A Normal Distribution-Based Over-Sampling Approach to Imbalanced Data Classification

Huaxiang Zhang[1,2] and Zhichao Wang[1,2]

[1] Department of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China
[2] Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan, 250014, China
huaxzhang@hotmail.com

Abstract. This study proposes a normal distribution-based over-sampling approach to balance the number of instances belonging to different classes in a data set. The balanced training data are used to learn unbiased classifiers for the original data set. Under some conditions, the proposed over-sampling approach generates samples with expected mean and variance similar to that of the original minority class data. As the approach tries to generate synthetic data with similar probability distributions to the original data, and expands the class boundaries of the minority class, it may increase the minority class classification performance. Experimental results show that the proposed approach outperforms alternative methods on benchmark data sets most of the times when implementing several classical classification algorithms.

Keywords: imbalanced classification, over-sampling, normal distribution.

1 Introduction

Imbalance data sets [1] refer to those that the number of one class instances is far more than that of another class instances for each concept in the training data. It is quite common in practice and automatic concept learning from imbalanced data sets usually produces biased classifiers with high predictive accuracy on the majority class data, but poor predictive accuracy on the minority class data [2].

Learning from imbalanced datasets has been explored extensively, and the existing works can be divided into data and algorithmic levels. The algorithm-level approaches use the original training data to construct new algorithms suitable to imbalanced data sets, and data-level approaches[3-6] generate new training datasets from the original datasets to make their class distribution approach balanced. As under-sampling may discard potentially useful data that could be important for classification, and over-sampling may change the original minority class sampling distribution.

Over-sampling refers to the process of generating more training instances for the minority class to balance the class distribution. As research results show that over-sampling with replication does not significantly improve the classification accuracy of the minority class data. A novel approach SMOTE [7] is proposed to

J. Tang et al. (Eds.): ADMA 2011, Part I, LNAI 7120, pp. 83–96, 2011.

overcome this issue. SMOTE is a linear over-sampling method which adopts linear interpolations between two near samples to generate synthetic samples. SMOTE ignores the change of the underlying probability distribution of the minority class data after synthetic samples have been included in the training data. Borderline-SMOTE methods [8] in which only the minority samples near the borderline are over-sampled. Borderline-SMOTE produces little change in performance and sometimes hurts the generalization of an algorithm, as it changes the minority data distribution. DataBoost-IM [9] combined synthetic data generation and an ensemble learning algorithm to tackle the imbalanced data classification problem. Random Over-Sampling (RO-Sampling) is a non-heuristic method that aims to balance class distribution through randomly duplicating the minority class instances. After the minority class instances have been processed, there will be several exact copies of some minority class instances in the new training data set, and this increases the likelihood of occurring over-fitting. Combination of over-sampling and under-sampling is often performed to resolve the imbalance problems [10] [11].

It is reported that under-sampling produces a reasonable sensitivity to changes in class distribution, and over-sampling often produces little or no change in performance as the training data distribution has been changed [12]. We just focus on the over-sampling approaches to handling imbalanced data classification, and ignore attribute processing methods [13] [14] in this work.

All the above mentioned data processing methods, whether they perform over-sampling, under-sampling or combinations of the both on the original datasets, will result in changing the sampling distribution of the original data, thus leads to a biased classifier that is not quite suitable to classify the original data.

We propose a novel over-sampling approach called normal distribution-based over-sampling (NDO-Sampling). To our knowledge, it differs from all the proposed over-sampling approaches in the related literature, such as non-heuristic over-sampling and heuristic over-sampling, and it neither performs linear interpolations between neighbors nor increases the number of minority class data with replacement, it generate synthetic samples through implementing a random walk starting from the original data point. Under some assumptions and when some conditions are satisfied, the expected value and the squared deviation of the synthetic samples can be proved to be probably approximate the same as that of the original minority class data. Our method decreases the risk of increasing the likelihood among instances in the minority class.

We organize the paper as follows. Section 2 describes the evaluation metrics commonly used in imbalanced data classification. Section 3 presents the normal distribution-based over-sampling model, and proves two theorems concerning the mean and the squared deviation of the synthetic samples. Section 4 evaluates NDO-Sampling on broad data sets from different aspects. Section 5 concludes the paper.

2 Performance Evaluation Metrics

The overall accuracy on a test dataset is commonly used to evaluate the performance of a classifier. But for imbalanced data, as the accuracy is profoundly dominated by the minority class data, alternative evaluation metrics are employed. These appropriate

metrics include Area Under the ROC Curve (*AUC*), *F-Measure*, Geometric Mean (*G-Mean*), *overall accuracy* and the accuracy rate for the minority class. We refer minority class to positive class and majority class to negative class. After the four values *TP*(the number of true positive), *FP*(the number of false positive), *TN*(the number of true negative) and *FN*(the number of false negative) have obtained, *precision* and *recall* are calculated as *TP/(TP+FP)* and *TP/(TP+FN)* respectively, positive accuracy(*pa*) and negative accuracy(*na*) are calculated as *TP/(TP+FN)* and *TN/(TN+FP)* respectively, and then *F-Measure* and *G-Mean* are defined as

$$F\text{-}Measure=\frac{(1+\beta^2)\times recall\times precision}{(\beta^2\times recall+precision)} \quad . \; G\text{-}Mean=\sqrt{pa\times na} \quad .$$

We set β to 1 in this paper.

The above metrics represent different aspects of the learning algorithms, and are extensively used in the field of imbalanced data classification problems. *F-measure* [15] integrates *precision* and *recall* into a single metric, and measures how "good" a learning algorithm on the interested class. It is high when both the *recall* and *precision* are high. The ROC curve indicates a balanced classification ability of a classifier by considering the tradeoffs between *TP Rate* and *FP Rate* [16]. *G-Mean* [17] measures whether the accuracy on each of the two classes are maximized.

We use *F-measures* for both majority class and minority class, the *overall accuracy*, *G-mean* and *TP rate* as the evaluation metrics in this work.

3 The Normal Distribution Model

The central limit theorem states that no matter what the real sampling distribution is, the sampling distribution of the mean approaches a normal distribution. This inspires us to create synthetic samples for the minority class without knowing the real sampling distribution. We aim to generate samples approximately obeying the real sampling distribution. After the newly generated samples are put together with the original minority class instances, these newly formed training data will keep the original sampling distribution approximately unchanged.

We first make the following attribute independence assumption: each attribute of the training data is considered as a random variable, and all the attributes are independent of each other. Given the m attributes, denoted as a_1, a_2,…, a_m, we have m random variables. Based on the given minority class data, we calculate the expected value and variance for each random variable, and the mean and the standard variance of a_i are denoted as μ_i and σ_i respectively, where $i\in\{1,2,..,m\}$.

Let μ_i' denote the mean of the unknown underlying distribution governing random variable a_i and σ_i' be the standard deviation. We say that all the values of attribute a_i for the minority training data are independent, identically distributed random variable values, because they represent independent experiments, and each obeying the similar underlying probability distribution. According the conclusion of central limit theorem,

we know that, as the number n of samples approaches infinite, the distribution governing the following random variable approaches a Normal distribution, with zero mean and standard deviation equals 1.

$$\frac{\mu_i - \mu_i^{'}}{\sigma_i^{'}/\sqrt{n}} \xrightarrow{P} N(0,1) \quad . \tag{1}$$

n is the number of minority class instances. Inspired by (1), given the value r_i of a random variable obeying distribution $N(0,1)$, we have the following equation:

$$\mu_i^{'} = \mu_i - r_i \bullet \sigma_i^{'}/\sqrt{n} \quad . \tag{2}$$

In (2), μ_i is the mean of attribute a_i for the given training minority class data, and we consider it as the representative of the original minority class data. μ_i' is the mean of attribute a_i for the unknown minority class data, and we consider it as the representative of the unknown minority class data. So for any instance and its given value of a_i, we can generate a synthetic value for this attribute through the following calculation.

$$a_i^{'} = a_i - r_i \bullet \sigma_i^{'}/\sqrt{n}\,, \quad i \in \{1,2,...,m\} \quad . \tag{3}$$

In (3), a_i' is a new value of attribute a_i. σ_i' is unknown, we use σ_i to approximate it, and obtain equation (4)

$$a_i^{'} = a_i - r_i \bullet \sigma_i/\sqrt{n}\,, \quad i \in \{1,2,...,m\} \quad . \tag{4}$$

We call (4) a normal distribution model. Based on this model, we propose a normal distribution-based over-sampling approach.

We obtain two conclusions from (4).

Theorem 1. *As n approaches infinite, the expected mean of the random variable values of attribute a_i obtained using (4) equals μ_i.*

Theorem 2. *The expected standard deviation of the random variable values of attribute a_i obtained using (4) equals σ_i as n approaches infinite.*

The conclusions of the above two theorems (duo to the limit to the range of pages, the related proofs were not shown in this paper) tell us, if we generate a synthetic value for each attribute according to (3), we get m synthetic values, and the m values form a vector, which can be considered as a synthetic training instance. Given the over-sampling rate, we create required number of instances for the minority class, and construct a classifier using the minority class training data together with the synthetic data. As the expected mean and the expected squared deviation of the synthetic samples equal the mean and the standard deviation of the original minority class data correspondingly, the constructed classifier will be suitable to the training data well and will be an unbiased classifier.

4 Experimental Results

All the experiments are conducted on 8 benchmark datasets extensively used in classification tasks. These datasets are selected from the UCI Machine Learning Repository[1] and are summarized in table 1. They have different data sizes and degrees of skew, and come from different domains, thus making the experimental results much more convincing. We do several sets of experiments in order to evaluate NDO-Sampling from different aspects. We use *TP rate*, *F-measure*, *G-mean* and *overall accuracy* to evaluate the results of our experiments, and study the impacts of varying imbalance ratios. In the first set of experiments, we employ three baseline classifier algorithms to compare NDO-Sampling with SMOTE and RO-Sampling. These algorithms are C4.5, NB (Naive Bayes) and KNN(k=3). Each experiment is done 10 times independently for each baseline algorithm, and in each time, a ten-fold cross validation is applied. The end results for each algorithm are the average of the 10 independent ten-fold cross validation experiments. In the second set of experiments, we study the impacts of different over-sampling approaches on multi-class classification problems.

The minority class is over-sampled at 100%, 200% and 300% of its original size, and the number of nearest neighbors is set to 5 for SMOTE. For the Glass, Satimage, Segment-challenge and Vehicle, we choose the smallest class as the positive class and convert the rest of the classes into a single class as the negative class to increase the degree of skew. The datasets are described in table one.

Table 1. Summary of datasets (including dataset, number of instances, number of minority class instances, number of attributes and minority class label)

dataset	# of inst.	# of min. inst.(%)	# of feat.	target
Breast-w	699	241(34.48)	9	class=malignant
Diabetes	768	268(34.90)	8	class=tested positive
Glass	214	9(4.21)	9	type=tableware
Ionosphere	351	126(35.90)	34	class=b
Satimage	6430	625(9.72)	36	class=2.
Segment-challenge	1500	205(13.67)	19	class=brick face
Sonar	208	97(46.63)	60	class=Rock
Vehicle	846	199(23.52)	18	class=van

We use Weka[2] to implement the experiments. Results for the above data sets are shown in table 2. Different baseline algorithms are executed at different over-sampling rates. We use F-min to denote *F-measure* for the minority class, F-maj to denote *F-measure* for the majority class, O-acc. to denote the *overall accuracy*, Alg. to denote algorithms, NDO, SMO and RO to denote the over-sampling strategies normal distribution over-sampling, SMOTE and random over-sampling respectively, and O.S to denote Over-Sampling strategies in the following tables and figures.

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html
[2] http://www.cs.waikato.ac.nz/ml/weka

Table 2. (1). Results on 8 datasets with continuous attributes (over-sampling rate at 100% and 200%)

dataset	Alg.	O.S	100%					200%				
			F-min (%)	F-maj (%)	O-acc(%)	G-mean (%)	TPrate (%)	F-min (%)	F-maj (%)	O-acc (%)	G-mean (%)	TPrate (%)
Breast-w	C4.5	NDO	**93.12**	**96.24**	**95.14**	**95.23**	**95.52**	**92.95**	**96.10**	**94.98**	**95.22**	**96.02**
		SMO	92.80	96.11	94.95	94.83	94.47	92.71	96.02	94.85	94.89	95.02
		RO	92.39	95.94	94.71	94.35	93.22	92.14	95.73	94.47	94.37	94.05
	NB	NDO	94.38	96.89	95.99	96.35	97.51	94.38	96.89	95.99	96.35	97.51
		SMO	94.38	96.89	95.99	96.35	97.51	94.38	96.89	95.99	96.35	97.51
		RO	94.38	96.89	95.99	96.35	97.51	94.38	96.89	95.99	96.35	97.51
	3-NN	NDO	95.94	97.77	97.12	97.45	98.51	96.39	98.00	97.42	**97.96**	**99.75**
		SMO	**96.36**	**98.01**	**97.42**	**97.73**	**98.76**	**96.50**	**98.08**	**97.52**	97.90	99.17
		RO	95.56	97.62	96.90	96.88	96.82	95.21	97.38	96.61	96.82	97.51
Diabetes	C4.5	NDO	**67.31**	**78.52**	**74.08**	**74.61**	**76.49**	**66.43**	**77.04**	**72.73**	**73.71**	77.31
		SMO	66.30	77.78	73.22	73.73	75.50	64.84	74.50	70.44	71.98	**78.11**
		RO	60.52	76.88	70.83	69.06	64.05	60.18	75.66	69.79	68.70	65.42
	NB	NDO	65.93	79.73	74.58	73.57	**70.49**	**66.74**	**77.78**	**73.36**	**74.07**	**76.60**
		SMO	**66.27**	**80.25**	**75.09**	**73.84**	70.15	65.59	76.71	72.22	73.01	75.87
		RO	65.57	79.95	74.65	73.26	69.15	65.35	77.36	72.61	72.93	74.00
	3-NN	NDO	65.03	75.61	71.26	72.38	**76.57**	**74.42**	**78.38**	**76.56**	**79.83**	**97.69**
		SMO	**65.14**	**76.12**	**71.66**	**72.56**	75.87	65.73	73.68	70.23	72.37	81.84
		RO	60.07	74.21	68.66	68.40	67.54	60.54	69.69	65.71	67.55	75.37
Glass	C4.5	NDO	94.74	99.76	99.53	99.76	100.00	94.74	99.76	99.53	99.76	100.00
		SMO	94.74	99.76	99.53	99.76	100.00	94.74	99.76	99.53	99.76	100.00
		RO	94.74	99.76	99.53	99.76	100.00	94.74	99.76	99.53	99.76	100.00
	NB	NDO	68.44	97.93	96.12	**97.95**	**100.00**	66.91	97.78	95.84	**97.81**	**100.00**
		SMO	**69.44**	**98.18**	**96.57**	94.65	92.59	**71.64**	**98.44**	**97.04**	93.05	88.89
		RO	64.86	97.85	95.95	92.50	88.89	60.76	97.43	95.17	92.11	88.89
	3-NN	NDO	73.30	98.55	97.24	**93.70**	**90.00**	**74.69**	**98.49**	**97.15**	**98.50**	**100.00**
		SMO	**73.85**	**98.61**	**97.35**	93.20	88.89	69.57	98.27	96.73	92.89	88.89
		RO	71.64	98.44	97.04	93.05	88.89	62.16	97.69	95.64	90.48	85.19

Table 2. (1). (*Continued*)

Ionosphere	C4.5	NDO	86.71	92.72	90.60	89.38	85.48	**86.40**	**92.82**	**90.60**	88.78	83.17
		SMO	85.45	91.73	89.46	88.72	**86.24**	86.43	91.76	89.74	**90.02**	**91.01**
		RO	**87.60**	**93.26**	**91.26**	**90.01**	85.98	83.99	90.92	88.41	87.54	84.66
	NB	NDO	**82.34**	**89.62**	**86.92**	**86.47**	84.92	**82.63**	**89.84**	**87.18**	**86.66**	84.92
		SMO	80.54	87.79	85.00	85.32	86.51	81.44	88.56	85.85	85.99	86.51
		RO	77.09	84.46	81.48	82.54	**86.77**	76.74	83.95	81.01	82.25	**87.30**
	3-NN	NDO	85.79	92.78	90.43	87.90	80.48	**95.57**	**97.40**	**96.72**	**97.10**	**98.49**
		SMO	**90.61**	**94.97**	**93.45**	**92.18**	**88.10**	95.02	97.05	96.30	96.75	98.41
		RO	82.11	91.43	88.41	84.52	74.07	84.00	92.03	89.36	86.34	77.78
Satimage	C4.5	NDO	63.38	95.56	92.08	**81.61**	**70.56**	**68.70**	**96.05**	**92.99**	**86.50**	**79.20**
		SMO	**63.39**	**95.64**	**92.20**	81.07	69.44	63.91	95.58	92.13	82.23	71.68
		RO	55.57	95.18	91.30	72.98	56.00	57.25	95.25	91.45	74.77	58.88
	NB	NDO	47.98	88.78	81.54	**84.18**	**87.60**	47.59	88.57	81.23	84.03	87.68
		SMO	**48.87**	**89.40**	**82.44**	84.15	86.35	**48.76**	**89.32**	**82.33**	**84.16**	86.51
		RO	47.96	88.81	81.58	84.10	87.36	47.52	88.50	81.13	84.06	**87.89**
	3-NN	NDO	68.90	95.74	92.50	89.27	85.44	**74.67**	**96.21**	**93.41**	**96.28**	**100.00**
		SMO	**72.17**	**96.08**	**93.12**	**92.52**	**91.79**	71.40	95.58	92.34	94.98	98.40
		RO	67.18	95.75	92.47	86.26	79.25	65.55	95.08	91.39	88.12	84.27
Segment-challenge	C4.5	NDO	**97.23**	**99.56**	**99.24**	98.59	97.71	97.82	99.65	99.40	**98.97**	**98.39**
		SMO	96.92	99.51	99.16	98.27	97.07	**97.96**	**99.68**	**99.44**	98.65	97.56
		RO	96.86	99.50	99.13	**98.61**	**97.89**	96.84	99.50	99.13	98.26	97.07
	NB	NDO	**65.47**	**91.79**	**86.73**	**89.55**	**98.37**	**66.55**	**92.22**	**87.38**	**89.23**	**99.02**
		SMO	62.44	89.69	83.82	88.91	92.03	61.90	89.41	83.42	89.36	98.54
		RO	61.24	89.08	82.96	89.06	98.04	61.67	89.22	83.18	89.38	98.87
	3-NN	NDO	**97.91**	**99.66**	**99.42**	**99.35**	**99.24**	**97.82**	**99.65**	**99.39**	**99.62**	**99.95**
		SMO	97.44	99.59	99.29	99.11	98.86	97.53	99.60	99.31	99.33	99.35
		RO	96.90	99.50	99.13	99.16	99.19	96.54	99.43	99.02	99.30	99.67
Sonar	C4.5	NDO	**82.90**	**80.62**	**81.83**	**81.77**	**94.43**	**84.40**	**82.07**	**83.32**	**83.21**	**96.80**
		SMO	78.15	77.94	78.04	78.22	84.19	81.75	80.72	81.25	81.39	90.03
		RO	70.63	76.48	73.88	73.21	67.35	72.70	76.67	74.84	74.60	71.82

Table 2. (1). (*Continued*)

	NB	NDO	**71.85**	63.77	**68.32**	67.31	**86.70**	**73.90**	**65.13**	**70.14**	**68.81**	**90.62**
		SMO	70.40	64.67	67.79	67.37	82.13	70.41	65.03	67.95	67.59	81.79
		RO	70.86	**64.78**	68.11	**67.60**	83.16	71.14	64.77	68.27	67.70	83.85
	3-NN	NDO	85.96	85.96	85.96	86.16	92.16	**89.23**	**88.43**	**88.85**	**88.98**	**99.07**
		SMO	**88.31**	**88.61**	**88.46**	**88.65**	**93.47**	87.66	87.34	87.50	87.69	95.19
		RO	87.31	87.99	87.66	87.82	91.07	86.54	86.54	86.54	86.74	92.78
Vehicle	C4.5	NDO	**88.47**	**96.25**	**94.34**	**93.64**	**92.36**	**87.17**	**95.70**	**93.56**	**93.39**	**93.07**
		SMO	86.80	95.72	93.54	92.39	90.28	86.40	95.56	93.30	92.30	90.45
		RO	86.76	95.79	93.62	91.95	88.94	86.96	95.91	93.77	91.80	88.27
	NB	NDO	55.66	72.06	65.72	72.71	91.46	**56.63**	**71.98**	**65.96**	**73.50**	**94.47**
		SMO	**56.26**	**72.35**	**66.12**	**73.27**	**92.63**	56.27	71.87	65.76	73.18	93.63
		RO	54.52	71.93	65.29	71.72	88.44	55.20	71.82	65.41	72.28	90.62
	3-NN	NDO	86.40	95.26	92.97	**93.64**	**94.92**	**88.36**	**95.78**	**93.80**	**95.86**	**100.00**
		SMO	**87.00**	**95.56**	**93.38**	93.64	94.14	86.97	95.42	93.22	94.22	96.15
		RO	85.29	94.98	92.51	92.44	92.29	84.48	94.43	91.80	92.82	94.81

Table 2 (2). Results on 8 datasets with numerical attributes(over-sampling rate at 300%)

dataset	Alg.	O.S	300%					times of win				
			F-min (%)	F-maj (%)	o-acc. (%)	G-mean (%)	TPrate (%)	F-min (%)	F-maj (%)	O-acc (%).	G-mean (%)	TP rate (%)
Breast-w	C4.5	NDO	93.04	96.13	95.02	95.36	**96.47**	2	2	2	2	3
		SMO	**93.52**	**96.46**	**95.42**	**95.52**	95.85	1	1	1	1	0
		RO	92.40	95.83	94.61	94.71	95.02	0	0	0	0	0
	NB	NDO	94.38	96.89	95.99	96.35	97.51	-	-	-	-	-
		SMO	94.38	96.89	95.99	96.35	97.51	-	-	-	-	-
		RO	94.38	96.89	95.99	96.35	97.51	-	-	-	-	-
	3-NN	NDO	96.39	98.00	97.42	**97.96**	**99.75**	-	-	-	2	2
		SMO	96.39	98.00	97.42	97.92	99.59	2+-	2+-	2+-	1	1
		RO	95.44	97.48	96.76	97.16	98.48	0	0	0	0	0
Diabetes	C4.5	NDO	**66.86**	**77.76**	**73.39**	**74.16**	76.94	3	3	3	3	1
		SMO	66.34	73.72	70.49	72.80	**83.33**	0	0	0	0	2
		RO	60.84	76.26	70.44	69.27	65.80	0	0	0	0	0

Table 2. (2). (*Continued*)

	NB	NDO	**66.44**	74.95	71.32	73.24	81.38	2	1	1	1	2
		SMO	66.33	74.48	70.96	73.03	**81.97**	1	1	1	1	1
		RO	66.39	**76.14**	**72.09**	**73.50**	78.98	0	1	1	1	0
	3-NN	NDO	**73.53**	**76.92**	**75.34**	**78.70**	**98.13**	2	2	2	2	3
		SMO	67.82	73.55	70.96	73.73	87.69	1	1	1	1	0
		RO	61.43	68.79	65.49	67.81	78.73	0	0	0	0	0
Glass	C4.5	NDO	94.74	91.53	99.59	99.22	99.59	-	-	-	-	-
		SMO	94.74	94.74	99.76	99.53	99.76	1+-	1+-	1+-	1+-	-
		RO	94.74	94.74	99.76	99.53	99.76	-	-	-	-	-
	NB	NDO	68.44	64.98	97.58	95.47	**97.61**	0	0	0	2	3
		SMO	**69.44**	**71.64**	**98.44**	**97.04**	93.05	3	3	3	1	0
		RO	64.86	64.00	97.77	95.79	92.42	0	0	0	0	0
	3-NN	NDO	73.30	**73.77**	**98.41**	**97.01**	**98.43**	1	2	2	3	3
		SMO	**73.85**	67.57	98.02	96.26	94.49	2	1	1	0	0
		RO	71.64	65.75	97.94	96.11	92.58	0	0	0	0	0
Ionosphere	C4.5	NDO	**86.22**	**92.80**	**90.54**	88.54	82.46	2	2	2	0	0
		SMO	85.50	90.87	88.79	**89.48**	**92.06**	0	0	0	2	3
		RO	83.22	90.82	88.13	86.65	82.01	1	1	1	1	0
	NB	NDO	**82.91**	**90.07**	**87.44**	86.86	84.92	3	3	3	2	0
		SMO	82.71	89.45	86.89	**86.98**	**87.30**	0	0	-	1	1
		RO	76.62	83.74	80.82	82.13	87.57	0	0	-	0	2
	3-NN	NDO	**95.42**	**97.31**	**96.61**	**97.01**	98.49	2	2	2	2	1
		SMO	95.04	97.04	96.30	96.86	**98.94**	1	1	1	1	2
		RO	84.99	92.43	89.93	87.22	79.37	0	0	0	0	0
Satimage	C4.5	NDO	**67.20**	**95.70**	**92.40**	**86.67**	**80.16**	2	2	2	3	3
		SMO	65.13	95.59	92.17	84.08	75.20	1	1	1	0	0
		RO	57.04	95.30	91.53	74.17	57.81	0	0	0	0	0
	NB	NDO	47.15	88.34	80.89	83.84	87.68	0	0	0	1	3
		SMO	**48.82**	**89.31**	**82.32**	**84.27**	86.77	3	3	3	2	0
		RO	47.49	88.48	81.10	84.05	**87.89**	0	0	0	0	2

Table 2. (2). (*Continued*)

	3-NN	NDO	**74.18**	**96.11**	**93.23**	**96.18**	**100.00**	2	2	2	2	2
		SMO	69.51	95.06	91.50	95.02	99.63	1	1	1	1	1
		RO	64.87	94.80	90.94	88.73	86.08	0	0	0	0	0
Segment-challenge	C4.5	NDO	**97.75**	**99.64**	**99.38**	**99.07**	**98.63**	2	2	2	2	2
		SMO	97.18	99.55	99.22	98.66	97.89	1	1	1	0	0
		RO	97.24	99.56	99.24	98.53	97.56	0	0	0	1	1
	NB	NDO	**67.19**	**92.47**	**87.75**	**89.41**	91.76	3	3	3	3	2
		SMO	61.15	89.03	82.89	89.01	98.54	0	0	0	0	0
		RO	60.81	88.78	82.56	88.98	**99.02**	0	0	0	0	1
	3-NN	NDO	**97.59**	**99.61**	**99.33**	**99.59**	**99.95**	3	3	3	3	3
		SMO	97.45	99.59	99.29	99.38	99.51	0	0	0	0	0
		RO	96.23	99.38	98.93	99.24	99.67	0	0	0	0	0
Sonar	C4.5	NDO	**84.17**	**81.14**	**82.79**	**82.51**	**98.14**	3	3	3	3	3
		SMO	82.12	79.93	81.09	81.07	93.13	0	0	0	0	0
		RO	69.86	75.15	72.76	72.28	67.70	0	0	0	0	0
	NB	NDO	**73.89**	65.02	**70.10**	**68.73**	**90.72**	3	1	3	2	3
		SMO	70.43	**65.39**	68.11	67.81	81.44	0	1	0	0	0
		RO	70.33	63.20	67.15	66.43	83.51	0	1	0	1	0
	3-NN	NDO	**89.27**	**88.49**	**88.89**	**89.03**	**99.07**	2	2	2	2	3
		SMO	88.65	87.93	88.30	88.45	97.94	1	1	1	1	0
		RO	87.94	87.70	87.82	88.01	95.19	0	0	0	0	0
Vehicle	C4.5	NDO	**86.70**	**95.57**	**93.36**	**92.90**	**92.06**	3	3	3	3	3
		SMO	85.69	95.35	92.99	91.67	89.28	0	0	0	0	0
		RO	85.55	95.33	92.95	91.46	88.78	0	0	0	0	0
	NB	NDO	**56.46**	71.73	65.72	73.30	**94.47**	2	1	1	1	2
		SMO	56.45	**71.90**	**65.84**	**73.34**	94.14	1	2	2	2	1
		RO	56.38	71.71	65.68	73.24	94.30	0	0	0	0	0
	3-NN	NDO	**88.12**	**95.67**	**93.66**	**95.76**	**100.00**	2	2	2	3	3
		SMO	87.29	95.45	93.30	94.82	97.82	1	1	1	0	0
		RO	84.87	94.49	91.92	93.40	96.31	0	0	0	0	0

Note: “n+-” denotes winning n times, and drawing with the other two approaches the left 3-n times; ‘-’ indicates drawing with the other one or two over-sampling approaches

The results described in table 2 reveal that the performance of each over-sampling strategy varies when implementing different baseline algorithms on a same data set, and the over-sampling rate influences the performance of an over-sampling approach too. We select dataset Diabetes as an example, if the over-sampling rate is set to 100%, NDO-Sampling outperforms SMOTE and RO-Sampling in terms of the listed five evaluation metrics when using C4.5 as the baseline classifier, but SMOTE performs the best in terms of the first four metrics when NB is executed, and NDO-Sampling win in term of TP rate. When the baseline algorithm is fixed to 3-NN, we find that SMOTE outperforms NDO-Sampling on the first four metrics when the over-sampling rate is set to 100%, but it lost when the over-sampling rate is set to 200%. The results reveal that the metric values are influenced by both the over-sampling rate and the classification algorithm. The above phenomenon occurs on almost all the datasets listed in table two.

The results shown in table 2 indicate that NDO-Sampling performs well against imbalanced data sets with continuous attributes. In many cases, our over-sampling strategy yields results in terms of the *F-measure* for minority, the *G-mean* and *TP rate* comparable or slightly higher than that produced by the other two approaches. It also achieves good results when the *F-measure* for majority class and the *overall accuracy* are considered. For the highest dimensional data set Sonar, the five metrics surpass that of SMOTE and random over-sampling all the time when C4.5 is conducted, the minority class *F-measure*, the *overall accuracy* and *TP rate* surpass that of SMOTE and random over-sampling all the time when NB is conducted. Our approach lost in terms of the five metrics only when the over-sampling rate is at 100% when implementing 3-NN. For the highly imbalanced data set Satimage, NDO-Sampling yields good results in terms of the five metric when C4.5 is implemented except that, it lost in terms of both of the *F-measures* and the *overall accuracy* when the over-sampling rate is set to 100%. SMOTE performs well when NB is conducted. When 3-NN is conducted, NDO-Sampling wins two times in terms of the five metrics except when the over-sampling rate is at 100%.

In order to facilitate the analysis of the results on the 8 data sets, we summarized the numbers of wins in table 2 for the three over-sampling approaches and show the results in table 3.

Table 3. Number of wins for 8 datasets with continuous attribute when implementing c4.5, NB and 3-NN

Alg.	O.S	F-min	F-maj	O-acc.	G-mean	TP rate
C4.5	NDO	**20**	**17**	**16**	**16**	**15**
	SMO	3	3	3	5	5
	RO	1	1	1	2	1
NB	NDO	**11**	8	**9**	**9**	**11**
	SMO	8	**10**	9	7	3
	RO	0	2	1	2	5
3-NN	NDO	**13**	**14**	**14**	**18**	**18**
	SMO	10	9	9	6	5
	RO	0	0	0	0	0

The results described in table 3 show that NDO-Sampling outperforms both SMOTE and RO-Sampling in terms of the *F-measures* of both the minority and majority class, the *overall accuracy*, *G-mean* and *TP rate* when C4.5 and KNN (k=3) are conducted. NDO-Sampling outperforms both SMOTE and RO-Sampling in terms of the four evaluation metrics (*F-measure for minority, Overall accuracy, G-mean and TP rate*) when NB is conducted, it only loses to SMOTE in *F-measure* for majority.

RO-Sampling performs worst among the three over-sampling strategies regardless of what classification algorithms are implemented. It loses all the time when 3-NN is conducted, because when RO-Sampling is employed to generate synthetic samples, it just duplicates the original samples randomly, and no real new samples are generated. When implementing NDO-Sampling and SMOTE to generate synthetic samples for the minority class, newly generated samples may become the nearest neighbors of a newly coming pattern, thus resulting in the increasing of the classification accuracy of KNN.

From the experimental results described in tables from table two and table three, we can conclude that NDO-Sampling has good potential in handling imbalanced data regardless of what learning algorithms have been used. NDO-Sampling also has good potential in handling imbalanced datasets with discrete attributes. Duo to the limit to the range of pages, the related work was not shown in this paper.

In order to evaluate the performance of NDO-Sampling on multiple-class classification problems, we conduct experiments on the original data set Satimage. Different over-sampling approaches are implemented on the class with the least number of samples. As the metrics such as *G-mean*, *F-measure* and *TP rate* have no meanings for multi-class classification problems, we just obtain the overall classification accuracies when different baseline algorithms are executed at different over-sampling rates. The averaged results of three experiments are shown in figure one.

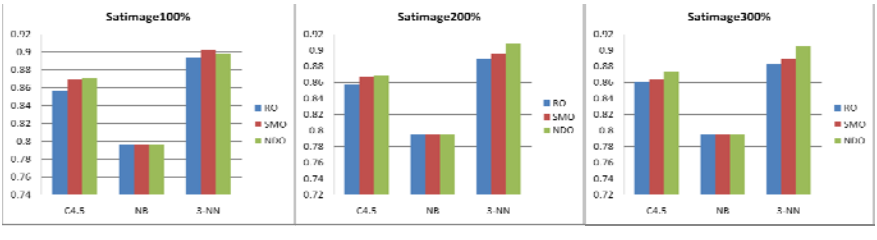

Fig. 1. Overall accuracies of multiple class classification on Satimage for different over-sampling strategies

For any given classification algorithm, obvious differences are not found in the overall accuracies at different over-sampling rates from figure two, because over-sampling is implemented on the minority class and the overall classification accuracies are mainly dominated by the other classes. Different over-sampling strategies obtain different classification accuracies, and NDO-Sampling performs the best most of the time.

5 Conclusions

Compared with SMOTE, NDO-Sampling reduces the computational complexity greatly, as it does not need to perform the nearest neighbor algorithm, a time-consuming approach, to obtain the k nearest neighbors before creating synthetic samples. SMOTE generates synthetic samples on the line between two neighbors, it can be considered as a linear interpolation approach. The newly generated samples does not hold the conclusions of the theorem one and two simultaneously, a necessary condition for the synthetic data and the original data to obey the same probability distribution. Random over-sampling re-samples the original minority class data with replacement, it is at the risk of over-fitting, but NDO-Sampling can avoid increasing the likelihood of occurring over-fitting. The key idea of NDO-Sampling is that it generates synthetic data that share an approximately similar probability distribution with that of the original minority data. In this paper, we just focus on the approach of over-sampling the minority class and ignore what a specific learning algorithm is conducted on the data sets. In our future work, considerations will be taken on algorithm-related normal distribution over-sampling approaches, such as just performing NDO-Sampling on the border samples for some specific algorithms.

Acknowledgments. This research is partially supported by the National Natural Science Foundation of China (No. 61170145), the Science and Technology Projects of Shandong Province, China (ZR2010FM021, and 2010G0020115) and Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology.

References

1. Barandela, R., Sanchez, J.S., Garcia, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36, 849–851 (2003)
2. Zhou, Z.-H., Liu, X.-Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18, 63–77 (2006)
3. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter 6(1), 20–29 (2004)
4. Sun, A., Lim, E.-P., Liu, Y.: On Strategies for Imbalanced Text Classification Using SVM: A Comparative Study. Decision Support Systems 48(1), 191–201 (2009)
5. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 301–312 (2006)
6. Yen, S.-J., Lee, Y.-S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications 36, 5718–5727 (2009)
7. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
8. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005)

9. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. SIGKDD Explorations Newsletter 6, 30–39 (2004)
10. Estabrooks, A., Jo, T., Japkowicz, N.: A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence 20(1), 18–36 (2004)
11. Peng, Y., Yao, J.: AdaOUBoost: Adaptive Over-sampling and Under-sampling to Boost the Concept Learning in Large Scale Imbalanced Data Sets. In: MIR 2010, Philadelphia, Pennsylvania, USA, pp. 111–118 (March 2010)
12. Drummond, C., Holte, R.C.: C4.5, Class Imbalance and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. In: Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets (2003)
13. Zheng, Z., Wu, X., Srihari, R.: Feature Selection for Text Categorization on Imbalanced Data. SIGKDD Explorations Newsletter 6(1), 80–89 (2004)
14. Chen, M.-C., Chen, L.-S., Hsu, C.-C., Zeng, W.-R.: An information granulation based data mining approach for classifying imbalanced data. Information Sciences 178, 3214–3227 (2008)
15. Wu, G., Chang, E.Y.: KBA: Kernel boundary alignment considering imbalanced data distribution. IEEE Transactions on Knowledge and Data Engineering 17(6), 786–795 (2005)
16. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 445–453. Morgan Kaufmann, San Francisco (1998)
17. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186. Morgan Kaufmann, San Francisco (1997)