Journal of Universal Computer Science, vol. 24, no. 6 (2018), 665-681
submitted: 22/5/17, accepted: 15/12/17, appeared: 28/6/18 © J.UCS

A Hybrid Machine Learning Scheme to Analyze the Risk Factors of Breast Cancer Outcome in Patients with Diabetes Mellitus

Linglong Ye
(School of Public Health, Xiamen University, Xiamen, Fujian Province, China
School of Economics, Xiamen University, Xiamen, Fujian Province, China
Graduate Institute of Business Administration, College of Management
Fu Jen Catholic University, New Taipei City, Taiwan
leyloria@gmail.com)

Tian-Shyug Lee*
(Graduate Institute of Business Administration, College of Management
Fu Jen Catholic University, New Taipei City, Taiwan
036665@mail.fju.edu.tw
*corresponding author)

Robert Chi
(Department of Information Systems, California State University
Long Beach, United States
rchi@csulb.edu)

Abstract: Along with the worldwide trend of rapidly aging populations, diabetes mellitus and its comprehensive complications have become major public health issues. Considerable evidence suggests patients with diabetes mellitus have a higher risk of breast cancer. However, the relationships between the complications of diabetes mellitus and occurrence of breast cancer have not been well characterized. Despite the higher risk of breast cancer among patients with diabetes mellitus, patients with breast cancer constitute only a relatively small proportion of the diabetes mellitus data, leading to an imbalanced data set. This study proposes a hybrid machine learning scheme to cope with imbalanced data in the analysis of risk factors of breast cancer in patients with diabetes mellitus. The scheme combines the undersampling based on the clustering algorithm, the k-means algorithm, and the extreme gradient boosting algorithm. The results identify that occlusion stroke, diabetes with peripheral circulatory disorders, peripheral angiopathy in diseases classified elsewhere, and other forms of chronic ischemic heart disease are risk factors. This study provides an application of advanced methods in health care and shows the epidemiologic and informatics value of the proposed hybrid machine learning scheme.

Key Words: Machine Learning, Risk factor, Diabetes Mellitus, Breast Cancer, Complication

Category: I.2, J.3, L.2

1 Introduction

Population aging becomgs a global phenomenon, as the number of people aged 60 years and older is expected to more than double between 2015 and 2050 [PDDESAUN 2015]. The aging process of Taiwan is more advanced than most regions in the world. Between 2015 and 2050, the proportion of Taiwan's population aged over 65 is predicted to increase from 12.5% to 38.9%—a three-fold increase over just 35 years—and it is expected to exceed the population aged under 14 years old in 2017 [NDC 2016]. This trend is projected to cause a dramatic increase in the prevalence of chronic diseases [PDDESAUN 2015, Hsu and Hsu 2016, Nie et al. 2008]. As a common chronic disease among the elderly that has increasing prevalence worldwide [Hou et al. 2013], diabetes mellitus and its comprehensive complications are major public health issues, which cause a heavy burden to health care systems [Hsu et al. 2011]. In 2015 in Taiwan,, diabetes mellitus was the fifth leading cause of death, and its treatments accounted for 4.4% of health care expenses [MHW 2017].

Considerable evidence suggests that patients with diabetes mellitus might have a higher risk of cancer, including cancer of the pancreas, endometrium, liver, bladder, colorectum, and breasts [Giovannucci et al. 2010]. In Taiwan, cancer was the leading cause of death in 2015 for the 33rd consecutive year, and breast cancer was the fourth highest cause of death for females with cancer [MHW 2017]. Several meta-analyses, which have used traditional statistical methods such as the statistical hypothesis testing and the generalized linear models, have shown that patients with diabetes mellitus might have an approximately 20% higher risk of breast cancer than patients without diabetes mellitus [Larsson et al. 2007, Liao et al. 2011, Boyle et al. 2012]. Previous studies have also reported the association between diabetes mellitus and breast cancer risk among different ethnic populations in differing geographic locations [Maskarinec et al. 2017, Larsson et al. 2007, Boyle et al. 2012]. Epidemiological studies suggest that drug exposure might create a link between diabetes mellitus and breast cancer [Redaniel et al. 2012, Sieri et al. 2012, Bosco et al. 2010, Tseng 2015]. Several mechanisms relating to obesity might also explain part of the increased incidence of breast cancer for patients with diabetes mellitus [Jiralerspong et al. 2013, Tait et al. 2014]. Nevertheless, as the association between the comprehensive complications of diabetes mellitus and the occurrence of breast cancer has not been well characterized, powerful machine learning tools are needed to analyze the differences in complications among diabetes mellitus patients with and without breast cancer.

Although the risk of breast cancer is higher for patients with diabetes mellitus, the number of patients with breast cancer constitutes only a relatively small percentage of the diabetes mellitus data, leading to an imbalanced data set. Imbalance is common for health care data, and it significantly impacts the performance of most standard machine learning algorithms [Liu et al. 2009, He and Edwardo 2009, Bach et al. 2017]. Balanced class distributions or equal misclassification costs are the bases of most standard machine learning algorithms [He and Edwardo 2009]. Hence, adopting these algorithms would fail to appropriately obtain the distribution characteristics of imbalanced health care data, and result in poor classification [Bach et al. 2017, He and Edwardo 2009]. This study therefore proposes a hybrid machine learning scheme to cope with the challenge of imbalance in data for analyzing the risk factors of breast cancer occurrence in patients with diabetes mellitus.

The remainder of this paper is structured as follows. Section 2 outlines the machine learning methods, including undersampling based on a clustering (SBC) algorithm, the k-means algorithm, and the extreme gradient boosting (XGBoost) algorithm. Section 3 presents the hybrid machine learning scheme, including the data source, study population, definitions of variables, and the process of analysis. Section 4 discusses the results of the scheme, and Section 5 summarizes the main findings of the study and suggests future research directions.

2 Research methods

2.1 SBC

SBC is a random undersampling method based on clustering to improve the prediction of the minority class in an imbalanced data set [Yen and Lee 2009]. Let N be the number of elements in the imbalanced data set with majority class elements (MA) and minority class elements (MI), let $Size_{MA}$ be the number of elements in MA, and $Size_{MI}$ be the number of elements in MI. In an imbalanced data set, $Size_{MA}$ is considerably larger than $Size_{MI}$. In SBC, the imbalanced data set is first grouped into k clusters $C = \{c_i\}\,(i = 1, \cdots, k)$, where $Size^i_{MA}$ and $Size^i_{MI}$ are, respectively, the number of MA and MI elements in the cluster c_i. Let the ratio of $Size_{MA}$ to $Size_{MI}$ in the undersampling dataset be $m : 1$ $(m \geq 1)$. Then the number of selected MA elements in the cluster c_i is defined as

$$SSize^i_{MA} = (m \times Size_{MI}) \times \frac{Size^i_{MA}/Size^i_{MI}}{\sum_{i=1}^{k} Size^i_{MA}/Size^i_{MI}}$$

where $m \times Size_{MI}$ is the total number of selected MA elements, $Size^i_{MA}/Size^i_{MI}$ is the ratio of the number of MA elements to the number of MI elements in cluster c_i, and $\sum_{i=1}^{k} Size^i_{MA}/Size^i_{MI}$ is the total ratio of the number of MA elements to the number of MI elements in all clusters. After obtaining the number of MA elements selected in cluster c_i, the MA elements are randomly picked from each cluster. Finally, the selected MA elements and all MI elements are combined to obtain resampled data sets.

2.2 K-means

The K-means algorithm has been widely used since it was first published in 1955, despite a great number of subsequent clustering algorithms being published [Aljawarneh et al. 2016, Jain 2010, Tonejc 2016]. The long-term use of the k-means algorithm shows not only the difficulty in designing clustering algorithms with general purpose, but also the stability and comprehensive applicability of the k-means algorithm.

In the k-means clustering algorithm, the data set is divided into k clusters such that the elements in a given cluster are closer to that cluster's center than to any other cluster's center, with distance typically measured by the Euclidean metric. For example, a data set of n elements with m features, such as $X = \{x_i\}\,(i = 1, \cdots, n)$ is clustered into a set of k clusters, such as $C = \{c_j\}\,(j = 1, \cdots, k)$, with the aim of finding a partition that minimizes the squared error between the empirical mean of a cluster and the elements in the cluster. Let μ_j be the mean of a given cluster c_j, then the squared error between μ_j and the points in the given cluster c_j is defined as

$$f\left(c_j\right) = \sum_{x_i \epsilon c_j} \left\|x_i - \mu_j\right\|^2,$$

and the minimization of the sum of the squared error over all k clusters is

$$f\left(c_j\right) = \min_{C} \sum_{j}^{k} \sum_{x_i \epsilon c_j} \left\|x_i - \mu_j\right\|^2.$$

The main process of the k-means algorithm is as follows [Aljawarneh et al. 2016, Jain 2010, Tonejc 2016]:

(1) Set an initial partition with k clusters for all points.

(2) Generate a new partition by assigning each element to its closest cluster center.

(3) Measure the center of the new cluster.

(4) Repeat steps (2) and (3) until cluster membership stabilizes.

The silhouette function provides a measure of compactness and separation of clusters that is widely used to choose the number of clusters k [Rousseeuw 1987, Benmouiza and Ali 2013, Abualhaj et al. 2017]. For the element x_i in a cluster c_j, the silhouette function is defined as

$$s\left(i\right) = \frac{b\left(i\right) - a\left(i\right)}{max\left\{a\left(i\right), b\left(i\right)\right\}}$$

where $a\left(i\right)$ is the average dissimilarity of x_i compared with other elements within the same cluster c_j, and $b\left(i\right)$ is the minimum average similarity of x_i compared with elements in other clusters, expressed as $\min d\left(x_i, c_m\right)\,(m = 1, \cdots, k; m \neq j)$.

The result can be interpreted as how similar an element is to other elements in the same cluster when compared to elements in other clusters.

The silhouette width $s(i)$ of the element x_i in a cluster c_j lies in the range $[-1, 1]$. If $s(i)$ is close to -1, then $a(i) \gg b(i)$, indicating that element x_i is badly matched for the given cluster and might be more appropriately matched to the neighboring cluster. An $s(i)$ near zero indicates that the elements may be on the border of two clusters and could be reasonably assigned to either cluster. If $s(i)$ is close to 1, then $a(i) \ll b(i)$, indicating that element x_i is appropriately clustered. Therefore, good clustering leads to a high mean silhouette width $s(i)$. A graphical representation of the mean silhouette value can be used to test various sets of clusters.

2.3 XGBoost

XGBoost is a new scalable end-to-end tree boosting system for classification that is based on the gradient boosting model (GBM) proposed by Friedman [Friedman 2001, Chen and Guestrin 2016]. Compared with traditional tree algorithms, XGBoost speeds up tree construction and includes a new distributed algorithm for tree searching [Torlay et al. 2017].

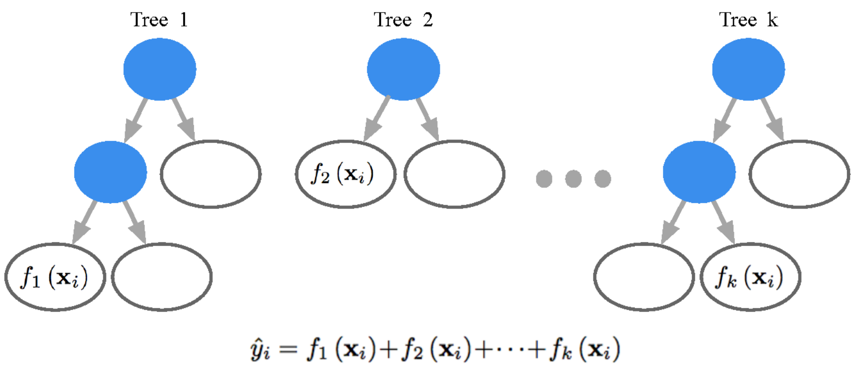

Figure 1: Tree ensemble model

For a given data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ $(|\mathcal{D}| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R})$ with n examples and m features, K additive functions are adopted in a tree ensemble model (shown in Figure 1) to predict the output, with the prediction defined as

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}$$

where $\mathcal{F} = \left\{f(\mathbf{x}) = w_{q(\mathbf{x})}\right\}\left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T\right)$ is the space of regression trees, q is the structure of each tree that maps an example to the corresponding leaf index, T is the number of leaves in the tree, each f_k corresponds to an independent tree structure q and leaf weight w, and w_i is defined as the score on the i-th leaf.

The objective function of XGBoost refers to training loss and regularization, and can be expressed as follows

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$where\, \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 .$$

Here, l is a differentiable convex loss function used to estimate the differences between the prediction $\hat{y}_i$ and the target y_i, Ω is a regularization function that penalizes the complexity of the model, and γ and λ are the penalty parameters in the regularization.

To quickly optimize the objective in the general setting, XGBoost uses the second-order approximation as follows

$$\mathcal{L}^{(t)} \simeq \sum_i \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(\mathbf{x}_i) \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_k)$$

where $\hat{y}_i^{(t)}$ is the prediction of the i-th instance at the t-th iteration, $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$.

XGBoost has been widely used in many machine learning challenges and has achieved state-of-the-art results [Chen and Guestrin 2016]. Compared with traditional machine learning methods, the advantages of XGBoost include regularization, parallel processing, high flexibility, handling missing values, and built-in cross-validation. The regularization term serves to control the complexity of the model by smoothing the final learnt weight, which also helps to avoid overfitting. The block structure in XGBoost provides the basis for parallel processing, leading to faster learning. The high flexibility of XGBoost is due to a number of adjustable parameters that can be customized for different problems. XGBoost adopts the algorithm of sparsity-aware split finding to learn the best path for missing values. A cross-validation at each iteration of the boosting process is provided in XGBoost, and it is easy to obtain the exact optimum number of boosting iterations in a single run [Jain 2016]. However, the XGBoost is memory intensive and traditionally slower than its new challenger, lightGBM.

3 Hybrid machine learning scheme

This study proposes a hybrid machine learning scheme, represented in Figure 2, to analyze the risk factors of breast cancer occurrence in patients with diabetes mellitus. The scheme was developed as follows:

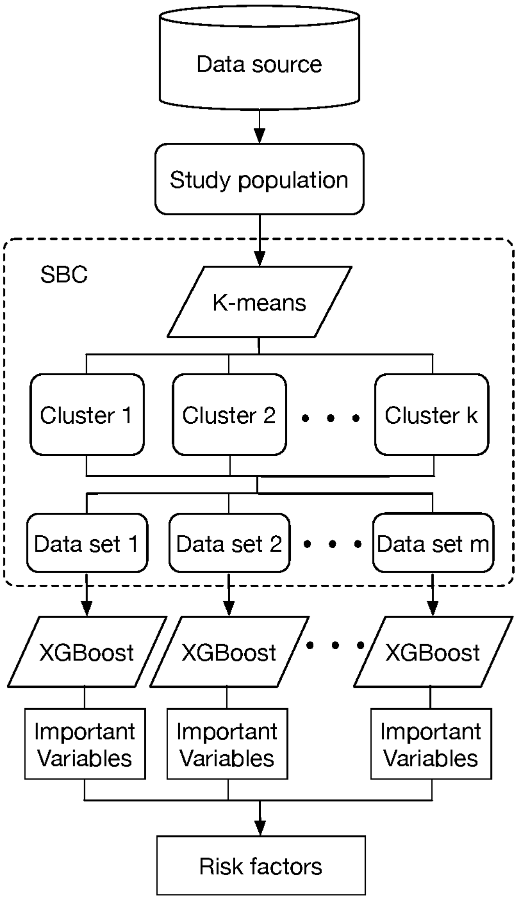

Figure 2: Hybrid machine learning scheme

(1) Given the definitions of diabetes mellitus, its complications and breast cancer, the study population was obtained from the data source.

(2) The SBC algorithm was adopted to address the problem of imbalance in the data set. This study adopted the k-means algorithm to cluster the imbalanced data and determined the number k of clusters in k-means clustering using the silhouette function. We set $m = 1, 2, 3$ to calculate the numbers of selected MA elements in each cluster.

(3) The XGBoost algorithm was applied to analyze important predictive variables for the occurrence of breast cancer based on the results of SBC. This study sampled five times to build five XGBoost models for each ratio of $m : 1$,

and select important variables according to the mean gain.

(4) The important predictive variables for each ratio of $m : 1$ were integrated to obtain the risk factors of breast cancer occurrence in patients with diabetes mellitus.

This study used Taiwan's nationwide population-based health insurance data [Cheng 2015]. Patients with diabetes mellitus were identified according to the definition of diabetes mellitus proposed by the National Health Research Institutes, Taiwan [NHRI 2014a]. Patients with breast cancer were identified according to the registry for catastrophic illness patients [NHRI 2014b]. This study included the female patients with diabetes mellitus before January 1, 2010, and excluded those with breast cancer before January 1, 2010. The patients that died from any cause before January 1, 2010 were excluded. Confirmation of the death and the date of death were obtained from the withdrawal records of the patient from the National Health Insurance program [Wu et al. 2012]. In addition, since previous studies showed that type 1 diabetes mellitus might not be associated with higher risk of breast cancer [Wolf et al. 2005], this study excluded patients with type 1 diabetes mellitus. The exclusions led to a study population of 30,025 patients. The breast cancer occurrence was defined since January 1, 2010.

This study adopted the comprehensive complications of diabetes mellitus as the predictive variables. Previous studies have pointed out the association of breast cancer with several complications of diabetes mellitus, such as cardiovascular diseases, peripheral circulatory disorders, and neurological manifestations. Boyle et al. noted that menopause and post-menopausal obesity increase the risk of breast cancer. Ramezani et al. showed that menopause incurs cardiometabolic risk. Peripheral circulatory disorders are common diseases of the circulatory system and frequent complications of diabetes. Xie et al. found that microvessel density is correlated to hMAM mRNA as valuable markers for the micrometastases of breast cancer. Pereira et al. discussed the impact of neurological manifestations on breast cancer. Hence, this study adopted cardiovascular diseases, peripheral circulatory disorders, and neurological manifestations as the predictive variables. The definitions of complications (35 diseases) were based on a patient having at least three related outpatient records or at least one related inpatient record between January 1, 2005 and December 31, 2009. Table 1 lists the relevant complications.

Complication	Disease	Variable definition
Cardiovascular disease	Transient cerebral ischemia	CVD01
Cardiovascular disease	Atherosclerosis	CVD02
Cardiovascular disease	Other acute and subacute forms of ischaemic heart disease	CVD03
Cardiovascular disease	Angina pectoris	CVD04
Cardiovascular disease	Other forms of chronic ischemic heart disease	CVD05
Cardiovascular disease	Cardiovascular disease, unspecified	CVD06
Cardiovascular disease	Hemorrhagic stroke	CVD07
Cardiovascular disease	Occlusion stroke	CVD08
Cardiovascular disease	Acute cerebrovascular disease	CVD09
Cardiovascular disease	Actue myocardial infarction	CVD10
Cardiovascular disease	Paroxysmal supraventricular tachycardia	CVD11
Cardiovascular disease	Atrial fibrillation	CVD12
Cardiovascular disease	Old myocardial infarction	CVD13
Cardiovascular disease	Heart failure	CVD14
Cardiovascular disease	Atherosclerosis of the extremities with ulceration	CVD15
Cardiovascular disease	Aortic aneurysm and dissection	CVD16
Neurological manifestations	Hereditary and idiopathic peripheral neuropathy	NM01
Neurological manifestations	Myasthenic syndromes in diseases classified elsewhere	NM02
Neurological manifestations	Injury to other cranial nerve(s)	NM03
Neurological manifestations	Mononeuritis of upper limb and mononeuritis multiplex	NM04
Neurological manifestations	Arthropathy associated with neurological disorders	NM05
Neurological manifestations	Polyneuropathy in diabetes	NM06
Neurological manifestations	Neurogenic bladder	NM07
Neurological manifestations	Cardiovascular autonomic neuropathy	NM08
Neurological manifestations	Functional diarrhea	NM09
Neurological manifestations	Orthostatic hypotension	NM10
Peripheral circulatory disorders	Diabetes with peripheral circulatory disorders	PCD01
Peripheral circulatory disorders	Of artery of lower extremity aneurysm	PCD02
Peripheral circulatory disorders	Peripheral angiopathy in diseases classified elsewhere	PCD03
Peripheral circulatory disorders	Open wound of foot except toe(s) alone	PCD04
Peripheral circulatory disorders	Peripheral vascular disease, unspecified	PCD05
Peripheral circulatory disorders	Lower extremity	PCD06
Peripheral circulatory disorders	Gangrene	PCD07
Peripheral circulatory disorders	Gas gangrene	PCD08
Peripheral circulatory disorders	Ulcer of lower limbs, except pressure ulcer	PCD09

Table 1: Definitions of complications of diabetes mellitus.

The ratios of $m : 1$ in the process of SBC were determined in subjective manner. Hence, this study also adopted another ratio of $m : 1 (m = 5)$ for sensitivity analysis.

4 Results

The MI elements of the study population in this study were 228 patients with breast cancer between 2010 and 2013, comprising 7.59% of female patients with diabetes mellitus. In comparison, the MA elements of the student population were 29,797 patients without breast cancer. The ratio of MA to MI elements was found to be very high in this study, indicating an imbalance. The prevalence of breast cancer in the general female population was 4.34% during the same period in Taiwan, which was lower than that of patients with diabetes mellitus. This result aligns with the higher risk of breast cancer for patients with diabetes mellitus suggested by previous studies [Larsson et al. 2007, Liao et al. 2011, Boyle et al. 2012]. Further, diabetes mellitus patients with complications had an 8.18% risk of breast cancer.

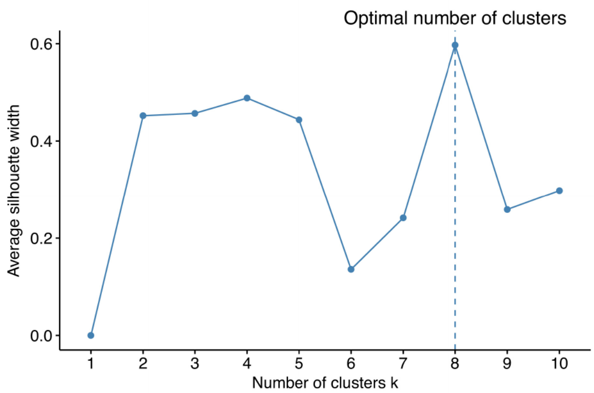

Figure 3: Optimal number of clusters with silhouette method for k-means clustering

In the process of SBC, the results of the silhouette method for the k-means clustering showed that eight clusters had the highest average silhouette width (close to 0.6) among the options of 1 to 10 clusters, as shown in Figure 3. Therefore, this study adopted eight as the optimal number of clusters for the imbalanced data set, with the clustering results of k-means for MA and MI shown in Table 2. Cluster 4 had 12,296 patients as the largest cluster, while Cluster 1 had 493 patients as the smallest cluster. Then, we set $m = 1, 2, 3$, and calculated the numbers of selected MA elements in each cluster, as listed in Table 3. The total sizes of SBC for each ratio of $m:1$ were 455, 685 and 911 for $m = 1, 2, 3$, respectively.

Cluster	$Size^j$	$Size^j_{MA}$	$Size^j_{MI}$	$Size^j_{MA}/Size^j_{MI}$
1	493	489	4	122.25
2	3385	3361	24	140.04
3	8600	8549	51	167.63
4	12296	12189	107	113.92
5	1051	1040	11	94.55
6	566	562	4	140.50
7	1800	1793	7	256.14
8	1834	1814	20	90.70
Total	30025	29797	228	1125.72

Table 2: Clustering results of k-means for MA and MI

Ratio		1:1	2:1	3:1
$SSize_{MI}$	$SSize^1_{MA}$	25	50	74
	$SSize^2_{MA}$	28	57	85
	$SSize^3_{MA}$	34	68	102
	$SSize^4_{MA}$	23	46	69
	$SSize^5_{MA}$	19	38	57
	$SSize^6_{MA}$	28	57	85
	$SSize^7_{MA}$	52	104	156
	$SSize^8_{MA}$	18	37	55
	Total	227	457	683
$Size_{MI}$		228		
Total		455	685	911

Table 3: Sample sizes of SBC for MA and MI

Figures 4, 5 and 6 show, respectively, the importance of variables on XGBoost for the ratios 1:1, 2:1, 3:1 sampling at five times, where variables with zero gain for each ratio were excluded. The results of the five-time-resampling in each ratio were similar, and, in accordance with the findings of previous studies, they indicated that elderly people are more likely to suffer from breast cancer [Wolf et al. 2005]. Figure 7 shows the mean importance of variables on XGBoost for sampling with the ratios 1:1, 2:1, and 3:1. The results for each ratio were similar. In addition to age, the variables with mean importance over 0.005 included CVD08, PCD01, PCD03, CVD05, NM04, CVD14, CVD04, CVD09, CVD01, CVD11, CVD02, NM01, NM06, and CVD07. This study adopted the top five important variables to build the XGBoost model for the resampling data sets, and the mean of accuracy and F-measure reached 79.92% and 67.53%, respectively. Then, we set $m = 5$ for sensitivity analysis. The results of sensitivity analysis also showed that age, CVD08, PCD01, PCD03, and CVD05 are the top five important variables. Hence, variables CVD08, PCD01, PCD03, and CVD05 showed important impacts. The findings indicate that occlusion stroke, diabetes with peripheral circulatory disorders, peripheral angiopathy in diseases classified elsewhere, and other forms of chronic ischemic heart disease are important risk factors for breast cancer occurrence in patients with diabetes mellitus.

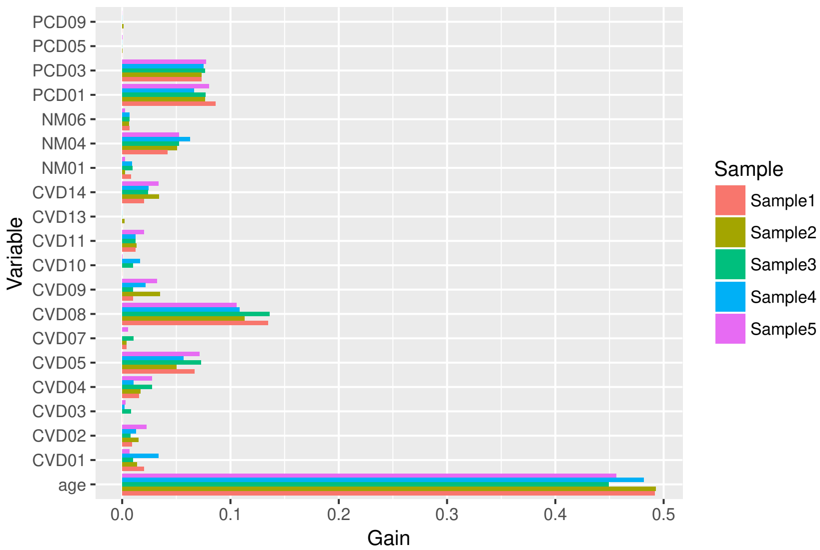

Figure 4: Importance of variables on XGBoost for the ratio of 1:1 sampling five times

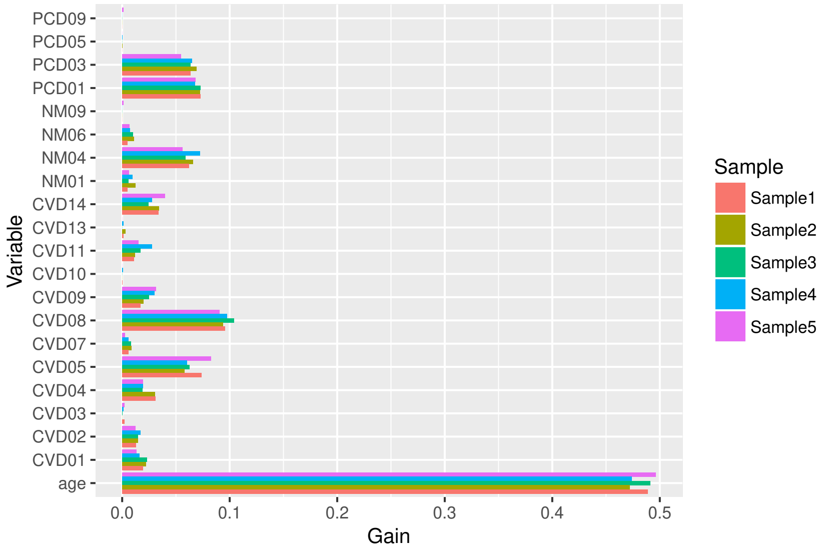

Figure 5: Importance of variables on XGBoost for the ratio of 2:1 sampling five times

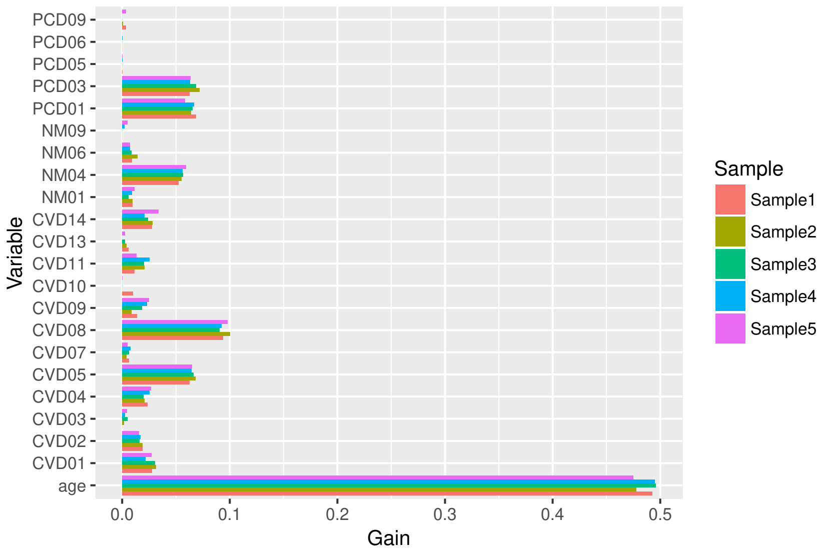

Figure 6: Importance of variables on XGBoost for the ratio of 3:1 sampling five times

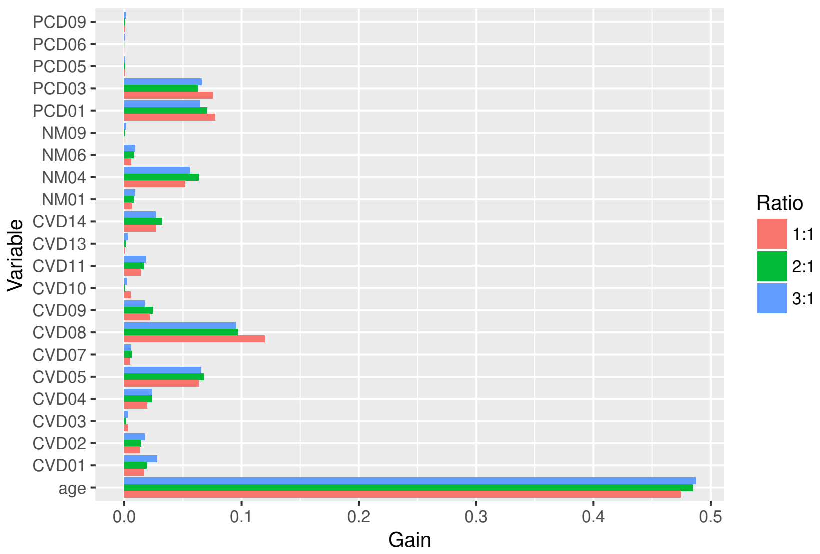

Figure 7: Mean importance of variables

5 Conclusion

With the trend of rapidly aging populations, diabetes mellitus and its comprehensive complications are major public health issues that place a heavy burden on global health care systems. This study proposed a hybrid machine learning scheme to analyze the risk factors of breast cancer occurrence in patients with diabetes mellitus, with consideration for the imbalance in the data. The proposed scheme combined the SBC algorithm, the k-means algorithm, and the XGBoost algorithm. The key findings of this study were (1) occlusion stroke, diabetes with peripheral circulatory disorders, peripheral angiopathy in diseases classified elsewhere, and other forms of chronic ischemic heart disease are risk factors of breast cancer occurrence in patients with diabetes mellitus, and (2) the proposed scheme addressed the issue of imbalanced data.

The findings of this study can assist health care providers to appropriately counsel patients on the risk of breast cancer and improve screening strategies. From the viewpoint of physicians, an increased risk of breast cancer in a patient with diabetes mellitus might necessitate greater consciousness of the physiological changes related to the development of complications and indicate early intervention to decrease the risk of breast cancer.

The hybrid machine learning scheme might also be applied to analyze the impact of complications of diabetes mellitus on other cancers. In addition to the epidemiologic value, the hybrid machine learning scheme provided an informatics value on the assessment of the imbalanced data that might be extended to other issues with imbalanced data, such as financial risk assessment.

This study has several data limitations. We used claim data which does not cover all health care records and might not identify all complications related to diabetes mellitus, which might reduce the generalizability of the findings in this study. This study also used retrospective data from different years of collection to assess the impact of complications on a future outcome. Further studies could thus measure the survival time of outcomes to better understand the associations between the complications of diabetes mellitus and the outcomes. Other future research could make comparisons with other methods, broadening the model to include more complications or applying the proposed scheme to other cancers.

Acknowledgements

This work was supported by the Ministry of Science and Technology, Taiwan [MOST 105-2221-E-030-010 to Mingchih Chen]. The authors would like to acknowledge Ting-Xuan Huang, Mingchih Chen, Ben-Chang Shia, Yefei Jiang, Pay Wen Yu, and Yi-Wei Kao for their valuable contributions to the project.

References

[Abualhaj et al. 2017] Abualhaj, B., Weng, G., Ong, M., Attarwala, A. A., Molina, F., Büsing, K., Glatting, G.: "Comparison of five cluster validity indices performance in brain [18F] FET-PET image segmentation using k-means." Medical physics 44.1 (2017): 209-220.

[Aljawarneh et al. 2016] Aljawarneh, S., A., Federica C., Abdelsalam M.: "Advanced Research on Software Security Design and Applications." Journal of Universal Computer Science 22.4 (2016): 453-458.

[Bach et al. 2017] Bach, M., Werner, A., Żywiec, J., Pluskiewicz, W.: "The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis." Information Sciences 384 (2017): 174-190.

[Benmouiza and Ali 2013] Benmouiza, K., and Ali C.: "Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models." Energy Conversion and Management 75 (2013): 561-569.

[Bosco et al. 2010] Bosco, J. L., Antonsen, S., Sorsensen, H. T., Pedersen, L., & Lash, T. L.: "Metformin and incident breast cancer among diabetic women: a population-based case-control study in Denmark." Cancer Epidemiology and Prevention Biomarkers (2010): cebp-0817.

[Boyle et al. 2012] Boyle, P., Boniol, M., Koechlin, A., Robertson, C., Valentini, F., Coppens, K., ..., Smans, M.: "Diabetes and breast cancer risk: a meta-analysis." British journal of cancer 107.9 (2012): 1608-1617.

[Chen and Guestrin 2016] Chen, T., and Guestrin, C.: "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, (2016).

[Cheng 2015] Cheng, T.-M.: "Reflections on the 20th anniversary of Taiwan's single-payer National Health Insurance System." Health Affairs 34.3 (2015): 502-510.

[Friedman 2001] Friedman, J., H.: "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

[Giovannucci et al. 2010] Giovannucci, E., Harlan, D. M., Archer, M. C., Bergenstal, R. M., Gapstur, S. M., Habel, L. A., Pollak, M., Regensteiner, J. G., Yee, D.: "Diabetes and cancer: a consensus report." CA: a cancer journal for clinicians 60.4 (2010): 207-221.

[He and Edwardo 2009] He, H., and Edwardo A. G.: "Learning from imbalanced data." IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[Hou et al. 2013] Hou, G., Zhang, S., Zhang, X., Wang, P., Hao, X., Zhang, J.: "Clinical pathological characteristics and prognostic analysis of 1,013 breast cancer patients with diabetes." Breast cancer research and treatment 137.3 (2013): 807-816.

[Hsu et al. 2011] Hsu, J. H., Chien, I. C., Lin, C. H., Chou, Y. J., Chou, P.: "Incidence of diabetes in patients with schizophrenia: a population-based study." The Canadian Journal of Psychiatry 56.1 (2011): 19-26.

[Hsu and Hsu 2016] Hsu, W.-C. and Hsu, Y.-P.: "Patterns of outpatient care utilization by seniors under the National Health Insurance in Taiwan." Journal of the Formosan Medical Association 115, 5 (2016), pp. 325--334.

[Jain 2016] Jain A.: "Complete Guide to Parameter Tuning in XGBoost (with codes in Python)."(2016). `https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/`

[Jain 2010] Jain, A., K.: "Data clustering: 50 years beyond K-means." Pattern recognition letters, 2010, 31, 8, 651-666.

[Jiralerspong et al. 2013] Jiralerspong, S., Kim, E. S., Dong, W., Feng, L., Hortobagyi, G. N., Giordano, S. H.: "Obesity, diabetes, and survival outcomes in a large cohort of early-stage breast cancer patients." Annals of oncology (2013): mdt224.

[Larsson et al. 2007] Larsson, S. C., Mantzoros, C. S., Wolk, A.: "Diabetes mellitus and risk of breast cancer: a meta-analysis." International journal of cancer 121.4 (2007): 856-862.

[Liao et al. 2011] Liao, S., Li, J., Wei, W., Wang, L., Zhang, Y., Li, J., Wang, C., Sun, S.: "Association between diabetes mellitus and breast cancer risk: a meta-analysis of the literature." Asian Pac J Cancer Prev 12.4 (2011): 1061-5.

[Liu et al. 2009] Liu, X.-Y., Wu J., Zhou Z.-H.: "Exploratory undersampling for class-imbalance learning." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39.2 (2009): 539-550.

[Maskarinec et al. 2017] Maskarinec, G., Fontaine, A., Torfadottir, J. E., Lipscombe, L. L., Lega, I. C., Figueroa, J., Wild, S.: "The Relationship between Type 2 Diabetes and Breast Cancer Incidence in Differing Ethnic Groups." Canadian Journal of Diabetes (2017).

[MHW 2017] Ministry of Health and Welfare: "Taiwan National Health Insurance Statistical Yearbook in 2015." (2017).

[NDC 2016] National Development Council: "Population Projections for Taiwan: 2016-2060." (2016).

[NHRI 2014a] National Health Research Institutes: "Specific subject datasets." (2016). `http://nhird.nhri.org.tw/en/Data_Subsets.html#S4`

[NHRI 2014b] National Health Research Institutes: "Data files" (2016). `http://nhird.nhri.org.tw/en/Data_Files.html`

[Nie et al. 2008] Nie, J. X., Wang, L. , Tracy, C. S., Moineddin, R., Upshur, R. E.: "Health care service utilization among the elderly: findings from the Study to Understand the Chronic Condition Experience of the Elderly and the Disabled (SUCCEED project)." Journal of evaluation in clinical practice 14, 6 (2008), pp. 1044--1049.

[Pereira et al. 2014] Pereira, S., Fontes, F., Sonin, T., Dias, T., Fragoso, M., Castro-Lopes, J., Lunet, N.: "Neurological complications of breast cancer: study protocol of a prospective cohort study." BMJ open 4.10 (2014): e006301.

[PDDESAUN 2015] Population Division, Department of Economic and Social Affairs, United Nations: "World Population Ageing 2015" (2015).

[Ramezani et al. 2014] Ramezani T., F., Behboudi-Gandevani, S., Ghanbarian, A., Azizi, F.: "Effect of menopause on cardiovascular disease and its risk factors: a 9-year follow-up study." Climacteric 17.2 (2014): 164-172.

[Redaniel et al. 2012] Redaniel, M. T. M., Jeffreys, M., May, M. T., Ben-Shlomo, Y., Martin, R. M.: "Associations of type 2 diabetes and diabetes treatment with breast cancer risk and mortality: a population-based cohort study among British women." Cancer Causes & Control 23.11 (2012): 1785-1795.

[Rousseeuw 1987] Rousseeuw, P., J.: "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.

[Sieri et al. 2012] Sieri, S., Muti, P., Claudia, A., Berrino, F., Pala, V., Grioni, S., ... & Krogh, V.: "Prospective study on the role of glucose metabolism in breast cancer occurrence." International journal of cancer 130.4 (2012): 921-929.

[Tonejc 2016] Tonejc, J., Güttes, S., Kobekova, A., Kaur, J.: "Machine Learning Methods for Anomaly Detection in BACnet Networks." Journal of Universal Computer Science, 2016, 22, 9, 1203-1224.

[Torlay et al. 2017] Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M.: "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy." Brain Informatics (2017): 1-11.

[Tseng 2015] Tseng, C.-H.: "Use of insulin and mortality from breast cancer among Taiwanese women with diabetes." Journal of diabetes research (2015).

[Tait et al. 2014] Tait, S., Pacheco, J. M., Gao, F., Bumb, C., Ellis, M. J., Ma, C. X. : "Body mass index, diabetes, and triple-negative breast cancer prognosis." Breast cancer research and treatment 146.1 (2014): 189-197.

[Wolf et al. 2005] Wolf, I., Sadetzki, S., Catane, R., Karasik, A., & Kaufman, B.: "Diabetes mellitus and breast cancer." The lancet oncology 6.2 (2005): 103-111.

[Wu et al. 2012] Wu, C. Y., Chen, Y. J., Ho, H. J., Hsu, Y. C., Kuo, K. N., Wu, M. S., Lin, J. T. : "Association between nucleoside analogues and risk of hepatitis B virus–related hepatocellular carcinoma recurrence following liver resection." Jama 308.18 (2012): 1906-1913.

[Xie et al. 2009] Xie, X. D., Qu, S. X., Liu, Z. Z., Zhang, F., Zheng, Z. D.: "Study on relationship between angiogenesis and micrometastases of peripheral blood in breast cancer." Journal of cancer research and clinical oncology 135.3 (2009): 413-419.

[Yen and Lee 2009] Yen, S.-J. and Lee Y.-S.: "Cluster-based under-sampling approaches for imbalanced data distributions." Expert Systems with Applications 36.3 (2009): 5718-5727.