International Journal of Research in Advent Technology, Vol.7, No.5, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

On Applying Sampling Methods and Rule Based Classifiers

Shalini Bhaskar Bajaj
Department of Computer Science and Engineering
Amity University Haryana, India
shalinivimal@gmail.com

Abstract: The objective of the study presented here is to assist oncologists in predicting metastasis of cervical cancer by applying sampling methods and evaluating their effect on the rule based classifiers. Cervical cancer dataset used here is unbalanced in nature. The sampling methods used are over-sampling, under-sampling and combine over-sampling. Rule based classifiers such as Conjunctive Rule, Decision Tree with Naive Bayes, Decision tree, PART and Repeated Incremental Pruning to Produce Error Reduction are applied on the dataset using WEKA data mining tool. From the results obtained experimentally, it has been found that the performance of the DTNB and over-fitting sampling approach in terms of AUC, G-mean and F-measure is the best when compared to different sampling strategies across five rule based classification algorithms named above on three different metrics. By applying the above results on the dataset of new patients we can classify them into metastasis or not metastasis category and thus can assist an Oncologist in short time by labelling the patient in one of the classes: metastasis and non-metastasis.

Keywords: Bone metastasis, rule based classifier, cervical cancer, sampling

1. INTRODUCTION

Knowledge discovery in datasets is to find different types of patterns and to discover relationship among the datasets by applying different data mining techniques. One of the important techniques that can be applied is the rule based classification methods to uncover the hidden patterns and to provide rules in the form of IF-THEN that are easy to understand and use. Applying the decision tree and classification rules on medical datasets gives efficient results and thus enhances the knowledge of progression of the disease and treatment.

Cervical cancer is one of the most commonly occurring cancer in women followed by breast cancer and liver cancer that causes early death in year 2002 [1]. In case of bone metastasis from cervical cancer, the median age of patient os found to be 46 years with a range of 15 to 76 years [2]. For treatable patients, radiation therapy provides moderately good palliation [3]. The dataset for cervical cancer is noisy and unbalanced (more data belonging to one class). Data preprocessing therefore becomes important to handle such noisy dataset. Preprocessing of data is independent of the classifier chosen. Over-sampling increases the minority class by randomly duplicating the positive samples. SMOTE (Synthetic Minority Over-sampling TEchnique) technique [4] adds synthetically generated samples of minority class rather than replacing the existing dataset values. In case of under sampling approach data of majority class too close to the decision boundary are removed with thep of Tomek link or TLink [5], Neighbourhood cleaning rule [7] and One sided selection [6]. The third approach is the hybrid approach wherein both oversampling and under sampling are used to select the final dataset using SMOTE and TLink [8]. Combined approach provides better performance as compared to oversampling and under sampling.

This paper presents predictive analytic model that uses rule based classification (PART, Repeated Incremental Pruning to Produce Error Reduction, Decision Tree with Naive Bayes, Decision Tree, Conjunctive Rule) with sampling technique (ROS, SMOTE, RUS, TLink, SMOTE+TLink) for the imbalance problem in bone metastasis of cervical cancer. Section 2 discusses rule based classifiers; section 3 focusses on different sampling approaches in preprocessing; section 4 give details on the proposed algorithm; section 5 highlights experimental results and discussion and finally section 6 discusses conclusion and future work.

2. RULE BASED CLASSIFICATION ALGORITMS

In Rule based classifiers rules are formulated using "IF-THEN" rules. In order to achieve good performance, a number of rule based classifiers have been proposed. Some of the rule based classifiers are discussed below:

A. Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER is a fast "IF-THEN" rules learning algorithm proposed by Cohen W. [9]. It has integrated reduced error pruning and a separate and conquer rule learning algorithm called incremental reduced pruning in which it greedily prepares a rule set by adding rules until all positive examples are covered. The algorithm initialises the ruleset as empty and for each class from the less prevalent to the most frequent one. It has two stages (1) Build stage, (2) Optimization stage. Build stage is further divided into: (i).Grow phase and (ii).Prune phase. It repeats the Grow phase and Prune phase until: (a) the description length of the rule set is sixty-four bits greater than the smallest description length met so far, (b) there are no +ve examples, (c) the error rate is >=50%. In the Grow Phase, one rule is added at a time by greedily adding antecedent until the rule becomes 100% accurate. Every possible value for each attribute is tried and the value with the highest information gain is finally selected. Information gain is calculated using the following formulae: p(log(p/t)-log(P/T)). In the Prune Phase, it incrementally prunes each rule and final sequences of the antecedents. The pruning metrics used is: (p-n)/(p+n). In the Optimization stage, after the initial rule set is generated, two variants of the rule are generated and pruned from each rule. From this, the variant is selected based on the minimal description length as the final representative of the rule examined. More rules are generated based on the residual positives from the existing rule set. Removal of the rules from the description length is done if the rule is already in the resultant rule set.

B. Decision Tree with Naive Bayes (DTNB)

Decision Tree with Naive Bayes [12] was proposed by M. Hall and E. Frank for building classes using decision tree. At each node in the tree traversal, the algorithm tries to evaluate whether the class can be divided into sub classes or not. After seeing the merit of the division of the attributes into disjoint sets, the attributes are divided into subsets: one for decision table and the other for the Naive Bayes. At each step of division of the attributes, selected attributes are first modelled using Naive Bayes and rest of the attributes are modelled using decision table or some of the attributes are dropped entirely from the model.

C. Decision Tree (DT)

Decision tree algorithm summarises the dataset in the form of a decision table that contains the same number of attributes as the dataset. Using simple decision table majority classifier, the algorithm builds decision rules.

D. Conjunctive Rule

In Conjunctive Rule a single rule learner is used to predict nominal as well as numeric class labels. In this algorithm, the antecedents and consequents are ANDed within themselves for the purpose of classification or regression. The antecedents are selected by the learner on the basis of the information gain value and the consequents can take the class value from the available classes. Finally, the generated rule is pruned on the basis of the simple pre-pruning or reduced error pruning based on antecedent number. For classification, weighted average of accuracy rate is used and for regression weighted average of mean squared error is used on the pruned data.

E. PART

E. Frank and H. Ian proposed PART algorithm [10] (uses features from C4.5 and RIPPER) which uses separate and conquer strategy to generate decision lists for producing rules. In each iteration, this algorithm builds a partial C4.5 decision tree and extends the best leaf obtained in the rule.

3. SAMPLING METHODS

Three methods have been discussed in this section: over-sampling, under-sampling and combined sampling.

A. Oversampling methods

Two methods are discussed in over sampling: Random oversampling and SMOTE. Random over sampling is a heuristic method to increase the number of minority classes by replicating the samples randomly. Since the replication of samples from minority classes is done therefore the approach comes under overfitting of minority classes. On the other hand, SMOTE [4] is used to prepare the classifier from unbalanced dataset. In this case synthetic samples are created rather than by performing oversampling with minority classes.

B. Undersampling methods

I. Tomek proposed Tomek links or TLink [5] that is based on removal of borderline or noisy samples in order to enhance the nearest neighbor rule. In case of random under sampling which is a non heuristic method decreases the number of majority class samples randomly. The drawback of this approach is that it can potentially remove useful data from the given dataset due to random selection.

C. Combined method

Tomek link and SMOTE can be combined ad used on the datasets. It is called combined method as it uses both oversampling and under sampling technique. In this algorithm, random oversampling is done on the dataset by applying SMOTE on the minority class. After application of SMOTE, TLink is used to detect noisy samples in the majority class.

4. PROPOSED METHODOLOGY

This section covers details regarding the cervical cancer dataset and the proposed methodology followed.

A. About Cervical Cancer Dataset

A dataset of 3864 patients was recorded. In the presented dataset 148 samples belong to metastasis class and 3716 samples belong to non metastasis class. Attributes recorded for the dataset used in the study are as follows:

i) Class: YES for metastasis and NO for not metastasis
ii) Mensuration: after, between, before
iii) Abortion: 0-8
iv) Age: number
v) Gravidity: number
vi) Pathological groups: squamous cell carcinoma, adenocarcinoma, other
vii) Tumor Size: number
viii) Stage: 1A, 2A, 3A, 4A, 1B, 2B, 3B, 4B
ix) prity: number
x) Keratinizing: Keratinizing, Non-keratinizing, small cell, other
xi) LrPelvic Wall: YES/NO
xii) LtParametrium: YES/NO
xiii) RtPelvic Wall: YES/NO
xiv) RtParametrium: YES/NO
xv) Heamoglobin: number
xvi) Anemia: YES/NO
xvii) Interval: number
xviii) Aim of Treatment: Radical, Irradiated from other Institute, Palliation,Palliation up to radical
xix) Type of Irradiation: Radiation alone, Pre operative radiation, Post operative radiation, post irradiation from other institute

In order to increase the performance of the classifiers it is important that the data be preprocessed. Table 1 shows that the number of minority classes in ROS and SMOTE have equal size. In case of oversampling, the number of majority classes in RUS and TLInk are different with having more number of majority classes as compared to RUS. In case of hybrid approach, there is balance in the class distribution. Over sampling rate can be defined as follows:

Rate = {(NOriginal - NSampling)/ NOriginal} * 100

Table 1. Original and sample dataset description

DATA	**ORIGINAL**	**ROS**	**SMOTE**	**RUS**	**TLINK**	**SMOTE and TLINK**
YES	148	3010	3010	148	148	3210
NO	3716	3716	3716	865	3540	3445
	3864	6726	6726	1013	3688	6655

Calculation of Oversampling rate is given in Table 2.

Table 2. Rate of Oversampling (given in %)

Dataset	RUS	SMOTE	SMOTE+TLink
Cervical Cancer	74.6	74.6	68.5

B. Proposed Methodology

Calculation of Oversampling rate is given in Table 2. Steps of the proposed Methodology for the bone metastasis of cervical cancer is given below:

I. Collection of Patient's data from the radiation unit
II. Pre-processing collected data using Sampling techniques to make it suitable for the data mining process
III. Applying Rule based Classifier (Conjunctive Rule, DT, DTNB, PART, RIPPER)
IV. Generating Rule sets
V. Predicting Metastasis or Non-metastasis
VI. Applying rules on the new patient data to predict the class (metastasis/non-metastasis)

C. Preparation of Evaluation matrix:

For measuring the performance of the classifier and to guide the learning algorithm performance measurement metrics are very important. For this it is important to record the following:
PT - True Positive cases
NT - True Negative cases
PF - False Positive cases
NF - False Negative cases

The confusion matrix for the 2-class problem is given in Table 3. Data imbalance leads to reduced accuracy of classification.

Table 3. Confusion matrix for a 2-class problem

	Predictive	
	Positive	Negative

Actual	Positive	True Positive (PT)	False Negative (NF)
	Negative	False Positive (PF)	True Negative (NT)

Sensitivity is given by :
S = PT/(PT+NF)

Specificity is given by:
Sp = NT/(NT+PF)

G-Mean is given by:
G-Mean = (S X Sp)1/2

Precision is given by:
P = PT/(PT+PF)

Recall is given by:
R = PT/(PT+NF)

F-measure is given by:
F-Measure = 2X (P X R) / (P + R)

5. EXPERIMENTAL RESULTS AND DISCUSSION

Results of the computation of the five classifiers on the dataset of Bone metastasis of cervical cancer are compiled in the Table 4. From the results given in Table 4, it can be concluded that the over-sampling and hybrid methods give better results as compared to under-sampling.i.e. best results are shown by ROS and SMOTE approach. In majority classes, most samples are misclassified as minority classes thus RUS and Tomek links give poor results. It can be concluded that in order to deal with data imbalance, pre-processing is the most important step. The DNTB algorithm is the most accurate classifier in the original dataset as well as with the sampled dataset when compared with other classifiers. Size of the training set is increased due to oversampling. Due to this, more time is required in classification and at the same time sample duplication leads to over fitting.

6. CONCLUSION ANF FUTURE WORK

A predictive modelling approach has been discussed in this paper based on five rule based classifiers. The experimental evaluations are discussed based on the performance of the different algorithms on the bone metastasis of cervical cancer dataset by applying different sampling approaches. it has been shown that the DTNB outperforms all other classification algorithms in terms of AUC, G-mean, F-measure and accuracy.

Table 4. Results of different sampling methods using five rule based classifiers on cervical dataset

Method	**Metrics used**	**Rule based Classifier**				
		Decision Tree	**Decision Tree with Naive Bayes**	**RIPPER**	**Conjunctive Rule**	**PART**
Original	G-Mean	0	0	0	0	0
	AUC	51	51	51	51	51
	F-Measure	0	0.14	0	0	0
	Average	17	17.05	17	17	17
SMOTE +TLink	G-Mean	92.4	97.4	89.9	49.3	88.8
	AUC	91.6	97.3	90.7	62.7	89.7
	F-Measure	91.2	97.2	90.5	68.1	89.2
	Average	91.73	97.3	90.4	60.03	89.23
SMOTE	G-Mean	92.6	97.9	91.4	53.8	90.2
	AUC	92.7	97.8	92.6	58.4	90.6
	F-Measure	92.5	97.7	91.2	62.1	89.4

	Average	92.6	97.8	91.73	58.1	90.07
RUS	G-Mean	8.8	24.82	0	0	21.2
	AUC	50.4	49.9	50.2	50.2	51.3
	F-Measure	1.6	9.6	0	0	6.7
	Average	20.27	84.32	16.73	16.73	26.4
ROS	G-Mean	96.6	97.6	96.3	48.7	95.9
	AUC	96.1	97.5	96.1	61.2	95.6
	F-Measure	96.4	97.4	96.2	67.8	95.3
	Average	96.37	97.5	96.1	59.23	95.6
Tomek Link	G-Mean	0	9.4	0	0	0
	AUC	50.2	51.3	50.2	50.2	50.2
	F-Measure	0	1.6	0	0	0
	Average	16.73	20.77	16.73	16.73	16.73

REFERENCES

[1] A. Nartthanarung and D. Thanapprapasr, "Comparison of outcomes for patients with cervical cancer who developed bone metastasis after the primary treatment with concurrent chemoradiation versus radiation therapy alone", Intl. Journal of Gynaecological Cancer, Vol. 20, no. 8, pp. 1386-1390, 2010

[2] D. Thanapprapasr, A. Nartthanarung, p. Likittanasombut, N. Israngura Na Ayudhya, C. Charakorn, U. Udomsubpayakul, T. Subhadarbandhu and S. Wilailak, "Bone metastasis in cervical cancer patients overa 1- year period", Intl. Journal of Gynaecological Cancer, Vol. 20, no. 3, pp. 373-378, 2010

[3] V. Ratanatharathon, W. E. Powers, N. Steverson, I. Hun, K. Ahmad and J. Grimm, "Bone metastasis from cervical cancer", Cancer, 1994, vol. 73, no. 9, pp. 2372-2379

[4] N. V. Chawala, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over sampling technique", Journal Artificial Intl. Research, 2002, vol. 16, no. 1, pp. 321-357

[5] T. Ivan, "An experiment with the edited nearest-neighbour rule", IEEE Trans. on System, Man and Cybernetics, 1976, vol. 6, no. 6, pp. 448-452

[6] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one sided selection", Proc. 14th Intl. Con. on Machine Learning, 1997, pp. 179-186

[7] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution", AIME LNAI 2001, Springer, pp. 63-66

[8] G. Batista, R. Prati and M. Monard, "A study of behaviour of several methods for balancing machine learning training data", SIGKDD Explorations, 2004, vol. 6, no. 1, pp. 20-29

[9] W. Cohen, "Fast effective rule induction", Proc. od the 12th Intl. Conf. on Machine Learning, California, 1995, pp. 115-123

[10] E. Frank, H. Ian, "Witten: Generating accurate rule sets without global optimisation", 15th Intl. Conf. on Machine Learning, 1998, pp. 144-151

[11] R. Kohavi, "The power of decision tables", 8th European Conf. on Machine Learning, 1995, pp. 174-189

[12] M. Hall and E. Frank, "Combining naive bayes and decision tables", Proc. of the 21st Florida Artificial Intelligence Society Conf., FLAIRS, 2008, pp. 318-319

[13] G.Y.Wong, F.H.F.Leung, S.H. Ling and I.H. Frank, "Data mining: Practical machine learning tools and techniques", 2nd edition, Morgan Kaufmann, San Francisco, 2005