BR+ for Addressing Imbalanced Multilabel Data Classification Combined with Resampling Technique

Nilam Novita Sari[1], Ismaini Zain[2], Kartika Fithriasari[2], Amri Muhaimin[4]
{nilamnovitasari2013@gmail.com[1], ismaini_z@statistika.its.ac.id[2*], kartika_f@statistika.its.ac.id[3]}

Department of Statistics, Faculty of Science and Data Analystics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1,2,3,4]

Abstract. BR+ is a multilabel method that transforms multilabel into binary single label and assumes label dependency. BR+ can use any different classification method such as random forest. Random forest is an advantageous classification method. But presence of imbalanced classes, random forest will result in poor performance. Hence, handling imbalanced data can be done by applying resampling techniques consisting of SMOTE-NC and T-Link. The dataset used was adolescent risk behavior of drug abuse and premarital sex based on SKAP. The dataset has two labels means there are multilabel problems and the dataset is imbalanced. Thus, the combination of BR+ (Stat) and resampling techniques will be compared in handling multilabel imbalanced data in the classification of adolescent risk behavior using random forest. The results show that the optimum Mtry is 7 and the combination of BR+ (Stat) and T-Link is the best method to handle the multilabel imbalanced data.

Keywords: Multilabel Imbalanced Data, BR+, SMOTE-NC, Tomek Link, Random Forest

1 Introduction

Classification is the process of finding a model that describes and distinguishes data classes or concepts used to predict the class label of objects for which the class label is unknown [1]. In general, classification is usually used for a single-label problem where each instance is associated with a class label from a set of disjoint labels $|L|$. However, sometimes each instance is associated with more than one class, which is called multilabel. Problem Transformation (PT) method is one way to solve the multilabel problem because of the flexibility of the Problem Transformation. PT is a method that transforms the multilabel classification into one or more single label [3]. Several PT methods are often used such as binary relevance (BR). BR is extended to BR+ to overcome the limitation of BR. BR+ is a method that transforms multilabel into $|L|$ number of the binary single label and assumes label dependency. There are three ways to do BR+, which are BR+ NU (No Update), BR+ Stat (Static Order), and BR+ Dyn (Dynamic Order). The difference between the three methods is in the prediction phase. The Transformation Problem methods such as BR, CC, and BR+ were compared by [4] using SVM, J48, and Naïve Bayes (NB) as the base classifier. BR+ (Stat) and BR+ (Dyn) transformation have the best performance using J48 and NB, while for SVM, the best performance is obtained using the CC transformation method. BR+ can use any classification method such as random forest.

AECon 2020, December 19-20, Purwokerto, Indonesia

DOI 10.4108/eai.19-12-2020.2309179

Random forest is one of the decision tree methods. Random forest is a group of un-pruned decision trees made from the random selection in samples of the training data and then the prediction is made by aggregating (majority vote for classification) the predictions of the trees [5]. Random forests have some advantages, such as the ability to handle thousands of variables without deletion or deterioration of accuracy, its speed and ease of implementation , accuracy of prediction results produced, etc. [6,7]. Random forest is considered to be one of the most accurate techniques available [8]. Although random forest has some advantages, just like other classification methods, random forest faced problems when the dataset is imbalanced. For an imbalanced dataset, most of the classification algorithms tend to produce a high accuracy rate for the majority class and produce a low prediction rate for the minority class with a low accuracy rate.

An imbalanced dataset is a condition where fewer training instances exist in one class (minority class) than another class (majority class) [8]. An imbalanced dataset will result in poor performance because it produces low accuracy in the minority class. Resampling approaches can be used to solve this problem. Resampling approaches are techniques that rebalance the distribution of data. Resampling approaches are divided into three categories which are over-sampling, under-sampling, and hybrid sampling.

Random over-sampling approach duplicates the minority samples so that the instances in the minority class equal the instances in the majority class. One of the most popular random over-sampling methods is SMOTE (Synthetic Minority Over-sampling Technique). SMOTE is a method that creates “synthetic” instances in the minority class [9]. SMOTE can be extended to SMOTE-NC to handle mixed datasets of continuous and nominal features. Random under-sampling is another way to deal with imbalance problems by removing some instances in the majority class to balance the distribution of datasets [10]. The advantage of the under-sampling method is that it can reduce the size of the data by eliminating some instances and decreasing the run-time cost especially in the case of big data [11]. One of the under-sampling methods is Tomek Link (T-Link). T-Link can be used as an under-sampling method which removes instances in the majority class or as a cleaning method to remove noise. Another resampling technique is hybrid sampling. Hybrid sampling is the combination of over-sampling and under-sampling approaches used to make the dataset more balanced and it can improve the accuracy of classification performance. SMOTE+T-Link is one of the hybrid sampling methods used to clean data. SMOTE, Tomek Link, and SMOTE+TL were compared by [12] using SVM as the base classifier. The results of this study show that hybrid sampling SMOTE+TL has better performance than using only SMOTE or Tomek Link, but in the case of extreme data imbalance (minority class less than 10%), the SMOTE+TL is no better than using Tomek Link. Then, in 2016 [11] used different imbalanced method and different classification methods to compare their effectiveness in addressing the imbalance data issue and the results show that the combination of SMOTE+TL and RUS+TL have the best performance as compared to other sampling methods.

In this paper, we will compare the classification of the multilabel imbalanced data using random forest with different parameter tuning and we propose BR+ Stat to solve the multilabel problem and combine it with resampling techniques which are SMOTE-NC, T-Link, and a combination of SMOTE-NC and T-Link (SMOTE-NC+TL) to handle imbalanced data to find the best method to classify the multilabel imbalanced data.

2 Research Method

2.1 Data

The multilabel imbalance data used in this research was the adolescent risk behavior consisting of drug consumption as the label 1 and pre-marital sex as the label 2 based on SKAP (Survei Kinerja dan Akuntabilitas Program KKBPK) of East Java in 2019 by BKKBN. The adolescent risk behavior has 5300 instances.

2.2 Research Variables

The variables used in this research are as follows:

Y_1 = drug consumption (0 = No, 1 = Yes)

Y_2 = having pre-marital sex (0 = No, 1 =Yes)

X_1 = age (0 = < 19 years old, 1 = $\geq$ 19 years old)

X_2 = sex (0 = Male, 1 = Female)

X_3 = education (0 = did not go to school or has completed either elementary or junior high school, 1 = completed either senior high school or college education)

X_4 = domicile (0 = urban, 1 = rural)

X_5 = knowledge of drugs (0 = no, 1 = yes)

X_6 = knowledge of the physical consequences of the drug (0 = no, one = yes)

X_7 = knowledge of the psychological consequences of the drug (0 = no, one = yes)

X_8 = knowledge of the socioeconomic consequences of the drug (0 = no, one = yes)

X_9 = knowledge of adolescent sexual and reproductive health (ASRH) (0 = no, 1 = yes)

X_{10} = knowledge of women's fertility (0 = no, 1 = yes)

X_{11} = knowledge of pregnancy (0 = no, 1 = yes)

X_{12} = knowledge of women's marriageable age (0 = $\geq$ 21 years old, 1 = others)

X_{13} = knowledge of men's marriageable age (0 = $\geq$ 25 years old, 1 = others)

X_{14} = knowledge of the consequences of early marriage (0 = no, 1 = yes)

2.3 Research Design

The classification in this study is using random forest and the transformation problem method is using BR+ (Stat). In this research, we try to classify the data using BR+ (Stat) with the order $y_1 \prec y_2$ and $y_2 \prec y_1$ with the combination of SMOTE-NC, T-Link and SMOTE-NC+T-Link. The random forest modelling uses parameter tuning. The numbers of "M try" that are used here are 2, 4, and 7, and the numbers of the tree used are 50, 100, 250, and 500 trees.

The multilabel imbalanced data are partitioned into training and testing data using 5-fold cross validation. The performance measures used here are accuracy, sensitivity, specificity, precision, macro f-measure and hamming loss. The multilabel imbalanced classification process consisted of several stages as seen in Figure 1.

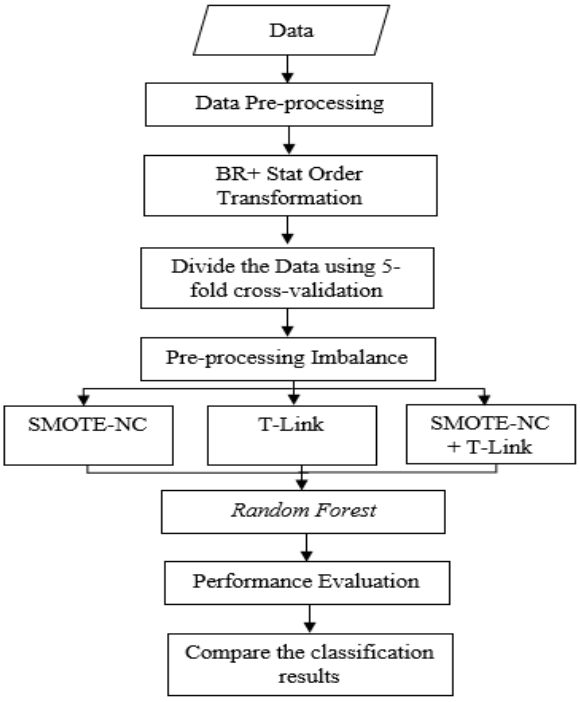

Fig. 1. Flowchart Research

3 Results and Discussion

The dataset has extremely imbalanced data the data in minority class are less than 5% for both labels, where in label 1 the minority class is 3.62% and for label 2 the minority class is 0.28%. The results are shown in the following Table.

Table 1. Performance of BR+ (Stat) using SMOTE-NC

	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**F-Measure**	**N Tree**	**M Try**
$y_1 \prec y_2$	0.8692	**0.9981**	0.0944	0.8667	0.9254	50	2
	0.8727	0.9975	0.0977	0.8708	0.9275	100	2
	0.8740	0.9975	0.0956	0.8721	0.9286	250	2
	0.8726	0.9975	0.0943	0.8707	0.9275	500	2
	0.9306	0.9973	0.1548	0.9309	0.9625	50	4
	0.9324	0.9977	0.1586	0.9324	0.9635	100	4
	0.9303	0.9974	0.1562	0.9305	0.9623	250	4

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N Tree	M Try
	0.9308	0.9973	0.1566	0.9311	0.9626	500	4
	0.9623	0.9979	0.2412	0.9632	0.9801	50	7
	0.9633	0.9979	0.2523	0.9643	0.9807	100	7
	0.9626	0.9977	0.2463	0.9638	0.9804	250	7
	0.9645	0.9977	**0.2543**	**0.9658**	**0.9814**	500	7
	0.8653	**0.9969**	0.0849	0.8637	0.9235	50	2
	0.8679	0.9964	0.0860	0.8669	0.9253	100	2
	0.8695	0.9964	0.0871	0.8684	0.9259	250	2
	0.8703	0.9968	0.0880	0.8689	0.9264	500	2
	0.9320	0.9968	0.1541	0.9329	0.9634	50	4
$y_2 \prec y_1$	0.9298	0.9966	0.1488	0.9309	0.9622	100	4
	0.9303	0.9968	0.1503	0.9312	0.9625	250	4
	0.9297	0.9965	0.1482	0.9309	0.9622	500	4
	0.9594	0.9968	0.2307	0.9613	0.9787	50	7
	0.9640	0.9967	**0.2516**	**0.9662**	**0.9811**	100	7
	0.9624	0.9968	0.2427	0.9643	0.9803	250	7
	0.9625	0.9968	0.2435	0.9644	0.9803	500	7

Table 1 shows the performance of random forest with BR+ (stat) when SMOTE-NC is applied. The M try 7 is the optimum parameter for the optimum model for both orders. The number of trees for the optimum model for order $y_1 \prec y_2$ is 500 trees while for order $y_2 \prec y_1$ is 100 trees. The classification using SMOTE-NC as a method of handling the imbalance is produced good model. Because the model produces high accuracy, sensitivity, precision, F-Measure. Also, increase the specificity, which is mean that the model with SMOTE-NC can capture data in the minority class.

Table 2. Performance of BR+ (Stat) using Tomek Link

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N Tree	M Try
	0.9805	0.9805	NaN	**1.0000**	0.9901	50	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	100	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	250	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	500	2
	0.9827	0.9827	NaN	**1.0000**	0.9912	50	4
$y_1 \prec y_2$	0.9827	0.9827	NaN	**1.0000**	0.9912	100	4
	0.9825	0.9824	NaN	**1.0000**	0.9911	250	4
	0.9824	0.9823	NaN	**1.0000**	0.9910	500	4
	0.9871	0.9885	0.6887	0.9984	0.9934	50	7
	0.9872	0.9880	0.8403	0.9989	0.9934	100	7
	0.9879	**0.9889**	**0.8501**	0.9988	**0.9938**	250	7
	0.9876	0.9887	0.8436	0.9987	0.9937	500	7
	0.9805	0.9805	NaN	**1.0000**	0.9901	50	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	100	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	250	2
	0.9805	0.9805	NaN	**1.0000**	0.9901	500	2
$y_2 \prec y_1$	0.9825	0.9824	NaN	**1.0000**	0.9911	50	4
	0.9825	0.9824	NaN	**1.0000**	0.9911	100	4
	0.9823	0.9822	NaN	**1.0000**	0.9910	250	4
	0.9822	0.9821	NaN	**1.0000**	0.9909	500	4

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N Tree	M Try
	0.9862	0.9870	0.7592	0.9990	0.9929	50	7
	0.9866	0.9876	0.9224	0.9987	0.9931	100	7
	0.9866	0.9875	**0.9258**	0.9989	0.9931	250	7
	0.9868	**0.9877**	0.9153	0.9988	**0.9932**	500	7

The optimum M try for both orders is 7, and the optimum number of trees for order $y_1 \prec y_2$ is 250 trees and for order $y_2 \prec y_1$ is 500 trees. Tomek Link also produces good models because the specificity is higher compared to SMOTE-NC. Because the model produces high accuracy, sensitivity, precision, F-Measure. Also, increase the specificity from 0 to more than 90%, but in some conditions, Tomek Link cannot capture data in the minority class or cannot correctly classify data in the minority class.

Table 3. Performance of BR+ (Stat) using SMOTE-NC+T-Link

	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N Tree	M Try
$y_1 \prec y_2$	0.8659	0.9974	0.0898	0.8641	0.9241	50	2
	0.8742	0.9977	0.0960	0.8722	0.9287	100	2
	0.8742	0.9975	0.0991	0.8723	0.9288	250	2
	0.8707	0.9976	0.0939	0.8685	0.9263	500	2
	0.9331	0.9973	0.1584	0.9335	0.9639	50	4
	0.9330	0.9977	0.1609	0.9331	0.9639	100	4
	0.9309	0.9973	0.1567	0.9313	0.9627	250	4
	0.9302	0.9973	0.1548	0.9305	0.9623	500	4
	0.9625	**0.9979**	0.2441	0.9634	0.9802	50	7
	0.9628	0.9975	**0.2460**	**0.9642**	**0.9805**	100	7
	0.9614	0.9978	0.2439	0.9624	0.9797	250	7
	0.9624	0.9978	0.2499	0.9634	0.9802	500	7
$y_2 \prec y_1$	0.8660	0.9967	0.0866	0.8645	0.9237	50	2
	0.8679	0.9963	0.0857	0.8668	0.9249	100	2
	0.8732	0.9965	0.0893	0.8721	0.9282	250	2
	0.8713	0.9967	0.0880	0.8701	0.9272	500	2
	0.9295	0.9964	0.1478	0.9306	0.9620	50	4
	0.9300	0.9966	0.1497	0.9311	0.9624	100	4
	0.9305	0.9964	0.1488	0.9318	0.9626	250	4
	0.9308	**0.9969**	0.1511	0.9316	0.9628	500	4
	0.9622	0.9966	0.2417	0.9643	0.9802	50	7
	0.9620	0.9966	0.2429	0.9641	0.9801	100	7
	0.9627	0.9968	**0.2461**	**0.9647**	**0.9804**	250	7
	0.9621	0.9968	0.2438	0.9640	0.9801	500	7

Table 3 shows the performance of random forest with BR+ (stat) using hybrid sampling consist of SMOTE-NC and Tomek Link. The SMOTE-NC is applied to the imbalanced data and after the data balanced, then Tomek Link is applied to the balanced data. The results from Table 3 show that the optimum model for both orders is when using M try 7 and 100 trees for order $y_1 \prec y_2$ and 250 trees for order $y_2 \prec y_1$. The classification using hybrid sampling also produces good model. Because the model can capture the data in the minority class.

The performance comparison of the three resampling methods is shown in the following Table.

Table 4. Performance Comparison

			$y_1 \prec y_2$				
	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N tree	M try
SMOTE-NC	0.9645	**0.9977**	0.2543	0.9658	0.9814	500	7
T-Link	**0.9879**	0.9889	**0.8501**	**0.9988**	**0.9938**	250	7
SMOTE-NC+T-Link	0.9628	0.9975	0.2460	0.9642	0.9805	100	7
			$y_2 \prec y_1$				
	Accuracy	Sensitivity	Specificity	Precision	F-Measure	N tree	M try
SMOTE-NC	0.9640	0.9967	0.2516	0.9662	0.9811	100	7
T-Link	**0.9868**	0.9877	**0.9153**	**0.9988**	**0.9932**	500	7
SMOTE-NC+T-Link	0.9627	**0.9968**	0.2461	0.9647	0.9804	250	7

Table 4 shows the performance comparison of the optimum model from both orders and each resampling method. The model with SMOTE-NC and hybrid sampling to handle the imbalanced data produce high sensitivity compared to the model with Tomek Link, but Tomek Link has the highest accuracy, specificity, precision, and F-measure. The optimum model of Tomek Link can correctly classify the data in minority class as much as 100%, but in some conditions, Tomek Link cannot capture the data in the minority class. This is according to research by [12] where for extreme cases in imbalanced data (minority class less than 10%), Tomek Link shows the best performance as compared to hybrid sampling. Therefore, it can be concluded that the best method for order $y_1 \prec y_2$ is Tomek Link with 250 trees and M try is 7 and the best method for order $y_2 \prec y_1$ is also Tomek Link with M try is 7 and 500 trees.

4 Conclusion

According to the analysis that has been done before, it can be concluded that the optimum M try for the classification of multilabel data is 7 with a varying number of trees. Using the different order, the combination of BR+ (stat) and Tomek Link is the best method to overcome the multilabel imbalanced data as compared to the combination of BR+ (Stat) and SMOTE-NC and the combination of BR+ (stat) and hybrid sampling because the former has the highest accuracy, specificity, precision, and F-measure.

Acknowledgments. This research was funded by BKKBN. The authors thanks to BKKBN for funding which support this research and all individuals associated with this research work.

References

[1] Han J, Kamber M, Pei J. Data Mining Techniques, Third Edition 2011:847.

[2] Daniels ZA, Metaxas DN. Addressing imbalance in multi-label classification using structured hellinger forests. 31st AAAI Conf Artif Intell AAAI 2017 2017:1826–32.
[3] Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recognit 2012;45:3084–104. https://doi.org/10.1016/j.patcog.2012.03.004.
[4] Alvares-Cherman E, Metz J, Monard MC. Incorporating label dependency into the binary relevance framework for multi-label classification. Expert Syst Appl 2012;39:1647–55. https://doi.org/10.1016/j.eswa.2011.06.056.
[5] Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees. Int J Comput Sci Issues 2012;9:272–8.
[6] Breiman L. Random forests. Mach Learining 2001;45:5–32. https://doi.org/10.1201/9780429469275-8.
[7] Biau G. Analysis of a Random Forests Model. J Mach Learn 2012;13:1063–95. https://doi.org/10.5603/AIT.a2017.0074.
[8] Gu Q, Wang XM, Wu Z, Ning B, Xin CS. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. J Digit Inf Manag 2016;14:92–103.
[9] Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. Deep synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[10] Jian C, Gao J, Ao Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing 2016;193:115–22. https://doi.org/10.1016/j.neucom.2016.02.006.
[11] Elhassan T, Aljurf M, Al-Mohanna F, Shoukri M. Classification of Imbalance Data using Tomek Link(T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. J Informatics Data Min 2016;1:1–12. https://doi.org/10.21767/2472-1956.100011.
[12] Sain H, Purnami SW. Combine Sampling Support Vector Machine for Imbalanced Data Classification. Procedia Comput Sci 2015;72:59–66. https://doi.org/10.1016/j.procs.2015.12.105.