IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2019, accepted May 26, 2019, date of publication June 17, 2019, date of current version July 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923524

A New Approach for Developing Segmentation Algorithms for Strongly Imbalanced Data

KAZUKI FUJIWARA, MAIKO SHIGENO, AND USHIO SUMITA
Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan
Corresponding author: Kazuki Fujiwara (s1720600@yahoo.co.jp)

ABSTRACT During the past two decades, the problem of how to develop efficient segmentation algorithms for dealing with strongly imbalanced data has been drawing much attention of researchers and practitioners in the field of data mining. A typical approach for this difficult problem is represented by a random under-sampling approach, where the cardinality of the majority set is reduced to that of the minority set through random sampling, thereby enabling one to utilize standard classifiers such as Logistic Regression, Support Vector Machine (SVM) and Random Forest. When the resulting segmentation algorithm is applied to a set of testing data with the original imbalanced-ness, however, its performance could be rather limited. So as to improve the performance, a bagged under-sampling (BUS) approach has been introduced where a random under-sampling is repeated M times, though the effect of BUS turns out to be still not quite satisfactory. The first purpose of this paper is to enhance the performance of BUS by developing a novel way where BUS is employed in a repetitive manner. While the performance improvement of this approach (R-BUS) over BUS is recognizable, it is still not sufficient enough from a practical point of view, especially when the dimension of underlying binary profile vectors is quite large. The second purpose of this paper is to establish a rank reduction (RR) approach for reducing this large dimension. The combined use of R-BUS with RR provides an excellent performance, as we will see through a real-world application of large magnitude.

INDEX TERMS Binary profile vectors, rank reduction approach, repetitive bagged under-sampling, strongly imbalanced data.

I. INTRODUCTION

Rare events may be defined as those events that happen with much less frequency than commonly occurring events. In real-world management problems, such minority events carry much more important and useful knowledge than common events, despite their rareness. Accordingly, from a data mining point of view, it becomes of vital importance to develop learning algorithms for predicting rare events with speed and accuracy. Since widely accepted machine learning algorithms assume balanced data in that the sizes of considered classes are approximately similar, the problem of dealing with rare events has been posing a tremendous difficulty.

One of early challenges to this problem can be traced back to early 1990's, where Anand et al. [1] proposed a novel algorithm for improving convergence rates of neural networks trained via backward-propagation based on deep analysis of imbalanced data. Since then, many methods have been developed. Following Krawczyk [13], we classify such methods into three approaches: (A) approach based on data pre-processing, where imbalanced data would be converted into a collection of small sets of balanced data through pre-processing so that standard segmentation algorithms can be applied; (B) approach for modifying or making ensemble of existing segmentation algorithms; and (C) hybrid approach where two approaches in (A) and (B) are combined.

Type (A) approaches can be further classified into two subcategories: (A-1) over-sampling where new objects for minority groups are generated according to underlying distributions, e.g. Chawla et al. [3]; and (A-2) under-sampling where examples are removed from majority groups, represented by Stefanowski [20]. One of the most prevalent approaches in (B) would be cost-sensitive, as in Zhou and Liu [27], where different penalties are assigned to different groups of examples in applying existing segmentation algorithms. Another approach in (B) is to apply a one-class segmentation algorithm that focuses on target groups so as to create a data description, see e.g. Japkowicz et al. [9]. Further modification would be needed to cope with more complex problems, as shown in Krawczyk et al. [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu.

Type (C) approaches combine previously mentioned approaches so as to extract their strong points and compensate their weaknesses, as represented by Woźniak et al. [23]. Merging data-level solutions with ensemble of segmentation methods is also popular so as to establish robust and efficient learning mechanism for imbalanced data, see e.g. Krawczyk et al. [11], Wozniak et al. [24] and Wang et al. [22].

The scope of applications involving strongly imbalanced data has been also expanded, including the problem of how to deal with software defects by Rodriguz et al. [19], natural disasters by Maalouf and Trafalis [15], fraudulent credit card transaction by Panigrahi et al. [18], telecommunications fraud by Olszewski [17], attention based neural networks for online advertising by Zhai et al. [26] and cancer gene expression by Yu et al. [25], to name only a few. As for understanding the scope of the imbalanced learning field, He and Garcia [7] provided a succinct summary of metrics and algorithm-level approaches, while Sun et al. [21] focused on the classification aspect of imbalanced learning. He and Ma [8] edited a book by collecting a variety of papers in the field, covering such important issues as sampling strategies, active learning and streaming data, among others. In García et al. [5], the topics of data preprocessing were discussed. Further references include Japkowicz and Stephen [10], Nguwi and Cho [16], Galar et al. [4], López et al. [14], and Branco et al. [2]. The reader is also referred to two excellent survey papers by Haixiang et al. [6] and Krawczyk [13] for further discussions of methodologies and applications associated with strongly imbalanced data.

The purpose of this paper is two-fold. The first purpose is to enhance the performance of a bagged under-sampling approach (BUS) by developing a repetitive BUS denoted by R-BUS. (See, e.g., the references in [4] for BUS.) While the performance improvement of R-BUS over BUS is recognizable, it is still not sufficient enough from a practical point of view, especially when the dimension of underlying binary profile vectors is quite large. The second purpose of this paper is to establish a rank reduction approach (RR) for reducing this large dimension. The combined use of R-BUS with RR provides an excellent performance, as demonstrated through a real-world application.

The structure of this paper is as follows. In Section 2, a succinct summary of the general structure of segmentation algorithms is provided for establishing notation to be employed throughout the paper. Section 3 describes a real-world problem to be studied by applying the new methods proposed in this paper. More specifically, considered are sessions which access the website of a housing equipment company. A problem of interest is to identify a session that would make a conversion based on only information available upon arrival at the website. This problem presents the difficulty associated with strongly imbalanced data. In Section 4, BUS is first introduced and then R-BUS is developed. The two methods are compared by applying them to the real-world problem discussed in Section 3, showing that R-BUS strongly enhances the learning quality of BUS. While R-BUS is an attempt to better deal with strongly imbalanced data, its effectiveness is still not satisfactory from a practical point of view, especially when the dimension of underlying binary profile vectors is quite large, as for the case of the real-world problem of Section 3. In order to overcome this difficulty, Section 5 is devoted to development of a rank reduction approach (RR) for profile vectors expressed as high dimensional binary vectors. In Section 6, the power of the combined use of R-BUS and RR is demonstrated through the real-world problem of Section 3. Finally, some concluding remarks are given in Section 7.

II. GENERAL STRUCTURE OF SEGMENTATION ALGORITHMS

We consider a data set $D = D_L \cup D_V \cup D_T$, where D_L, D_V and D_T are mutually exclusive and describe the set of learning data, the set of validating data and the set of testing data, respectively. Suppose that one has a flag function $flg : D \to \{0, 1\}$, that is, either $flg(x) = 0$ or $flg(x) = 1$ for each $x \in D$. The data set D can then be decomposed into $D = D_0 \cup D_1$, where $D_i = \{x : x \in D \text{ and } flg(x) = i\}$, for $i \in \{0, 1\}$. The decompositions $D_L = D_{L:0} \cup D_{L:1}$, $D_V = D_{V:0} \cup D_{V:1}$ and $D_T = D_{T:0} \cup D_{T:1}$ are defined similarly. Each $x \in D$ is characterized by its profile vector $\underline{y}(x) \in \Omega$. Corresponding to D_L, we define $\Omega_L = \{\underline{y}(x) : x \in D_L$, and Ω_V and Ω_T are defined in a similar manner.

Let ALG: $\Omega \to [0, 1]$ be an algorithm for estimating the probability that $P[x \in D_1]$. Given a threshold $z \in (0, 1)$, the associated segmentation algorithm SEG_z:$D \to \{0, 1\}$ can then be defined as

$$SEG_z(x) = \begin{cases} 1 & \text{if } ALG\left(\underline{y}(x)\right) \geq z \\ 0 & otherwise \end{cases} \tag{1}$$

The problem of interest is then how to find z^* which is optimal in some sense.

Given $z \in (0, 1)$, let $X_{L:ij} = \{x | SEG_z(x) = i, x \in D_{L:j}\}$ for $i, j \in \{0, 1\}$ and define $n_{L:ij} = \left|X_{L:ij}\right|$, where $|A|$ denotes the cardinality of a set A. A foundation for establishing 'optimality' can be provided by the confusion matrix given by

		flg		
		0	1	
SEG	0	$n_{L:00}$	$n_{L:01}$	$N_{L:SEG,0}$
	1	$n_{L:10}$	$n_{L:11}$	$N_{L:SEG,1}$
		$N_{L:flg,0}$	$N_{L:flg,1}$	N_L

(2)

where $N_{L:SEG,i}$ and $N_{L:flg,j}$ are the row sum and the column sum respectively, and $N_L = N_{L:SEG,0} + N_{L:SEG,1} = N_{L:flg,0} + N_{L:flg,1} = |D_L|$. The traditional measures for assessing the performance of $SEG_z(x)$ on D_L are:

$$Acc\left(D_L | SEG_z\right) = \frac{n_{L:00} + n_{L:11}}{N_L}; \tag{3.a}$$

$$Rec\left(D_L | SEG_z\right) = \frac{n_{L:11}}{N_{L:flg,1}}; \tag{3.b}$$

and

$$Pre\left(D_L | SEG_z\right) = \frac{n_{L:11}}{N_{L:SEG,1}}. \tag{3.c}$$

Here, $Acc\,(D_L|SEG_z)$ provides the overall accuracy of $SEG_z\,(x)$, while $Rec\,(D_L|SEG_z)$ denotes the portion of those x in $D_{L:1}$ which are picked up correctly by the segmentation algorithm, that is, $SEG_z\,(x) = 1$. $Pre\,(D_L|SEG_z)$ describes the portion of those x with $SEG_z\,(x) = 1$ which are judged correctly by the segmentation algorithm, that is, $SEG_z\,(x) = 1$ with x in $D_{L:1}$.

In general, as the segmentation criteria is tightened by increasing z, $Pre\,(D_L|SEG_z)$ increases while $Rec\,(D_L|SEG_z)$ decreases. In this paper, we optimize z by maximizing $Pre\,(D_V|SEG_z)$ subject to $Rec\,(D_V|SEG_z) \geq \alpha$ for a prespecified value $\alpha \in (0, 1)$. More specifically, given $z \in (0, 1)$, a segmentation algorithm SEG_z is firstly constructed on D_L. This segmentation algorithm is applied to the data set D_V so as to determine z^* in such a way that

$$z^* = \text{argmax}\{Pre\,(D_V \mid SEG_z) : Rec(D_V|SEG_z) \geq \alpha\}. \quad (4)$$

III. DATA DESCRIPTION AND BASIC FEATURES OF PAGE BLOCK ACCESSES BY SESSIONS AT THE WEBSITE

In this section, we describe a set of data to be employed for the study. The data set is provided by a housing equipment company, which has a website consisting of 33 blocks of pages, such as Top Page, Customer Service, Products, Reform and Showroom, among others. There are 90,121 pages spread over the 33 blocks. An access initiates a session, which begins with arrival at a landing page and is considered to end when it remains inactive for 30 minutes. The website has about 25,000 sessions per day. A session is said to make a conversion if it accesses the designated page in Customer Service, indicating that the session caller is likely to take a positive action soon for the businesses of the company. Accordingly, it is important to understand the characteristic of sessions making a conversion.

The data set provided by the company consists of 3,060,512 sessions over the period May through August 2018. An access to the website initiating a session may be made through key reference words provided by a search engine such as Google, Yahoo and the like. There are 3, 304 types of key reference words found in the data set, where one type is designated to indicate that the corresponding access is made without key reference words.

The conversion rate for this data set is extremely low at 0.3%. Equivalently, sessions without conversion and those with conversion are 3,050,817 and 9,695, respectively, constituting a set of strongly imbalanced data. Given a session x, let $\underline{v}\,(x)$ be the associated profile vector, where $\underline{v}\,(x)$ consists of the following binary vectors available upon access at the website. The domain of n dimensional binary vectors is denoted by B^n.

1) the day of a week on which the access is made: B^7

2) the hour of the day when the access is made: B^{24}

3) the type of device (PC, smart phone or tablet) from which the access is made: B^3

4) the block of the landing page on which the access is made:B^{33}

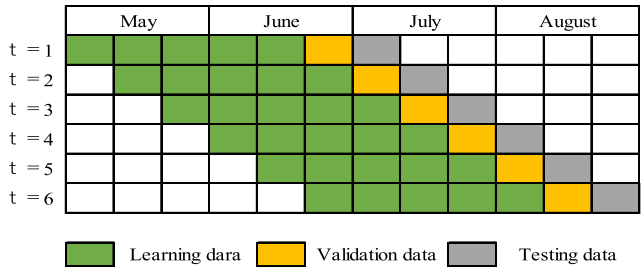

FIGURE 1. Construction of $D = D_L \cup D_V \cup D_T$ based on rolling horizon approach.

5) the title of the landing page: B^{90121}

6) the type of key reference words employed for the access:B^{3304}

By constructing $D = D_L \cup D_V \cup D_T$ from the above data set and using $\underline{v}\,(x)$, the performance of any probability estimation function ALG and the resulting segmentation algorithm SEG_{z^*} from (1) and (4) can be tested. For demonstrating the stability of such performances, a scheme based on a rolling horizon approach is adapted as shown in Fig.1. Here, each month is decomposed into three periods of almost equal length. Five consecutive periods constitute D_L, followed by one adjacent period for D_V and the period next to it for D_T. Such a set of periods would be repeated six times, thereby enabling one to test the stability of the underlying segmentation algorithms.

In what follows, we challenge the difficult problem of estimating the probability that a session would make a conversion upon access at the website based on the scheme in Fig.1, so as to test the effectiveness of a new approach to be developed for dealing with strongly imbalanced data.

IV. DEVELOPMENT OF REPETITIVE BAGGED RANDOM UNDER-SAMPLING APPROACH FOR STRONGLY IMBALANCED DATA

We refer to a dataset $D = D_0 \cup D_1$ as *strongly imbalanced* if the cardinality of D_0 is substantially larger than that of D_1, i.e. $|D_0| \gg |D_1|$. In this case, it is known that the development of the segmentation algorithm SEG_z in (1) for judging whether $x \in D$ belongs to $D_1$is extremely difficult. In order to overcome this difficulty, given $D = D_L \cup D_V \cup D_T$, a random under-sampling approach has been proposed. In the approach, D_L would be reconstructed as $\widehat{D}_L = D_{L:1} \cup \widehat{D}_{L:0}$ with $\widehat{D}_{L:0} \subset D_{L:0}$ and $|D_{L:1}| \approx |\widehat{D}_{L:0}|$ by randomly choosing elements from $D_{L:0}$ to form $\widehat{D}_{L:0}$. A probability estimation algorithm ALG is established over $\widehat{D}_L$ and the associated segmentation algorithm SEG_z is determined as in (1). Then, by applying SEG_z over D_V, the segmentation algorithm SEG_{z^*} can be finalized.

It is known, however, that the performance of SEG_{z^*} over D_T based on the random under-sampling approach above may not be necessarily satisfactory, since the sampling might miss important data in D_1 and/or there is a possibility to over-fitting to sampling data. The bagged random under-sampling approach (BUS) has been proposed to solving

TABLE 1. (a) Logit regression. (b) SVM. (c) Random forest. (d) Ensemble model.

(a)

t	1	2	3	4	5	6
Precision	0.035	0.018	0.031	0.029	0.008	0.010
Recall	0.564	0.576	0.506	0.595	0.587	0.568
F-value	0.065	0.036	0.058	0.056	0.016	0.019

(b)

t	1	2	3	4	5	6
Precision	0.026	0.012	0.032	0.019	0.017	0.030
Recall	0.562	0.526	0.487	0.607	0.622	0.563
F-value	0.050	0.023	0.060	0.037	0.033	0.057

(c)

t	1	2	3	4	5	6
Precision	0.040	0.013	0.032	0.020	0.019	0.019
Recall	0.538	0.525	0.538	0.584	0.602	0.591
F-value	0.074	0.026	0.060	0.039	0.037	0.036

(d)

t	1	2	3	4	5	6
Precision	0.042	0.016	0.034	0.025	0.025	0.021
Recall	0.572	0.524	0.495	0.584	0.613	0.534
F-value	0.078	0.032	0.064	0.048	0.048	0.041

TABLE 2. (a) Logit regression with bus. (b) SVM with bus. (c) Random forest with bus. (d) Ensemble model with bus.

(a)

t	1	2	3	4	5	6
Precision	0.116	0.080	0.114	0.140	0.064	0.097
Recall	0.570	0.554	0.487	0.607	0.581	0.541
F-value	0.187	0.128	0.181	0.199	0.110	0.150

(b)

t	1	2	3	4	5	6
Precision	0.120	0.084	0.122	0.136	0.067	0.103
Recall	0.551	0.532	0.497	0.613	0.622	0.569
F-value	0.192	0.138	0.180	0.210	0.114	0.155

(c)

t	1	2	3	4	5	6
Precision	0.132	0.087	0.127	0.145	0.069	0.109
Recall	0.567	0.509	0.533	0.590	0.621	0.568
F-value	0.194	0.139	0.183	0.215	0.120	0.168

(d)

t	1	2	3	4	5	6
Precision	0.135	0.090	0.131	0.148	0.074	0.113
Recall	0.578	0.509	0.491	0.567	0.590	0.556
F-value	0.190	0.144	0.195	0.215	0.127	0.167

such problems. Here, the random under-sampling is repeated M times by choosing $\widehat{D}_{L:0}(m)$, $m = 1, \cdots, M$ through re-sampling. For each $\widehat{D}_L(m) = D_{L:1} \cup \widehat{D}_{L:0}(m)$, one derives $ALG_m : \widehat{\Omega}_m \rightarrow [0, 1]$ where $\widehat{\Omega}_m = \{\underline{\nu}(x) : x \in \widehat{D}_L(m)\}$. Once ALG_m is established, it can be expanded to $ALG_m : \Omega_L \rightarrow [0, 1]$ by applying the same coefficients involved in ALG_m to $\underline{\nu}(x)$ for all $x \in D_L$. One can then unify ALG_m by taking the average, that is, for $x \in D_L$,

$$ALG_{AVE}\left(\underline{\nu}(x)\right) = \frac{1}{M}\sum_{m=1}^{M} ALG_m\left(\underline{\nu}(x)\right). \tag{5}$$

The segmentation algorithm can then be established in a unified manner based on ALG_{AVE} in (5) and (4), yielding $SEG_{AVE:z^*}$.

In Tables 1(a) through 1(d), the results of the segmentation algorithms associated with ALGs being Logit Regression, SVM (Support Vector Machine), Random Forest and Ensemble Model thereof are exhibited, when they are applied to the data set described in Section 3. We employ $\alpha = 0.6$ in (4) throughout our calculations, because it is appropriate for our data experimentally. Here, the results for Ensemble Model are obtained via the linear regression using the results of Logit Regression, SVM and Random Forest as independent variables. Tables 2(a) through 2(d) show similar results when BUS is employed. It can be seen that, because of the constraint in (4) applied to D_V, the overall results of Recall for D_T are more or less satisfactory. One also observes that BUS improves the results in every case. Among the three methods of Logit Regression, SVM and Random Forest, Random Forest seems to be superior to the other two. In general, Ensemble Model proves to perform best over the three other methods.

Although BUS provided a recognizable improvement, there seems to exist some room for further improvement, especially for Precision. In this paper, we propose a repetitive bagged random under-sampling approach, denoted by R-BUS hereafter. In R-BUS, one would start with BUS, and the resulting segmentation function is employed to reduce $D_{L:0}$ to the set of those in $D_{L:0}$ which are judged mistakenly as they belong to $D_{L:1}$. Then BUS is repeated with $D_{L:1}$ and the reduced $D_{L:0}$. This process continues until the further improvement cannot be expected. The idea behind R-BUS is to identify those in $D_{L:0}$ which are more difficult to be judged correctly. By concentrating on the segmentation of $D_{L:1}$ from those difficult elements in $D_{L:0}$ in a condensed manner, one may expect the improvement of the overall performance by R-BUS. More specifically, R-BUS can be described as follows.

R-Bus Algorithm

Input: $D = D_0 \cup D_1 = D_L \cup D_V \cup D_T$;

$D_L = D_{L:0} \cup D_{L:1} = (D_0 \cap D_L) \cup (D_1 \cap D_L)$

Output: $SEG_{CUM:\theta^*_{CUM:n}}(x)$ as the final segmentation algorithm, where CUM stands for the cumulative effect

[0] $n \leftarrow 1$; $D^1_L \leftarrow D_L$; $D^1_{L:1} \leftarrow D_{L:1}$; $D^1_{L:0} \leftarrow D_{L:0}$

[1] Apply BUS to $D^n_L = D^n_{L:1} \cup D^n_{L:0}$, yielding $ALG^n_{AVE}\left(\underline{\nu}(x)\right)$ in (5)

TABLE 3. (a) Logit regression with R-bus. (b) SVM with R-bus. (c) Random forest with R-bus. (d) Ensemble model with R-bus.

(a)

t	1	2	3	4	5	6
Precision	0.178	0.121	0.176	0.203	0.092	0.145
Recall	0.564	0.532	0.512	0.584	0.599	0.552
F-value	0.271	0.197	0.262	0.302	0.160	0.230

(b)

t	1	2	3	4	5	6
Precision	0.185	0.128	0.183	0.210	0.099	0.152
Recall	0.563	0.532	0.512	0.584	0.598	0.552
F-value	0.278	0.206	0.269	0.309	0.170	0.238

(c)

t	1	2	3	4	5	6
Precision	0.191	0.134	0.190	0.217	0.106	0.159
Recall	0.561	0.531	0.512	0.584	0.597	0.551
F-value	0.285	0.214	0.277	0.316	0.179	0.246

(d)

t	1	2	3	4	5	6
Precision	0.194	0.140	0.195	0.221	0.109	0.164
Recall	0.561	0.530	0.511	0.584	0.595	0.551
F-value	0.288	0.221	0.282	0.320	0.184	0.253

[2] Find $SEG^n_{AVE:z^*_n}$ by solving

$$z^*_n = \underset{0\le Z_n\le 1}{arg\,max} \left\{Pre\left(D_V \mid SEG^n_{AVE:z_n}(x)\right) : Rec\left(D_V \mid SEG^n_{AVE:z_n}(x)\right) \ge 1\right\}$$

[3] Set $ALG^n_{CUM}\left(\underline{v}(x)\right) = \{\prod_{l=1}^n SEG^l_{AVE:z^*_l}(x)\} \times ALG^n_{AVE}\left(\underline{v}(x)\right)$

[4] Define $SEG^n_{CUM:\theta}(x) = \begin{cases} 1 & if\ ALG^n_{CUM}\left(\underline{v}(x)\right) \ge \theta \\ 0 & else \end{cases}$

[5] Find $\theta^*_{CUM:n}$ by solving

$$\theta^*_{CUM:n} = \underset{0\le\theta\le 1}{arg\,max} \left\{Pre\left(D_V \mid SEG^n_{CUM:\theta}(x)\right) : Rec\left(D_V \mid SEG^n_{CUM:\theta}(x)\right) \ge \alpha\right\}$$

[6] Define

$$SEG^n_{CUM:\theta^*_{CUM:n}}(x) = \begin{cases} 1 & if\ ALG^n_{CUM}\left(\underline{v}(x)\right) \ge \theta^*_{CUM:n} \\ 0 & else \end{cases}$$

[7] Stop at $n\,(\ge 3)$ if $Pre\left(D_V \mid SEG^n_{CUM:\theta^*_{CUM:n}}\right) \ge Pre\left(D_V \mid SEG^{n-j}_{CUM:\theta^*_{CUM:n}}\right)$, for $j = 1, 2, 3$. In this case, use $SEG_{CUM:\theta^*_{CUM:n}}(x) = SEG^{n-j^*}_{CUM:\theta^*_{CUM:n}}(x)$ as the final segmentation algorithm, where $j^* = \text{argmin}_{1\le j\le 3}\{Pre\left(D_V \mid SEG^{n-j}_{CUM:\theta^*_{CUM:n}}\right)\}$. Otherwise, set $D^{n+1}_{L:0} = \{x \in D^n_{L:0} : SEG^n_{AVE:z^*_n}(x) = 1\}$, $D^{n+1}_L = D_{L:1} \cup D^{n+1}_{L:0}$; $n \leftarrow n+1$, and go to [1].

Corresponding to Tables 2(a) through 2(d), the results of the segmentation algorithms with R-BUS are exhibited in Tables 3(a) through 3(d). As before, because of the constraint in (4), the overall results for Recall are more or less comparable. However, R-BUS achieved around 50% improvement over BUS in the results of Precision and F-value in every comparison of (a) through (d). Among the four different methods involving R-BUS, Ensemble Model uniformly dominates the other three with improvement ranging from 1% to 18%. Despite such improvements by R-BUS, the performances for Precision and F-value are still rather limited, with the values in the range between 0.109 and 0.221 for Precision and that between 0.184 and 0.320 for F-value, as shown in TABLE 3(d). In the next section, we propose an additional device for better managing strongly imbalanced data. This additional device involves a new rank reduction technique for high dimensional binary vectors and improves the overall performance substantially when it is combined with R-BUS, as we will see.

TABLE 4. Example of rank reduction approach with $K = 3$.

	Category 1			Category 2		Category 3			$flg(x)$
	A	B	C	I	II	a	b	c	
$\underline{v}(x_1)$	1	0	0	1	0	1	0	0	1
$\underline{v}(x_2)$	1	0	0	1	0	0	1	0	0
$\underline{v}(x_3)$	1	0	0	1	0	0	0	1	1
$\underline{v}(x_4)$	1	0	0	1	0	1	0	0	0
$\underline{v}(x_5)$	1	0	0	0	1	0	1	0	1
$\underline{v}(x_6)$	1	0	0	1	0	0	0	1	0
$\underline{v}(x_7)$	0	1	0	0	1	1	0	0	1
$\underline{v}(x_8)$	0	1	0	0	1	0	1	0	0
$\underline{v}(x_9)$	1	0	0	1	0	0	0	1	1
$\underline{v}(x_{10})$	0	1	0	0	1	1	0	0	0
$\underline{v}(x_{11})$	0	1	0	1	0	0	1	0	1
$\underline{v}(x_{12})$	0	1	0	0	1	0	0	1	0
$\underline{v}(x_{13})$	0	0	1	1	0	1	0	0	0
$\underline{v}(x_{14})$	0	0	1	1	0	0	1	0	0
$\underline{v}(x_{15})$	0	0	1	0	1	0	1	0	0
$\underline{v}(x_{16})$	0	0	1	0	1	0	0	1	0

V. RANK REDUCTION APPROACH FOR PROFILE VECTORS EXPRESSED AS HIGH DIMENSIONAL BINARY VECTORS

In many cases, the segmentation algorithm is constructed based on a profile vector $\underline{v}(x)$ associated with $x \in D_L$, which may be defined over K different categories. When the domain of the k-th category is given by a set of $W(k)$ finite values for $k = 1, \ldots, K$, the segment of $\underline{v}(x)$ corresponding to the k-th category can be represented by a binary vector of length $W(k)$. More specifically, for $k = 1, \ldots, K$, we define

$$\underline{v}(x,k) = \left(v_1(x,k), .., v_{W(k)}(x,k)\right); \quad v_i(x,k) \in \{0,1\}; \sum_{i=1}^{W(i)} v_i(x,k) = 1, \tag{6}$$

so that the whole profile vector can be written as

$$\underline{V}(x) = \left(\underline{v}(x,1), .., \underline{v}(x,K)\right). \tag{7}$$

The dimension of $\underline{v}(x)$ is given by $\sum_{k=1}^K W(k)$, which may be quite large for many applications and the computational burden for developing $SEG{:}D \to \{0,1\}$ based on $\underline{v}(x), x \in D$ could be substantial. The purpose of this section is to establish a rank reduction approach for reducing the dimension of $\underline{v}(x)$ for achieving the computational efficiency but still without losing much information so that the accuracy of SEG can be assured.

TABLE 5. Result of rank reduction for example in Table 4.

$\mathcal{A}_m$	{1}	{2}	{3}	{1,2}	{1,3}	{2,3}	{1,2,3}	$flg(x)$
m	1	2	3	4	5	6	7	
$\tilde{V}(x_1)$	4/7	4/9	2/5	3/6	1/2	1/3	1/2	1
$\tilde{V}(x_2)$	4/7	4/9	2/6	3/6	1/2	1/3	0/1	0
$\tilde{V}(x_3)$	4/7	4/9	2/5	3/6	2/3	2/3	2/3	1
$\tilde{V}(x_4)$	4/7	4/9	2/5	3/6	1/2	1/3	1/2	0
$\tilde{V}(x_5)$	4/7	2/7	2/6	1/1	1/2	1/2	1/1	1
$\tilde{V}(x_6)$	4/7	4/9	2/5	3/6	2/3	2/3	2/3	0
$\tilde{V}(x_7)$	2/5	2/7	2/5	1/4	1/2	1/2	1/2	1
$\tilde{V}(x_8)$	2/5	2/7	2/6	1/4	1/2	1/2	0/2	0
$\tilde{V}(x_9)$	4/7	4/9	2/5	3/6	2/3	2/3	2/3	1
$\tilde{V}(x_{10})$	2/5	2/7	2/5	1/4	1/2	1/2	1/2	0
$\tilde{V}(x_{11})$	2/5	4/9	2/6	1/1	1/2	1/3	1/1	1
$\tilde{V}(x_{12})$	2/5	2/7	2/5	1/4	0/1	0/2	0/1	0
$\tilde{V}(x_{13})$	0/4	4/9	2/5	0/2	0/1	1/3	0/1	0
$\tilde{V}(x_{14})$	0/4	4/9	2/6	0/2	0/2	1/3	0/1	0
$\tilde{V}(x_{15})$	0/4	2/7	2/6	0/2	0/2	1/3	0/1	0
$\tilde{V}(x_{16})$	0/4	2/7	2/5	0/2	0/1	0/2	0/1	0

TABLE 6. (a) Logit Regression without R-bus with RR. (b) SVM without R-bus with RR. (c) Random forest without R-bus with RR. (d) Ensemble model without R-bus with RR.

(a)

t	1	2	3	4	5	6
Precision	0.100	0.116	0.081	0.074	0.084	0.101
Recall	0.576	0.569	0.674	0.545	0.547	0.603
F-value	0.170	0.193	0.145	0.131	0.145	0.173

(b)

t	1	2	3	4	5	6
Precision	0.088	0.119	0.168	0.121	0.194	0.108
Recall	0.642	0.547	0.609	0.551	0.575	0.608
F-value	0.155	0.196	0.264	0.199	0.290	0.183

(c)

t	1	2	3	4	5	6
Precision	0.076	0.187	0.131	0.195	0.066	0.128
Recall	0.560	0.539	0.638	0.544	0.575	0.656
F-value	0.133	0.278	0.217	0.286	0.118	0.214

(d)

t	1	2	3	4	5	6
Precision	0.078	0.190	0.133	0.200	0.069	0.130
Recall	0.583	0.524	0.574	0.516	0.540	0.621
F-value	0.137	0.278	0.216	0.288	0.123	0.215

Let $B(k)$ be a set of binary vectors of length $W(k)$ given by

$$B(k) = \{\underline{b}_1(k), \underline{b}_2(k), \ldots, \underline{b}_{W(k)}(k)\}, \tag{8}$$

where

$$\underline{b}_j(k) = \left(b_{j:1}(k), \ldots, b_{j:W(k)}(k)\right);$$
$$b_{j:r}(k) \in \{0,1\}; \sum_{r=1}^{W(k)} b_{j:r}(k) = 1. \tag{9}$$

Let $d_j(k) = \sum_{r=1}^{W(k)} 2^{W(k)-r} \times \delta\left\{b_{j:r}(k) = 1\right\}$, denoting the decimal interpretation of $\underline{b}_j(k)$. Here $\delta\{STATEMENT\} = 1$ if $STATEMENT$ is true and $\delta\{STATEMENT\} = 0$ otherwise.

TABLE 7. (a) Logit regression with R-bus and RR. (b) SVM with R-bus and RR. (c) Random forest with R-bus and RR. (d) Ensemble model with R-bus and RR.

(a)

t	1	2	3	4	5	6
Precision	0.611	0.803	0.678	0.814	0.712	0.666
Recall	0.618	0.559	0.636	0.540	0.564	0.622
F-value	0.615	0.659	0.656	0.649	0.629	0.643

(b)

t	1	2	3	4	5	6
Precision	0.618	0.810	0.685	0.821	0.718	0.673
Recall	0.618	0.558	0.634	0.540	0.564	0.621
F-value	0.618	0.661	0.659	0.651	0.632	0.646

(c)

t	1	2	3	4	5	6
Precision	0.625	0.816	0.692	0.828	0.725	0.680
Recall	0.615	0.556	0.633	0.538	0.564	0.619
F-value	0.620	0.661	0.661	0.652	0.634	0.648

(d)

t	1	2	3	4	5	6
Precision	0.630	0.818	0.695	0.830	0.729	0.685
Recall	0.614	0.552	0.629	0.537	0.563	0.614
F-value	0.622	0.659	0.660	0.652	0.635	0.648

We assume that $\underline{b}_j(k)$ in (8) are ordered in such a way that $d_j(k) \geq d_{j+1}(k), j = 1, \ldots, W(k-1)$.

Let $\beta(Y)$ be the power set of a set Y excluding the empty set, that is, the set of all the subsets of Y excluding the empty set, and define $X(K) = \{\underline{a} = (a_1, \ldots, a_K) : a_j \in \{1, \ldots, W(k)\}, j = 1, \ldots, K\}$. We note that the cardinality of $X(K)$ is given by $|X(K)| = \prod_{k=1}^{K} W(k)$. For $\mathcal{A} \in \beta(\{1, \ldots, K\})$, let

$$D_L\left(\mathcal{A}, \underline{a}\right) = \left\{x : \underline{v}(x,k) = \underline{b}_{a_k}(k) \in B(k) \text{ for } k \in \mathcal{A}\right\}. \tag{10}$$

For notational convenience, we write $\beta(\{1, \ldots, K\}) = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{2^K-1}\}$. We are now in a position to establish a way of reducing the cardinality of $\underline{v}(x)$. For any $A_m \in \mathcal{B}(\{1, .., K\})$ and $\underline{a} \in X(K)$, we define

$$\underline{\tilde{V}}\left(\underline{a}\right) = \left\{\tilde{v}_1\left(\underline{a}\right), \ldots, \tilde{v}_w\left(\underline{a}\right)\right\}; \mathrm{W} = 2^K - 1, \tag{11}$$

where

$$\tilde{v}_m\left(\underline{a}\right) = \frac{|\{x : x \in D_L\left(\mathcal{A}_m, \underline{a}\right) \wedge flg(x) = 1\}|}{|D_L\left(\mathcal{A}_m, \underline{a}\right)|},$$
$$m = 1, \ldots, 2^K - 1. \tag{12}$$

The dimension of $\underline{v}(x)$ is given by $\sum_{k=1}^{K} W(k)$, which can now be reduced to that of $\underline{\tilde{V}}\left(\underline{a}\right)$, i.e. $2^K - 1$. This reduction could be substantial when individual $W(k)$'s is quite large while K is relatively small. For the data set

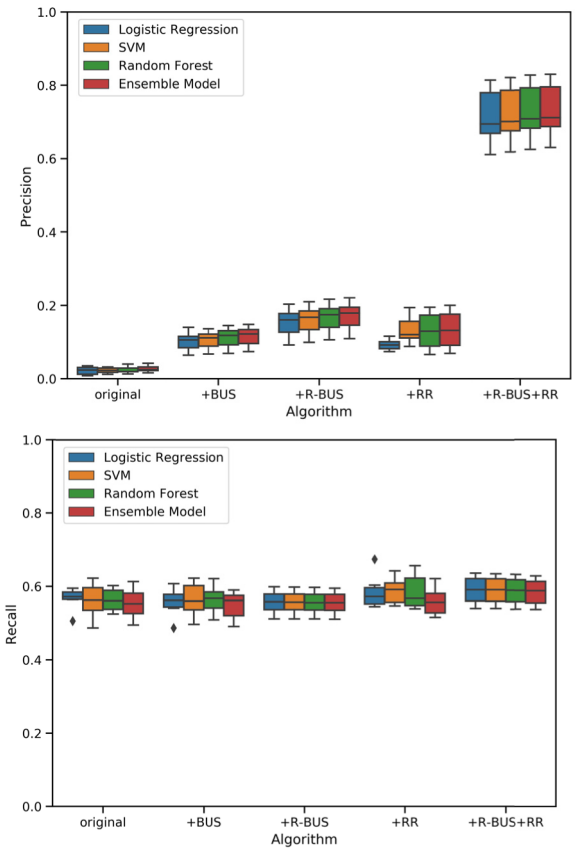

FIGURE 2. Comparing results of precision and recall of approaches we discussed.

described in Section 3, for example, one has $\sum_{k=1}^{6} W(k) = 7 + 24 + 3 + 33 + 90{,}121 + 3{,}304 = 93{,}492$, which is reduced to $2^6 - 1 = 63$. This reduction process is illustrated in Tables 4 and 5 with $K = 3$, where $\tilde{\underline{V}}(x)$ means the corresponding $\tilde{V}(\underline{a})$ for $\underline{a}$ constructed from x.

VI. APPLICATION OF THE NEW ALGORITHM FOR IDENTIFYING CONVERSION SESSIONS BASED ON ONLY INFORMATION AVAILABLE UPON ARRIVAL AT THE WEBSITE FOR A HOUSING EQUIPMENT COMPANY

In Tables 6(a) though 6(d), exhibited are the results of the four segmentation algorithms associated with Logit Regression, SVM, Random Forest and Ensemble Model, combined with RR (Rank Reduction) but without R-BUS, when they are applied to the data set described in Section 3. By comparing them against the corresponding counterparts in Tables 3(a) though 3(d), one observes that R-BUS in general outperforms RR for both Precision and Recall. Since RR is designed to reduce the dimension of binary profile vectors with intention of retaining characteristic information associated with the flags 0 and 1, RR alone cannot overcome the difficulty arising from imbalanced data. R-BUS is developed to deal with imbalanced data, and accordingly it could perform relatively better than RR. However, when the dimension of binary profile vectors is extremely large, the effect of R-BUS for handling imbalanced data could be still limited, as demonstrated in Tables 3(a) though 3(d).

When RR is employed together with R-BUS, one may expect that the two methods could exploit their strengths and compensate their weaknesses. This observation turns out to be true, as can be seen in Tables 7(a) though 7(d), demonstrating substantial improvement over the case of the single use of either R-BUS or RR for both Precision and F-value. One sees that the improvements for Precision between Tables 3(a) though 3(d) and Tables 7(a) though 7(d) range from 325% to 772%, while those for F-value lie between 206% and 394%.

Finally, we compare our whole approaches we discussed in Fig. 2, where we can see method with R-BUS and RR is improved drastically for precision, although values of recall do not so change.

VII. CONCLUSION

In this paper, a repetitive bagged under-sampling approach, denoted by R-BUS, is first developed for enhancing the learning capability through strongly imbalanced data. While the performance improvement of R-BUS over the existing bagged under-sampling approach is recognizable, it is still not sufficient enough from a practical point of view, especially when the dimension of underlying binary profile vectors is quite large. In order to overcome this difficulty, a rank reduction approach (RR) is also developed for profile vectors expressed as high dimensional binary vectors. The thrust of this paper can be found in the combined use of R-BUS and RR, demonstrating its power by applying them to a real-world problem.

More specifically, a housing equipment company is considered, which has a website consisting of 33 blocks of pages, such as Top Page, Customer Service, Products, Reform and Showroom, among others. There are 90,121 pages spread over the 33 blocks. An access initiates a session, which begins with arrival at the landing page and is considered to end when it remains inactive for 30 minutes. The website has about 25,000 sessions per day. A session is said to make a conversion if it accesses the designated page in Customer Service, indicating that the session caller is likely to take a positive action soon for the businesses of the company. Accordingly, it is important to identify sessions making a conversion upon their arrival at the website. Since a typical conversion rate is 0.3%, this problem involves the difficulty associated with strongly imbalanced data. The novel approach proposed in this paper revealed its power, achieving Recall around 0.6, and Precision and F-value over 0.6 despite the extraordinary rareness at 0.3%.

As for the future challenge, the combined use of R-BUS and RR may be tested involving segmentation algorithms other than Logistic Regression, SVM and Random Forest employed in this paper. When characteristic vectors are defined over different categories, they may be converted into binary profile vectors. Then the combined use of R-BUS and RR may be employed together with LSTM (Long Short Term Memory) for dealing with extreme rareness involving

sequential data. Similar approach may be devised for identifying rare events associated with image analysis. Such studies are in progress and will be reported in due course.

REFERENCES

[1] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Trans. Neural Netw.*, vol. 4, no. 6, pp. 962–969, Nov. 1993.

[2] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modelling under imbalanced distributions," *CoRR*, May 2015. [Online]. Available: https://arxiv.org/abs/1505.01658

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[4] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[5] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining* (Intelligent Systems Reference Library), vol. 72. Berlin, Germany: Springer, 2015.

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[8] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed. New York, NY, USA: Wiley, 2013.

[9] N. Japkowicz, C. Myers, and M. Gluck, "A novelty detection approach to classification," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, vol. 1, San Francisco, CA, USA: Morgan Kaufmann, 1995, pp. 518–523.

[10] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[11] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Appl. Soft Comput.*, vol. 14, pp. 554–562, Jan. 2014.

[12] B. Krawczyk, M. Woźniak, and F. Herrera, "On the usefulness of one-class classifier ensembles for decomposition of multi-class problems," *Pattern Recognit.*, vol. 48, no. 12, pp. 3969–3982, 2015.

[13] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[14] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[15] M. Maalouf and T. B. Trafalis, "Robust weighted kernel logistic regression in imbalanced and rare events data," *Comput. Statist. Data Anal.*, vol. 55, no. 1, pp. 168–183, 2011.

[16] Y.-Y. Nguwi and S.-Y. Cho, "An unsupervised self-organizing learning with support vector ranking for imbalanced datasets," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8303–8312, 2010.

[17] D. Olszewski, "A probabilistic approach to fraud detection in telecommunications," *Knowl.-Based Syst.*, vol. 26, pp. 246–258, Feb. 2012.

[18] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning," *Inf. Fusion*, vol. 10, no. 4, pp. 354–363, 2009.

[19] D. Rodriguez, I. Herraiz, R. Harrison, J. Dolado, and J. C. Riquelme, "Preliminary comparison of techniques for dealing with imbalance in software defect prediction," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng.*, 2014, Art. no. 43.

[20] J. Stefanowski, "Dealing with data difficulty factors while learning from imbalanced data," in *Challenges in Computational Statistics and Data Mining*. Springer, 2016, pp. 333–363.

[21] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[22] S. Wang, Z. Li, W. Chao, and Q. Cao, "Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Brisbane, QLD, Australia, Jun. 2012, pp. 1–8.

[23] M. Woźniak, *Hybrid Classifiers: Methods of Data, Knowledge, and Classifier Combination* (Studies in Computational Intelligence), vol. 519. Berlin, Germany: Springer, 2014.

[24] M. Wozniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Inf. Fusion*, vol. 16, pp. 3–17, Mar. 2014.

[25] H. Yu, J. Ni, Y. Dan, and S. Xu, "Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets," *Tsinghua Sci. Technol.*, vol. 17, no. 2, pp. 666–673, Dec. 2012.

[26] J. Zhai, S. Zhang, and C. Wang, "The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 3, pp. 1009–1017, 2017.

[27] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Comput. Intell.*, vol. 26, no. 3, pp. 232–257, 2010.

KAZUKI FUJIWARA received the master's degree in service engineering from the Graduate School of Systems and Information Engineering, University of Tsukuba, in 2019. His research interests include the areas related to big data analytics with emphasis on production.

MAIKO SHIGENO received the Ph.D. degree in science from the Tokyo Institute of Technology, in 1996. She is a Professor with the University of Tsukuba. Her specialized research areas are operations research and mathematical optimization. Especially, she has research interests in combinatorial optimization and network optimization.

USHIO SUMITA received the first Ph.D. degree from the University of Rochester, USA, in 1981, and the second Ph.D. degree from the Tokyo Institute of Technology, Japan, in 1987. He is a Professor Emeritus with the University of Tsukuba. He is also a Technical Advisor for some company. His research interests focus on both theoretical and functional areas, including applied probability, stochastic processes, financial engineering, marketing, production management, information technology, and big data analytics, among others. He has published more than 170 papers in leading archive journals in such areas.