Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data

Tomasz Maciejewski and Jerzy Stefanowski
Institute of Computing Science
Poznań University of Technology
60-965 Poznań, Poland
jerzy.stefanowski@cs.put.poznan.pl, tomek.maciejewski@gmail.com

***Abstract*—In this paper we discuss problems of inducing classifiers from imbalanced data and improving recognition of minority class using focused resampling techniques. We are particularly interested in SMOTE over-sampling method that generates new synthetic examples from the minority class between the closest neighbours from this class. However, SMOTE could also overgeneralize the minority class region as it does not consider distribution of other neighbours from the majority classes. Therefore, we introduce a new generalization of SMOTE, called LN-SMOTE, which exploits more precisely information about the local neighbourhood of the considered examples. In the experiments we compare this method with original SMOTE and its two, the most related, other generalizations Borderline and Safe-Level SMOTE. All these pre-processing methods are applied together with either decision tree or Naive Bayes classifiers. The results show that the new LN-SMOTE method improves evaluation measures for the minority class.**

I. Introduction

Some real-life data mining problems involve learning classifiers from *imbalanced data*, which means that one of the classes (further called a *minority class*) includes much smaller number of examples than the others (further referred to as *majority classes*). Typical such problems are medical diagnosing dangerous illness, analysing financial risk, detecting oil spills in satellite images, predicting technical equipment failures or information filtering [**?**], [**?**]. Class imbalance constitutes a difficulty for most learning algorithms, which are biased toward learning and recognition of the majority classes. As a result, minority examples tend to be misclassified.

Our experiments were carried out on 14 data sets coming from UCI repository. Their basic characteristics is listed in Table **??**. The imbalance ratio (IMB in Table **??**) is calculated as a ratio of the number of minority examples to the total number of examples in a data set. We chose these data as they are characterized by varying degree of imbalance and they were often used in related experimental studies. Data sets with higher imbalanced ratio, as Slovenia breast cancer, were also chosen as they contained many noisy or borderline minority class examples. Some of these data sets originally included more than two classes, however, to focus more on minority vs majority characteristics and to simplify calculations we decided to aggregate all majority classes into one. In our opinion this aggregation does not influence the work of compared algorithms.

TABLE I
PRESENCE OF *danger* AND *noisy* EXAMPLES.

Dataset	MIN	D3	D5	D7	N3	N5	N7
Balance scale	49	0	4	18	49	45	31
Breast cancer	85	50	56	63	23	11	6
Cleveland	35	13	19	25	22	16	10
CMC	333	207	225	232	94	60	43
Ecoli	35	20	20	21	7	5	3
Flags	17	10	11	11	7	5	5
German credit	300	191	221	228	71	32	14
Haberman	81	51	53	57	20	14	8
Hepatitis	32	21	22	20	6	4	4
Pima	268	169	165	161	42	23	14
Post-operative	24	13	19	22	11	5	2
Solar flare	43	25	28	31	18	14	12
Transfusion	178	97	100	103	41	27	20
Yeast	51	25	28	32	24	20	18

Finally, let us discuss tuning parameters of methods. As to tree classifiers we used standard options with choosing unpruned version to strengthen the minority class. For SMOTE methods we have two parameters: k – number of neighbours and o – amount of oversampling (expressing how many times the minority class should be increased by oversampling)[1]. In related papers on SMOTE and its extensions k was set to 5 and o was usually stepwise changed to 5. In our experiments we decided to study wider range of these parameters. In case of k we tested the following values 3, 5 and 7. However, for o we decided to check more values of oversampling depending on data characteristics – for data with imbalanced ratio larger than 15% we tested following o values 1, 2, 3, 4, 5 and 6 while in case of data with the smaller imbalance ratio we additionally considered o values equal to 10, 15, 20 and 25. Moreover, we found an additional value which led to balancing cardinalities of both minority and majority classes. We prepared a batch procedure for testing all these combinations of k and o parameters for a given method and data. Among these results we chose this combination which led to the best value of F measure for the minority class (as it

[1]In the paper introducing original version of SMOTE [**?**] it was expressed in percentage, e.g. 500%

TABLE II
F-MEASURE FOR THE MINORITY CLASS FOR ALL COMPARED METHODS USED TOGETHER WITH C4.5 TREE CLASSIFIER

	None	**SMO**	**BS1**	**BS2**	**SLS**	**LN1**	**LN2**
Balance scale	0.00	9.29	8.40	11.33	8.58	16.54	16.08
Breast cancer	39.83	43.83	43.02	44.37	45.15	43.83	45.64
Cleveland	19.29	26.71	25.27	28.33	26.03	29.27	29.70
CMC	40.81	41.64	42.05	44.16	41.64	44.95	45.94
Ecoli	58.86	64.31	62.38	64.02	63.98	62.01	66.96
Flags	30.89	44.51	41.35	42.68	43.15	39.46	42.03
Germ. credit	45.51	50.30	49.98	51.01	50.02	50.91	50.46
Haberman	30.36	43.70	41.84	43.58	40.08	44.56	42.59
Hepatitis	49.20	52.10	53.94	53.00	57.10	58.57	57.86
Pima	62.05	65.51	65.68	65.61	65.02	65.13	65.06
Post-operative	5.84	22.03	22.86	19.06	20.56	20.42	19.44
Solar flare	28.79	27.84	28.85	29.93	28.68	31.60	33.08
Transfusion	47.27	48.80	50.05	51.12	48.94	49.19	50.30
Yeast	35.02	39.64	42.23	42.02	40.07	41.39	42.58

TABLE III
G-MEAN FOR ALL COMPARED METHODS WITH THE TREE CLASSIFIER

	None	**SMO**	**BS1**	**BS2**	**SLS**	**LN1**	**LN2**
Balance scale	0.00	23.36	23.93	27.58	21.10	43.93	41.21
Breast cancer	53.27	57.17	55.77	57.93	58.10	57.18	58.73
Cleveland	31.46	45.89	38.84	45.86	41.41	46.25	50.04
CMC	57.33	60.33	60.25	62.92	60.29	63.23	64.71
Ecoli	73.28	83.77	76.89	84.26	80.71	80.67	82.88
Flags	41.61	61.46	58.53	56.31	59.14	53.80	58.06
Germ. credit	58.83	63.00	62.70	63.55	62.81	63.57	63.11
Haberman	44.53	59.00	57.15	59.20	55.78	59.94	58.13
Hepatitis	63.17	66.43	66.21	67.75	69.81	70.05	69.97
Pima	69.84	72.73	72.80	72.60	72.47	72.33	72.56
Post-operative	8.23	26.44	26.25	26.47	25.86	27.52	26.67
Solar flare	39.28	45.71	46.77	47.03	45.78	50.85	57.33
Transfusion	60.79	63.18	65.53	65.53	63.71	63.95	63.67
Yeast	50.17	66.23	67.08	70.64	61.04	64.70	66.03

was also the main criterion in [**?**], [**?**]). For this combination of parameters we also calculated the remaining measures. As for each method this best configuration was found separately on each data set, the chosen values for o and k may vary between methods and data sets.