應用虛擬樣本方法改善不平衡大數據分類性能

Virtual Sampling Method to Improve Imbalanced Big Data Classification Performance

林良憲
Liang-Sian Lin

摘要

在大數據的時代，企業經常能夠獲得大量的資料建構一個學習模型來進行決策。對大數據而言，如此的學習模型很有可能受到不平衡資料集(imbalanced data set)的影響而產生有偏差的訓練，造成模型傾向於數量較多的類別。因此，使用不平衡類別資料集來建構一個可靠的大數據學習模型是目前企業最重要的挑戰之一。為了解決這個問題，本研究提出一個新的增加少數抽樣(over-sampling)方法來增加少數量類別的數量，提出的方法是使用整體趨勢擴散(mega-trend-diffusion；MTD)技術生成虛擬樣本，以及應用可能性評估機制(plausibility assessment mechanism；PAM)來評估虛擬樣本合適性，其目的在降低分類上的偽陽性率(false positive rate；FPR)而不影響其他評估分類性能之指標如：分類正確性、geometric mean (Gmean)與F1-measure (F1)。在此，我們使用一個模擬資料集來建構支持向量機器(support vector machine；SVM)的分類模型，而實驗結果顯示所提出的方法能夠有效地改善不平衡大數據的分類性能。

Abstract

In the age of big data, enterprise normally can obtain numerous data to build a learning model to make a decision. For big data, such learning model tends to majority class due to imbalanced data set likely leads to a biased training. Hence, using an imbalanced data set to build a reliable learning model for big data is one of the most important challenges in enterprise. For solving this, this paper proposes a new over-sampling method to increase the data size in minority class. The proposed method is to use the mega-trend-diffusion (MTD) technology to generate virtual samples and the plausibility assessment mechanism (PAM) to access the suitability of virtual sample. In addition, this paper is to decrease the false positive rate (FPR) on classification and not to influence the other indices for accessing the classification performance, such as accuracy, geometric mean (Gmean), and F1-measure (F1). In this paper, a simulated data set is used to build the support vector machine (SVM) classification model, and the experiment results show that the proposed method can effectively improve classification performances for imbalanced big data sets.

關鍵詞(Key Words)

不平衡大數據(Imbalanced Big Data；IBD)
增加少數抽樣(Over-sampling)
虛擬樣本(Virtual Sample；VS)
偽陽性率(False Positive Rate；FPR)

1 · 前言

在全球化的競爭下，企業產品推陳出新的速度必須越來越快，而顧客對於產品的品質要求更是越來越高，因此產品的品質好壞嚴重影響顧客對產品的忠誠度以及對產品的黏著度。當顧客使用企業所生產的產品時，若是顧客拿到品質較差的產品，那麼顧客願意再次購買該企業生產的其他相關產品的意願便會降低，使得企業的收益受到損害。企業為了取得顧客的信任，必須不斷地改善技術以降低產品在生產上的不良率。當產品經由改良的技術進行大量生產後，大量的產品數據使得企業必須利用資料探勘方法來判斷產品品質之好壞。然而在品質不良的產品數量較少的情況下，傳統的資料探勘分類器可能將全部的產品歸類為具有好品質的產品，主要由於當使用不平衡資料集來建構分類器時，利用多數量類別建構的分類器，其學習規則不利於較少資料的類別，使得企業生產的少數量不良品判斷為良品的機率可能大大的提升。擁有大數據的企業便會經常遇到這類的問題，我們稱為不平衡大數據問題(imbalanced big data problem)，而該問題已經在企業界與學術界受到越來越多的重視，例如del Río et al. [1] 在專門處理大數據運算效率的MapReduce工具中使用random forest方法改善數個分類技術於不平衡大數據的學習，並且應用業界的三個實際案例來驗證其研究方法之有效性。在醫學基因領域，Galpert et al. [2]的研究中提及人類基因配對(gene pair features)資料數量通常是巨大的且存在不平衡大數據問題，而在生物資訊(bioinformatics)領域，伴隨著生物技術(biotechnology)的發展，研究者已經可以從生物獲得大量關於細胞、蛋白質與基因等數據[3]。

面對不平衡大數據問題，有些學者建議使用減少多數抽樣法(under-sampling)或增加少數抽樣法(over-sampling)，減少多數抽樣法是刪除多數量類別(majority class)的資料，有些學者建議隨機減少多數抽樣法(random under-sampling)，例如：Yen and Lee [4]、Liu [5]以及Tahir [6]，有些學者建議以分群方法為基礎的方式來刪除多餘的資料，例如：Yen and Lee [7]、Zhang et al. [8]與Fu et al. [9]，然而這樣的方式可能會刪除多數量類別的重要資訊，因此有學者建議使用增加少數抽樣法，其增加少數量類別(minority class)的樣本數量。目前以Chawla et al. [10]提出的SMOTE: synthetic minority over-sampling technique方法最常被使用，其主要是隨機地取出原始資料中少數量類別附近的線性資料點成為新的少數量類別資料以改善少數量類別的學習。Chawla et al. [11]更進一步地提出SMOTEBoost方法抽取少數量類別的資料成為新的資料來提升不平衡類別資料集的預測性能。在語意分析領域，Peng and Yao [12]提出AdaOUBoost方法來改善以視頻內容為主的語意學習性能，該學者使用SMOTE方法來生成少數量類別的新資料並且使用減少多數抽樣法縮減多數量類別的資料來建構不平衡數據下學習性能較佳的分類器。除了取出原始的樣本成為新的樣本外，有學者建議生成潛在於資料間隔(data gap)之間的資料點當作虛擬樣本(virtual sample)。例如Zhang and Wang [13]提出一個以常態分佈(normal distribution)為基礎的減少多數抽樣法來改善在非線性資料的不平衡類別分類性能。其研究方法為設定少數量類別的資料來自一個常態分佈，並使用估計的分佈生成虛擬樣本，接著虛擬樣本被放入至原始資料集以強化多種分類器的學習性能。

在虛擬樣本生成技術方面，Li et al. [14]提出整體趨勢擴散(mega-trend-diffusion；MTD)技術估計小樣本的母體值域，並在該值域內生成虛擬樣本以增強倒傳遞類神經網路模型的學習來改善解決初期製造系統時遭遇的產品數量不足之情況。Li et al. [15]、Li et al. [16]、Li et al. [17]與Luor [18]已經應用MTD技術在他們的研究當中來改善小資料集的學習問題。此外，虛擬樣本的合適性可能影響小樣本的學習情況，因此Li et al. [19]提出可能性評估機制(plausibility assessment mechanism；PAM)來評估生成的虛擬樣本其是否適用於目前的小資料集母體。由以上的研究可了解虛擬樣本生成技術對於資料量較少的學習能夠有顯著的改善。因此為了解決企業的不平衡大數據資料集問題，我們使用增加少數抽樣法來增加較少資料

量的類別其資料數量。本研究提出的增加少數抽樣法是去生成虛擬樣本並設定為少數量類別的新資料點。在此，虛擬樣本生成方式為應用容易且受歡迎的整體趨勢擴散技術，並且使用可能性評估機制來評估虛擬樣本的合適性。在此，我們使用一個模擬的資料集來驗證本研究所提方法能夠改善大數據的不平衡類別資料集的學習性能。我們所提出的方法主要在降低偽陽性率(false positive rate；FPR)值同時，不影響分類正確性、geometric mean (Gmean)、F1-measure (F1)。在分類模型選擇上，我們使用(support vector machine；SVM)分類模型的radial basis function (RBF) kernel function來進行分類，而實驗結果顯示本研究提出的方法比未使用增加少數抽樣法可以得到更好的分類性能。

剩餘的章節為第二章說明不平衡資料集分類性能評估以及學習模型；第三章詳細地說明所提出的方法與使用流程，以及整體趨勢擴散技術與可能性評估機制的介紹。第四章我們說明模擬的資料集與詳細的實驗步驟，並且呈現我們的實驗結果。第五章為本研究之結論。

2 · 文獻探討

本章節說明分類性能評估的計算方式，以及支援向量機的學習方式。

2.1分類性能評估

面對不平衡資料集，我們假設少量資料為負的類別(negative class)，較多數量的為正類別(positive class)。如表1，二元分類問題經常使用混淆矩陣(confusion matrix)於分析二元分類的學習性能。

表 1　Confusion matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

在表1，TP (true positive)為真陽性的數量、FP (false positive)為偽陽性的數量、FN (false negative)為偽陰性的數量、TN (true negative)為真陰性的數量，其用以計算分類正確性(Accuracy)、Gmean與F1其越高代表分類性能越好，FPR其越低代表分類性能越好。

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$\text{Gmean} = \sqrt{TPR \times TNR} \tag{2}$$

$$\text{TRP(true positive rate)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{TNP(true negative rate)} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1} = 2 \times \frac{R \times P}{R + P} \tag{5}$$

$$\text{R} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{P} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{FPR(false positive rate)} = \frac{FP}{TN + FP} \tag{8}$$

2.2支援向量機

支援向量機(support vector machine；SVM)由Cortes and Vapnik [20]提出的方法，主要用於處理高維度空間的非線性分類問題，其為一種最小化風險的演算法，在高維度空間中找到最佳分割超平面(hyperplane)使得二元分類資料能有效分開，同時二元分類資料存在最大邊界(maximum margin)，如圖1。

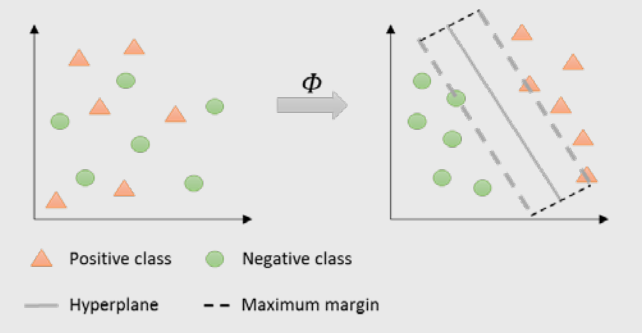

圖 1　透過Φ函數轉換資料至可分割空間

藉由使用kernel Φ函數將非線性二元分類資料轉換至可分割線性空間Campbell [21]，如此資料可在此空間中可被一個超平面所分割。此外，一般常用的kernel函數有linear, polynomial, radial basis function與sigmoid 函數。

3 ．本研究方法與流程

本研究使用虛擬樣本生成技術改善大數據不平衡類別資料集在SVM分類器上的學習性能。在此章節，我們將詳細地說明本研究的虛擬樣本生成流程並介紹整體趨勢擴散(MTD)技術與可能性評估機制(PAM)。

3.1 虛擬樣本生成流程

由於少數量類別的樣本數量不足，我們透過虛擬樣本產生法(virtual sample generation method)增加樣本數量。在生成虛擬樣本前我們必須先估計一個合適的樣本分佈，並從該分佈生成相當數量的樣本。本研究使用MTD技術估計樣本之可能範圍，接著利用PAM判斷生成的虛擬樣本其合適性。最後將多數量類別資料集、少數量類別資料集以及虛擬樣本集結合成一個新的訓練資料集，接著使用該訓練資料建構SVM模型，如圖2。

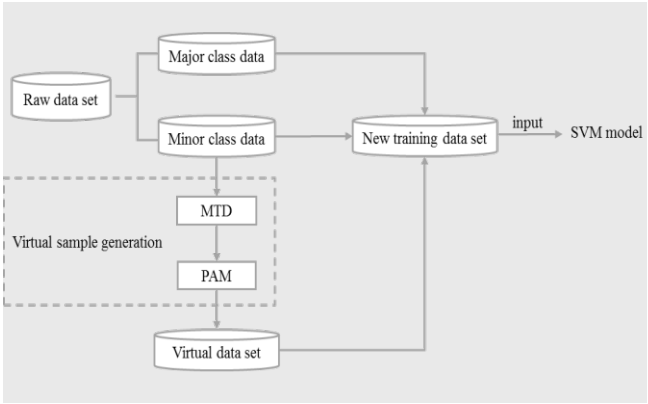

圖 2 本研究執行流程

3.2整體趨勢擴散技術

少數量類別的資料存在著資料結構不完整的問題，各樣本點間存在著資料間隙。Li等學者於2007年[14]提出整體趨勢擴散技術生成虛擬樣本來填補這些間隙，該技術以三角隸屬函數(membership function；MF)推估母體值域以生成虛擬樣本，如圖3。

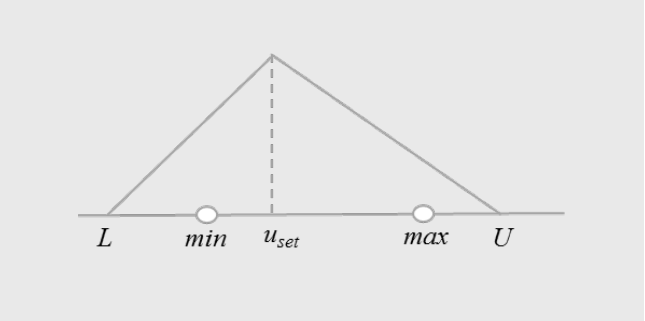

圖 3 整體趨勢估計

由式子(9)與(10)建構母體值域範圍[L, U。

$$L=\begin{cases}\dfrac{\min+\max}{2}-\dfrac{N_L}{N_L+N_U}\times\sqrt{-2\times\dfrac{\hat{s}_x^2}{N_L}\times\ln\left(10^{-20}\right)} & ,\ L\le\min\\ \min & ,\ L>\min\\ \dfrac{\min}{5} & ,\ N_L=0\end{cases}\tag{9}$$

$$U=\begin{cases}\dfrac{\min+\max}{2}+\dfrac{N_U}{N_L+N_U}\times\sqrt{-2\times\dfrac{\hat{s}_x^2}{N_U}\times\ln\left(10^{-20}\right)} & ,\ U\ge\max\\ \max & ,\ U<\max\\ \max\times 5 & ,\ N_U=0\end{cases}\tag{10}$$

其中max為觀測值中的最大值、min為觀測值中的最小值、$(\max+\min)/2$為中點、N_L代表大於中點之資料筆數、N_U代表小於中點之資料筆數。$N_L/(N_L+N_U)$為左偏係數，$N_U/(N_L+N_U)$為右偏係數，分別代表擴散函數中擴展幅度兩邊的比重，$\hat{s}_x^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)$為樣本變異數。

3.3可能性評估機制

整體趨勢擴散技術利用三角隸屬函數值隨機產生更多的樣本點以填補原始資料中的資料間隙，此技術的隨機生成方式是根據均等分配，與其原始資料集所推論的母體分配可能有所差異。Li et al. [19]提出可能性評估機制來解決這個問題，其機制是去模擬生成能符合估計的母體分配其虛擬樣本的隨機性機率值。藉由該機制可以檢視隨機產生的虛擬樣本是否合適保留下來。當估計出母體值域的上下界[L,U]，在[L,U]區間隨機生成的一個數值tv(temporary value)，並利用式子(11)計算tv的MF值

$$MF(tv)=\begin{cases}1, & tv=u_{set}\\ \dfrac{tv-L}{u_{set}-L}, & tv<u_{set}\\ \dfrac{U-tv}{U-u_{set}}, & tv>u_{set}\end{cases} \quad (11)$$

其中 $u_{set}=(\max+\min)/2$ ，如圖4。

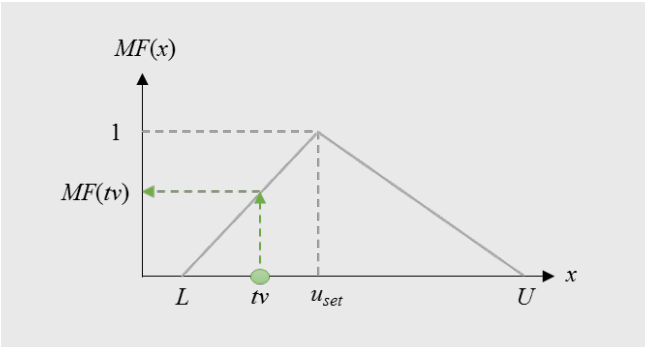

圖 4　MF(tv) 生成

該機制產生一介於[0,1]之間來自均等分配的隨機種子rs(random seed)。當rs小於tv的MF值，該tv會被保留成為適當的虛擬樣本v；若相反，則捨棄該tv並重新產生新的tv值，如式(12)。

$$v=tv\,|\,rs<MF(tv) \quad (12)$$

4 · 實例驗證

本研究模擬一個資料集來驗證所提出的方法之有效性，在此我們利用該資料建構SVM分類模型，驗證在SVM模型下所提出方法具有較佳的學習性能。

4.1資料說明

在此模擬一組二元分類資料，其中類別為判斷生產的產品是否具有合格的性能，符合標準的合格產品設定為positive class，而不合格產品則設定negative class。本研究產生1000筆資料，並設定每筆資料有5個屬性 $\{X_1, X_2, X_3, X_4$，並且positive class為 $\{Y\,|\,y_i=0\}$，而negative class為 $\{Y\,|\,y_i=1\}$，如表2。

表 2　1000筆模擬的訓練資料

No.	X_1	X_2	X_3	X_4	X_5	Y
1	-1.68	0.02	1.21	1.85	-0.05	0
2	-0.03	0.69	-0.91	-0.13	1.12	0
3	-0.87	-0.32	0.09	-0.85	1.68	1
4	0.55	0.14	-0.18	-0.81	0.76	1
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	y_i

4.2實驗設計

本研究使用python執行所提出的方法，並建構SVM學習模型，其中核函數選用RBF kernel function，其參數設定為調整常數（C）為1、gamma參數為0。假設之資料集的N筆資料，其中 N_{majo} 與 N_{minor} 分別為多數量類別的數量與少數量類別的數量。在此，我們設定 $r=N_{maj}/N_{min}$，然後以隨機抽取的方式抽取 $N_{train}=1000$ 筆為訓練樣本，其另外模擬1000筆資料當作測試樣本，並重複進行30次實驗，在此生成的虛擬樣本數量為 $N_{vir}=\{10,20,30,40,50,100,200,300\}$。

4.3實驗結果

本研究重複30次實驗用以檢驗在不同虛擬樣本數量之下對實驗結果的影響，並計算進行30次實驗Accuracy、Gmean、F1、FPR的平均值與標準差，如表3與4。

表 3　Accuracy、Gmean、F1、FPR 平均值

=0.7	Accuracy		Gmean		F1		FPR	
N_{vir}	RAW	PM	RAW	PM	RAW	PM	RAW	PM
10	0.74	0.74	0.52	0.54	0.83	0.83	0.70	0.68
20	0.74	0.73	0.50	0.53	0.83	0.83	0.73	0.68
30	0.74	0.73	0.51	0.56	0.83	0.83	0.72	0.65
40	0.74	0.73	0.51	0.56	0.83	0.82	0.72	0.65
50	0.74	0.73	0.50	0.56	0.83	0.82	0.73	0.64
100	0.74	0.73	0.51	0.58	0.83	0.82	0.71	0.61
200	0.74	0.73	0.52	0.58	0.83	0.82	0.71	0.62
300	0.74	0.73	0.50	0.58	0.83	0.82	0.72	0.62

表 4　Accuracy、Gmean、F1、FPR 標準差

=0.7	Accuracy		Gmean		F1		FPR	
N_{vir}	RAW	PM	RAW	PM	RAW	PM	RAW	PM
10	0.01	0.01	0.03	0.03	0.01	0.01	0.04	0.04
20	0.01	0.01	0.04	0.04	0.01	0.01	0.05	0.06
30	0.01	0.01	0.05	0.03	0.01	0.01	0.05	0.05
40	0.01	0.01	0.03	0.03	0.01	0.01	0.03	0.04
50	0.01	0.01	0.04	0.03	0.01	0.01	0.04	0.04
100	0.01	0.02	0.04	0.03	0.01	0.01	0.05	0.04
200	0.01	0.01	0.04	0.03	0.01	0.01	0.04	0.05
300	0.01	0.01	0.03	0.02	0.01	0.01	0.04	0.03

每次實驗過程中，我們設定訓練樣本數量為1000以及 r 為0.7，即較少類別僅有300筆資料。實驗結果如表3，所提出的方法(PM)增加虛擬樣本至原始資料集中與僅有原始資料(RAW)在Accuracy與F1分類性能評估指標上無顯著差異。此外，Gmean值由0.54增加到0.58與FPR值由0.68下降到0.62。因此在分類器為SVM下，所提出的方法可降低壞產品判斷為好產品的誤判率。

在表3的實驗結果證實增加虛擬樣本至原始資料集中可以讓不平衡資料集的學習獲得改善，然而若生成不適當的虛擬樣本可能反而導致拙劣的學習性能。為了穩定學習性能，這篇研究使用PAM來評估已生成的虛擬樣本之合適性。實驗結果如表4，所提出的方法與未使用虛擬樣本的方法在四個分類性能評估指標(Accuracy, F1, Gmean, FPR)上有相當接近的變異程度，因此所提出的方法能夠有效且穩定地改善不平衡資料集的分類性能。

5．結論

隨著技術的進步，企業的產品大量製造後，我們所生產的產品大部分為合格的產品，因此不合格產品所觀察到樣本數量便很少。因此當處理大數據資料集時，我們在大數據分類上經常會遇到不平衡類別問題，其容易造成分類性能的下降。根據過往的研究增加少數抽樣法已經被證實在解決不平衡類別分類的問題上為一個有效的方法。

在本研究，我們使用整體趨勢擴散技術生成虛擬樣本，並使用可能性評估機制來評估虛擬樣本的合適性以增加數量較少類別的資料量。由實驗結果顯示本研究驗證提出的方法比不加入虛擬樣本時，在不平衡類別分類上的學習性能有更好表現。在未來研究方面，我們將使用實際的案例驗證本研究所提出的方法比其他增加少數抽樣法可以得到更好的學習性能。

參考文獻

[1] S. del Río, V. López, J. M. Benítez, and F. Herrera, “On the use of MapReduce for imbalanced big data using Random Forest,” Information Sciences, vol. 285, pp. 112-137, 2014.

[2] D. Galpert, S. del Río, F. Herrera, E. Ancede-Gallardo, A. Antunes, and G. Agüero-Chapin, “An Effective Big Data Supervised Imbalanced Classification Approach for Ortholog Detection in Related Yeast Species,” BioMed research international, vol. 2015, 2015.

[3] I. Triguero, S. del Río, V. López, J. Bacardit, J. M. Benítez, and F. Herrera, “ROSEFW-RF: the winner algorithm for the ECBDL’14 big data competition: an extremely imbalanced big data bioinformatics problem,” Knowledge-Based Systems, vol. 87, pp. 69-79, 2015.

[4] S. J. Yen and Y. S. Lee, “Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset,” Intelligent Control and Automation, Springer, 2006, pp. 731-740.

[5] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class-imbalance learning,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 39, pp. 539-550, 2009.

[6] M. A. Tahir, J. Kittler, and F. Yan, “Inverse random under sampling for class

imbalance problem and its application to multi-label classification," Pattern Recognition, vol. 45, pp. 3738-3750, 2012.

[7] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," Expert Systems with Applications, vol. 36, pp. 5718-5727, 2009.

[8] Y. P. Zhang, L. N. Zhang, and Y. C. Wang, "Cluster-based majority under-sampling approaches for class imbalance learning," IEEE International Conference on Information and Financial Engineering, 2010, pp. 400-404.

[9] Y. Fu, G. Gao, and S. Liu, "Assessment based on Monte Carlo for sample rotation under stratified cluster sampling," Proceedings of the 2014 Annual Congress on Advanced Engineering and Technology, 2014, p. 265.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, pp. 321-357, 2002.

[11] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," Knowledge Discovery in Databases: PKDD 2003, Springer, 2003, pp. 107-119.

[12] Y. Peng and J. Yao, "AdaOUBoost: adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets," Proceedings of the international conference on Multimedia information retrieval, 2010, pp. 111-118.

[13] H. Zhang and Z. Wang, "A normal distribution-based over-sampling approach to imbalanced data classification," Advanced Data Mining and Applications, Springer, 2011, pp. 83-96.

[14] D. C. Li, C. S. Wu, T. I. Tsai, and Y. S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," Computers & Operations Research, vol. 34, pp. 966-982, 2007.

[15] D. C. Li, C. C. Chen, C. J. Chang, and W. K. Lin, "A tree-based-trend-diffusion prediction procedure for small sample sets in the early stages of manufacturing systems," Expert Systems with Applications, vol. 39, pp. 1575-1581, 2012.

[16] D. C. Li, C. W. Liu, and W. C. Chen, "A multi-model approach to determine early manufacturing parameters for small-data-set prediction," International Journal of Production Research, vol. 50, pp. 6679-6690, 2012.

[17] D. C. Li, L. S. Lin, and L. J. Peng, "Improving learning accuracy by using synthetic samples for small datasets with non-linear attribute dependency," Decision Support Systems, vol. 59, pp. 286-295, 2014.

[18] D. C. Luor, "A comparative assessment of data standardization on support vector machine for classification problems," Intelligent Data Analysis, vol. 19, pp. 529-546, 2015.

[19] D. C. Li, C. C. Chen, W. C. Chen, and C. J. Chang, "Employing dependent virtual samples to obtain more manufacturing information in pilot runs," International Journal of Production Research, vol. 50, pp. 6886-6903, 2012.

[20] C. Cortes and V. Vapnik, "Support-vector

networks," Machine learning, vol. 20, pp. 273-297, 1995.

[21] C. Campbell, "Kernel methods: a survey of current techniques," Neurocomputing, vol. 48, pp. 63-84, 2002.

作者簡介

林良憲

Liang-Sian Lin is a data engineer at the Information and Communications Research Laboratories, Industrial Technology Research Institute, Taiwan. As a data engineer, his current interests concentrate on small data set learning and big data analysis. His articles have appeared in European Journal of Operational Research and Decision Support Systems.