned using balanced data in order to avoid the class-imbalance problem.

Index Terms: OligoIS, Friedman's one-versus-one approach, multi-class DLDA

1. INTRODUCTION

Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest? The newest answer is data mining. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

The first and simplest analytical step in data mining is to describe the data — summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). The Data Mining process is collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. It must build a predictive model based on patterns determined from known results, and then test that model on results outside the original sample. A good model should never be confused with reality (you know a road map isn't a perfect representation of the actual road), but it can be a useful guide to understanding your business. The final step is to empirically verify the model.

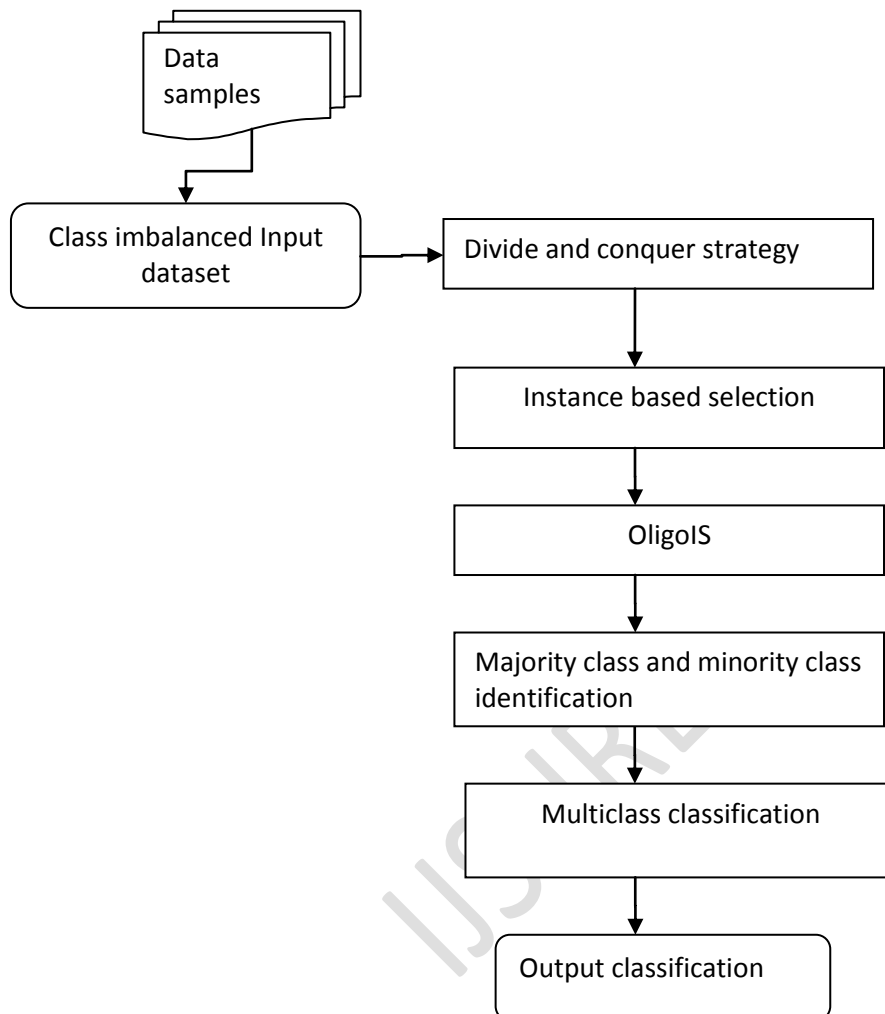
A main process in data mining is the one known as data reduction. In classification, it aims to reduce the size of the training set mainly to increase the efficiency of the training phase and even to reduce the classification error rate. Instance Selection is one of the most known data reduction techniques in data mining. Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database.

Frequent patterns are the patterns that occur frequently in data. Mining frequent patterns lead to the discovery of interesting associations and correlations within data. The degradation of classification performance in the presence of

imbalanced data has also been associated with aspects related to training data quality, such as the amount of class overlap and the lack of minority class representativeness.

Two main strategies developed to address the class imbalance problem are data preprocessing and

Fig-1: Proposed Architecture



Algorithmic. The objective of the first one is to modify (balance) the class prior distributions by sampling data in the input space, including oversampling of the minority class, under sampling of the majority class, or a combination of both. Most learning algorithms expect an approximately even distribution of instances among the different classes and suffer, to different degrees, when that is not the case. Dealing with the class-imbalance problem is a difficult but relevant task as many of the most interesting and challenging real-world problems have a very uneven class distribution. In that system not consider the multi class problem. Especially, many ensemble methods have been proposed to deal with such imbalance. However, most efforts so far are only focused on two-class imbalance problems. There are unsolved issues in multi-class imbalance problems, which exist in real-world applications. No existing methods can deal with multi-class imbalance problems efficiently and effectively.

The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new

understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation. Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant. In Subset selection methods, it assesses subsets of variables

according to their usefulness to a given predictor. Feature construction goal includes increasing the predictor performance and building more compact feature subsets.

We have selected a large collection of imbalanced data sets from KEEL-dataset repository and UCI Machine Learning Repository for developing our experimental analysis. In order to deal with the problem of imbalanced data sets we are using of a preprocessing technique, the "Synthetic Minority Over-sampling Technique" (SMOTE) to balance the distribution of training examples in both classes. Our Proposed architecture is shown in Fig-1

2. OLIGOIS

The oligarchic instance selection treats majority class instances unfairly; favoring minority class instances and provides scalability of data set with thousands of features

and millions of instances. In order to provide scalability it uses divide-and-conquer technique. Training set T with n instances will be partitioned into number of disjoint subsets D_j of equal size s as follows:

$$T = \bigcup_{j=1}^t D_j$$

Disjoint subsets contain instances from both majority and minority classes. Instances selection is applied for all subsets separately and the result of all subsets are recorded. The number of times each instances has been selected are also recorded. We can call this as votes of instances. This process is repeated for number of rounds and results of each rounds is recorded.

We focused mainly on DLDA because of its good behaviour in the two-class problems with high-dimensional class-imbalanced data; another reason for choosing DLDA was the straightforward generalization of the two-class DLDA to the multi-class situation (multi-class DLDA, mDLDA). Friedman's approach was chosen because of its wide applicability and simplicity, and because it is beneficial when the classes are imbalanced or when the number of classes is large.

2.1 Class Prediction

The number of samples with n , the number of variables with p and the number of variables selected and used in the classification rule with G , these variables are the most informative about class distinction; K is the number of classes while the class membership of the samples is indicated with integers from 1 to K ; the classes are non-overlapping and each sample belongs to exactly one class, the number of samples in Class k is denoted by n_k .

Let x_{ij} be the expression of j^{th} variable ($j = 1... p$) on i^{th} sample ($i = 1... n$). For sample i we denote the set of G selected variables by x_i . Let $\bar{x}_g^{(k)}$ denote the mean expression of the g^{th} selected variable in Class k . The mean expression of the g^{th} variable in Class k is defined as

$$\bar{x}_g^{(k)} = \frac{1}{n_k} \sum_{i \in c_k} x_{ig} \quad (2.1)$$

and let x^* represent the set of selected variables for a new sample.

2.2 Multi-class DLDA

Discriminant analysis methods are used to find linear combination of variables that maximize the between-class variance and at the same time minimize the within-class variance. Diagonal linear discriminant analysis (DLDA) is a special case of discriminant analysis that assumes that the variables are independent and have the same variance in all classes. The multi-class DLDA (mDLDA) classification rule for a new sample x^* is linear and is defined as

$$C(x) = \operatorname{argmin}_k \sum_{g=1}^G \frac{(x_g^* - \bar{x}_g^{(k)})^2}{S_g^2} \quad (2.2)$$

Where S_g^2 is the sample estimate of the pooled variance for variable g and x_g^* is the g^{th} selected variable of the new sample.

2.3 Friedman's approach

The class-prediction problem for $K > 2$ classes is divided in $\binom{K}{2}$ binary class-prediction problem, one for all pairs of classes. Within each binary class-prediction problem we build a rule for class-prediction (train a classifier) and a new

sample is classified in one of the two classes. The final class-prediction in one of the K classes is defined with majority voting, assigning the new sample to the class with most votes. Friedman's approach considers only the class-membership when deciding for the final classification of the new samples. The probabilities of belonging to a certain class rather than the class membership derived from the binary sub-problems could also be considered. The class probabilities using Friedman's approach is a function of the class probabilities from the binary class-prediction sub-problems. Consider a classification problem with $K = 3$ classes. The possible outcomes of the three possible pair wise comparisons among the three classes are listed in Table-1, where the class assignments derived using Friedman's approach are also given together with the notation used to denote the probabilities of each outcome (the subscripts indicate the winning class from each binary comparison). The class is chosen at random when each of the three classes receives a vote.

Table-1: Class assignments and probabilities using Friedman's approach.

Vote			Probability	Class assignment
1 vs 2	1 vs 3	2 vs 3		
1	1	2	p_{112}	Class 1
1	1	3	p_{113}	Class 1
1	3	2	p_{132}	Class 1,2 or 3
1	3	3	p_{133}	Class 3
2	1	2	p_{212}	Class 2
2	1	3	p_{213}	Class 1,2 or 3
2	3	2	p_{232}	Class 2
2	3	3	p_{233}	Class 3

The probabilities of assigning a new sample to each of the classes are (see Table-1 for the definition of p_{ijk})

$$P(C = 1) = p_{112} + p_{113} + \frac{1}{3}(p_{213} + p_{132})$$

$$P(C = 2) = p_{212} + p_{223} + \frac{1}{3}(p_{213} + p_{132})$$

$$P(C = 3) = p_{133} + p_{233} + \frac{1}{3}(p_{213} + p_{132})$$

Let us assume that in each pair wise comparison the new samples are equally likely to be assigned to both classes. In this case all the outcomes listed in the table are equally likely and it is therefore straightforward to show that the Friedman's approach would assign the new samples to each class with equal probability: $P(C = 1) = P(C = 2) = P(C = 3) = 1/3$. This behavior would be expected when there are no true differences between classes in the training set (null case).

2.4 Simple Undersampling

Simple undersampling (down-sizing) consists of obtaining a class-balanced training set by removing a subset of randomly selected samples from the larger class. In mDLDA undersampling consisted in using $\min(n_1, n_2, n_3)$ samples from each class, randomly selecting which samples from the majority class(es) should be removed. With Friedman's approach each pair wise comparison was undersampled if the size of the classes was not equal ($n_k \neq n_j$).

2.5 Variable Selection

The variable selection method used to reduce the number of variables can be different from the method used for mDLDA. The $G < p$ variables that were most informative about class distinction were selected on the training set and used to define the classification rules (Eq. 2.2). Variable selection was based on two sample t-test with assumed equal variances for the Friedman's approach, or F-test for the equality of more than two means for mDLDA. A limited set of simulations and real data analysis were carried out using the F-test with Friedman's approach. In the null case we also considered the situation where all the variables were used ($G = p$). We carried out also a limited set of simulations where the variables used with the Friedman's approach were selected with the F-test using all the classes. To reduce the computational burden in the reanalysis of the breast cancer gene-expression data sets we considered only the $p = 1000$ variables with the largest variances. The variable selection consisted in selecting on each training set the $G = 40$ variables with the smallest p-values.

2.6 Evaluation of the performance of the classifiers

The performance of the classifiers was evaluated on the independent test sets. It is well known that for imbalanced data the proportion of correctly classified samples can be a misleading measure of the performance of a classifier. For this reason four different measures of performance were considered: (i) overall predictive accuracy (PA, the number of correctly classified subjects from the test set divided by the total number of subjects in the test set), (ii) predictive accuracy of Class 1 (PA1, i.e., PA evaluated using only samples from Class 1), (iii) predictive accuracy of Class 2 (PA2 i.e., PA evaluated using only samples from Class 2) and (iii) predictive accuracy of Class 3 (PA3). Their standard deviations were also reported.

3. CONCLUSION

We have presented a new method of instance selection in class-imbalanced data sets that is applicable to very large data sets. The method consists of concurrently applying instance selection on small class-balanced subsets of the original data set and combining them by means of a voting method, setting a different threshold for minority and majority class samples. Our proposed system results show that the class-imbalance has a significant effect on the classification results also in multi-class problems and that its influence is magnified when the number of variables is large. The amount of bias depends also jointly on the magnitude of the differences between classes and on the sample size, i.e. the bias diminishes when the differences between the classes are larger or the sample size is increased. Also variable selection plays an important role in the class-imbalance problem and the most effective strategy depends on the type of differences that exist between classes. DLDA

seems to be among the least sensible classifiers to class-imbalance and its use is recommended also for multi-class problems. Whenever possible the experiments should be planned using balanced data in order to avoid the further complications arising from the class-imbalance.

REFERENCES

- [1]. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, Jan. 2002, "SMOTE: Synthetic minority over-sampling technique", *J. Artif. Intell. Res.*,
- [2]. C. J. Carmona, J. Derrac, S. García, F. Herrera and I. Triguero, Feb. 2012, "Evolutionary-based selection of generalized instances for imbalanced classification", *Knowl.-Based Syst.*,
- [3]. J. R. Cano, F. Herrera, and M. Lozano, Dec. 2003, "Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study", *IEEE Trans. Evol. Comput.*,
- [4]. A. Elisseeff and I. Guyon, Mar. 2003, "An introduction to variable and feature selection", *J. Mach. Learn. Res.*,
- [5]. A. Estabrooks, N. Japkowicz and T. Jo, Feb. 2004, "A multiple resampling method for learning from imbalanced data sets", *Comput. Intell.*,
- [6]. N. García-Pedrajas, Feb. 2009, "Constructing ensembles of classifiers by means of weighted instance selection", *IEEE Trans. Neural Netw.*,
- [7]. E. A. Garcia and H. He, Sep. 2009, "Learning from imbalanced data", *IEEE Trans. Knowl. Data Eng.*,
- [8]. N. García-Pedrajas, D. Ortiz-Boyer and J. A. Romero del Castillo, Mar. 2010, "A cooperative coevolutionary algorithm for instance selection for instance based learning", *Mach. Learn.*,
- [9]. L. Kuncheva and C. J. Whitaker, May 2003, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy", *Mach. Learn.*,
- [10]. T. R. Martinez and D. R. Wilson, Mar. 2000, "Reduction techniques for instance based learning algorithms", *Mach. Learn.*,
- [11]. F. Provost and G. M. Weiss, 2001, "The effect of class distribution on classifier learning: An empirical study", *Dept. Comput. Sci., Rutgers Univ., Newark*,
- [12]. D. L. Wilson, Jul. 1972, "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Trans. Syst.*,